



E
E
F
ducation
ndowment
oundation

Improving Numeracy and Literacy

Evaluation report and Executive summary

June 2015

Independent evaluators:

Jack Worth, Juliet Sizmur, Rob Ager & Ben Styles (NFER)



**Evidence for
Excellence in
Education**

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; and 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:

Peter Henderson
Research Officer
Education Endowment Foundation
9th Floor, Millbank Tower
21-24 Millbank
SW1P 4QP

p: 020 7802 1679
e: peter.henderson@eefoundation.org.uk
w: www.educationendowmentfoundation.org.uk

About the evaluator

The project was independently evaluated by a team from the National Foundation for Educational Research. The project director was Dr Ben Styles and the project leader was Jack Worth. The impact evaluation was carried out by Jack Worth and the process evaluation by Juliet Sizmur and Rob Ager.

Contact details:

Dr Ben Styles
The Mere
Upton Park
Slough
Berkshire
SL1 2DQ

p: 01753 637386

e: b.styles@nfer.ac.uk

Contents

Executive summary	4
Introduction	6
Methodology	9
Impact evaluation	14
Outcomes and analysis	18
Process evaluation	26
Outcomes	33
Conclusion	39
References	43
Appendix 1: Covering letter to headteacher	44
Appendix 2: Opt-out consent letter to parents	46
Appendix 3: Randomisation syntax	48
Appendix 5: Primary analysis	57
Appendix 6: Security classification of trial findings	78
Appendix 7: Cost rating	79

Executive summary

The project

The *Improving Numeracy and Literacy* project aimed to improve the numeracy and literacy abilities of pupils in Year 2 through two separate programmes of teacher training and accompanying teaching materials and computer games. The *Mathematics and Reasoning* programme aimed to develop children's understanding of the logical principles underlying mathematics, and the *Literacy and Morphemes* programme aimed to improve spelling and reading comprehension by teaching children about sentence structure and morphemes. Morphemes are components of words that are either stems, which can often appear as words on their own (such as 'fair'), or affixes, which cannot be words on their own (such as 'un-' or '-ly'). The programmes were originally developed (with the support of the ESRC-TLRP Research Programme) by Professor Terezinha Nunes and Professor Peter Bryant at the Department of Education, University of Oxford.

Both interventions were designed to last for 10 to 12 weeks with children receiving one hour of instruction per week as one of their normal literacy or numeracy lessons. Teachers in intervention schools attended a day of training aimed at introducing them to the programmes, explaining the concepts, and allowing them to explore the learning activities for themselves. This was followed by a visit from a member of the research team based at Oxford to support with programme implementation.

Fifty-five schools were recruited to participate in the evaluation by the University of Oxford team: 17 were allocated to the numeracy group, 19 to the literacy group and 19 to the control group.

Key Conclusions

1. This evaluation provided evidence that the *Mathematics and Reasoning* programme had a positive impact on pupils' numeracy ability equating to three additional months' progress.
2. There was no evidence to suggest that the *Literacy and Morphemes* programme had an impact on pupils' literacy ability overall.
3. There was an association between greater use of the accompanying computer games and greater impact in the numeracy intervention, suggesting the computer games were important to successful implementation.
4. All teachers were able to implement the programmes, but most agreed there was too much content to deliver in one hour per week and so made various adaptations to their delivery of the programme. In future trials of the programmes, teachers should be permitted to use and integrate the materials in their own way, as they would in a normal teaching situation.
5. A future trial could evaluate the programmes at scale in more than one location. When drawing up plans for bringing the programmes to scale, the Oxford team should consider whether training and ongoing technical support could be delivered remotely, rather than in person.

How secure is this finding?

Security rating awarded as part of the EEF peer review process

The findings from this trial have high security. The evaluation was well designed and successfully implemented. The programmes were evaluated using a randomised controlled trial that compared the progress of pupils who received the programmes to a 'business as usual' control group. This evaluation was an efficacy trial. Efficacy trials aim to test whether an intervention can work under ideal conditions, with intensive support from the intervention's developer.

Schools were randomised into the three groups and informed of which group they were in after the baseline testing was completed. No schools dropped out of the study, but around 12% of pupils were excluded from the analysis because of missing test data. The proportion of missing pupils was similar in intervention and control groups, but bias may have been introduced if the intervention and control pupils dropped out for different reasons. Analysis of the results using predictions of the missing pupil data suggested that the missing data is unlikely to have affected the result.

The teaching resources were restricted to the teachers in the intervention schools, so there was no risk of the control group schools implementing elements of the intervention and ‘contaminating’ the trial. Testing to measure the outcomes was independently administered by NFER test administrators who were not told which group (literacy intervention, numeracy intervention, or control) the school had been allocated to, and tests were marked externally. The findings of the process evaluation are also secure: teacher questionnaires were distributed and collected by test administrators during the school visits and, as a result, the response rates were very high at 82%.

Results

The evaluation of the *Mathematics and Reasoning* programme provided evidence that it had a positive impact on pupils’ numeracy ability: pupils who received the programme made, on average, the equivalent of three months’ additional progress over the course of a year. Although pupils who received the *Literacy and Morphemes* programme made slightly less progress than the control group, this difference was too small to confidently conclude that it was caused by the intervention. The evaluation therefore provided no evidence that *Literacy and Morphemes* had an impact on literacy ability. In both evaluations, pupils eligible for free school meals (FSM) made slightly more progress if they participated in the programmes, but we are not able to conclude that these observed effects were caused by the programmes themselves rather than occurring by chance.



The teacher questionnaire indicated that teachers found the interventions straightforward to implement. All teachers were able to use the teaching units to deliver the numeracy or literacy content in weekly sessions.

This evaluation builds on previous work conducted by the Oxford University team that suggested that the programmes had evidence of promise. In these studies, the interventions were delivered either one to one or in small groups by researchers familiar with the intervention materials and the underlying theory. In addition, the measures used were closely aligned to the concepts being taught. The purpose of this trial was to evaluate the efficacy of teacher delivery in a whole-class situation and examine the impact on pupils’ overall performance in numeracy and literacy.

How much does it cost?

Over the single year of this evaluation, each intervention cost £21 per pupil (£520 per class). This figure includes the cost of training and the school visit, as well as costs associated with implementing the programme. Teachers required one day of supply cover to enable them to attend the training for each programme during term time: this may or may not have incurred additional cost depending on how it was dealt with by the school.

The estimated cost of implementing the programme over three years is £10 per pupil per year (£257 per class per year). Access to computers and/or tablets is a factor in successful implementation, which may require additional investment for some schools in the future.

Group	Effect size	Estimated months’ progress	Evidence strength	Cost Rating*
Mathematics and Reasoning	0.20	+3 months		£
Literacy and Morphemes	-0.05	-1 months		£
Mathematics and Reasoning (FSM only)	0.14	+2 months		£
Literacy and Morphemes (FSM only)	0.10	+2 months		£

*More information about the cost rating can be found on p.79.

Introduction

The project, *Oxford Improving Numeracy and Literacy Programme*, was delivered by Oxford University Department of Education. This evaluation tested two different initiatives with Year 2 children: *Mathematics and Reasoning* and *Literacy and Morphemes*.

The *Mathematics and Reasoning* programme aimed to develop children's understanding of the logical principles underlying mathematics. The *Literacy and Morphemes* programme aimed to teach children about morphemic spelling rules with the aim of aiding their spelling and also their reading comprehension. Both interventions were designed to last for 12 weeks with children receiving one hour of instruction per week.

For each intervention, teachers from the respective groups received one day of training prior to implementation. For each intervention, lessons were designed to be delivered weekly, replacing one literacy or numeracy lesson. Teaching aims and lesson plans were included in the intervention materials. The lessons normally included teacher-led activities, supported by PowerPoint presentations, and worksheets for pupils. These learning activities were supported by online games that the children could access and do individually at school and at home. This meant that the number of times that children completed online activities could be logged. Ongoing technical support was provided to teachers as and when required by the University of Oxford team.

Background evidence

Both programmes to be evaluated had been tested in intervention studies. Nunes *et al.* (2007) showed that logical reasoning training had an effect of 1.2 standard deviations on mathematics achievement. Morphological training had been tested in a study with both active and no treatment control groups and children given morphological training showed improvements over and above both groups in terms of spelling (Nunes *et al.*, 2003). The intervention had an effect size of between 0.3–0.6 standard deviations, depending on how it was delivered.

Mathematics and Reasoning

Two types of ability have been identified by Nunes and Bryant as essential for mathematical understanding: number sense (what numbers mean and how they relate to each other) and quantitative reasoning (reasoning logically about numbers). In a five-year longitudinal study, Nunes and Bryant demonstrated that these two abilities, though interdependent, could be measured separately. Their research showed that each of the abilities measured at age eight to nine predicted later performance at ages 11 and 14 (Nunes *et al.*, 2012). This research also showed that the ability to reason about quantities was a far stronger predictor of mathematical achievement than arithmetic skill.

The *Mathematics and Reasoning* numeracy intervention is based on a programme previously used by Nunes and Bryant (Nunes *et al.*, 2007) that improved the reasoning ability of children considered at risk of having difficulty with learning mathematics, and improved their mathematics attainment. Year 1 children performing at the 20th percentile or lower who took part in the programme were placed just above the 50th percentile in the National Curriculum assessments at the end of Year 2. Pupils of the same ability from the Year 1 group who did not participate in the programme performed at the 28th percentile. The same programme was adapted to improve the mathematical attainment of deaf children.

In this 2007 research, Nunes and Bryant also demonstrated the importance of logical reasoning for mathematical understanding. This included the logical relationships between numbers, inverse relationships between different operations (such as adding and subtracting), and additive composition (the understanding that every number can be represented as the sum of two other numbers, or, inversely, that every number can be broken down into smaller numbers). Nunes and Bryant demonstrated that children's logical understanding of mathematics when they started school predicted

the progress that they made over a 16-month period. This relationship was still strong even after general cognitive ability and working memory had been controlled for.

Literacy and Morphemes

The current focus of early literacy learning in England is the connection between letter sounds as they are written (graphemes) and as they are said (phonemes). Understanding the connection between phonemes and graphemes helps children to read (Bradley and Bryant, 1983).

Another set of elements of written words are the morphemes from which they are composed. Morphemes are units of meaning, and are either stems (which can often appear as words on their own) or affixes which cannot be words on their own. For example, 'fair' is a morpheme, to which the affixes 'un-' and '-ly' can be added to alter its meaning. This intervention intended to teach children about morphemes and help them to learn how an awareness of morphemes could help them with their spelling and comprehension. For example, knowledge of the past tense morpheme 'ed' helps to distinguish 'missed' and 'mist'. The intervention's focus on morphemes was also expected to draw children's attention to grammatical categories and the structure of sentences.

In 2012, Nunes, Bryant and Barros also showed that children's skills in using both grapheme–phoneme correspondences (GPC) and morphemic units in reading and writing play an important part in learning to read. Children were assessed on these strategies and their reading comprehension at ages nine or ten years. Four years later their reading fluency was tested. Use of morphemic units predicted reading proficiency more strongly than the use of GPC, but both variables made significant and independent contributions to the children's reading, even after the effects of verbal IQ differences had been controlled.

In a further study, Bryant, Nunes and Barros (2014) looked beyond reading attainment to wider school attainment, and the connection to GPC and morphemic skills. The study used data from a longitudinal study linking data on children's use of GPC and morphemic units in reading and writing (at ages eight or nine years) and their achievement in Key Stage 2 and Key Stage 3 assessments (at the ages of 11 and 14) in English, mathematics and science. The study found a strong link between these skills and school achievement, even when the children's IQ was taken into account. Morphemic skills were more strongly predictive of school achievement than GPC skills, strongly mediated by reading ability—an ability that generally influences success in English, mathematics and science.

Morphological training had been tested in an earlier study (Nunes *et al.*, 2003) with both active and no treatment control groups. In this study, children given morphological training showed improvements over and above both groups in terms of spelling, with the intervention having an effect size of between 0.3 and 0.6 standard deviations, depending on how it was delivered. Nunes and Bryant demonstrated that morphemic understanding at ages eight and nine was a strong predictor of reading comprehension and fluency at ages 12 and 13, with several intervention studies showing that improving children's awareness of morphemes has a positive effect on their spelling and word recognition.

Evaluation objectives

The objectives of the impact evaluation were to measure the impact of the *Literacy and Morphemes* and *Mathematics and Reasoning* programmes on pupils' development of literacy and numeracy ability respectively, by comparing with a 'business as usual' control group. The purpose of the process evaluation was to assess how well the interventions were implemented and identify any barriers that may exist to its wider roll-out.

Project team

The *Literacy and Morphemes* and *Mathematics and Reasoning* programmes were designed by Professor Terezinha Nunes and Professor Peter Bryant from the University of Oxford, respectively.

The pre-testing was delivered by a group of postgraduate students and researchers from Oxford, co-ordinated by Deborah Evans.

The evaluation was directed by Dr Ben Styles. The impact evaluation was led by Jack Worth, with assistance from Katie Pyle and Michael Neaves, and Juliet Sizmur and Rob Ager carried out the process evaluation.

Ethical review

At the outset of the trial, the project director considered that all aspects of the trial, including the approach of gaining headteacher consent with a parental opt-out (see 'Eligibility' section), complied with NFER's Code of Practice. The project also underwent a separate ethical review by the University of Oxford.

Trial registration

This trial has been registered on the international standard randomised controlled trial number (ISRCTN) registry at <http://www.isrctn.com/ISRCTN20621819>.

Methodology

Trial Design

The trial used a cluster-randomised design to compare pupil outcomes in intervention schools with pupil outcomes in control schools. Eligible schools were randomised into one of three groups: a set of schools that received *Literacy and Morphemes* training (the literacy group), a set of schools that received *Mathematics and Reasoning* training (the numeracy group) and a set of schools that received no training and continued to teach literacy and numeracy in the way they would normally.

The cluster-randomised design was chosen because the intervention is designed to be taught to a whole class and consists of training and teaching materials for a classroom teacher. Three arms were chosen in preference to two (a literacy group and a numeracy group, each being used as the control group for the other) because of the potential for the literacy intervention to have an impact on numeracy, and vice versa. We tested the hypothesis of transfer effects as one of the secondary analyses.

Eligibility

Primary or infant schools with Year 2 pupils were eligible to participate in the trial. Special schools and independent schools were not eligible. All teachers of Year 2 pupils within the recruited schools were eligible for the training, and all pupils in those teachers' classes were eligible to participate in the trial.

Following initial contact with schools, we sought formal permission to participate in the trial from headteachers through a memorandum of understanding. We sent a letter to headteachers introducing the organisations carrying out the research, and setting out clearly the expectations of participating schools and explaining the data that would be collected (see Appendix 1).

A letter to the parents of each child gave them the opportunity for their child to (a) opt out of participating in the trial entirely, or (b) to opt out of the data collected about their child being matched to the National Pupil Database (see Appendix 2). The letter was provided to schools to send home and responses were collected when schools uploaded pupil information to NFER.

Intervention

Both interventions involved a 10- to 12-week programme of lessons, each of which provided trained teachers with lesson plans and materials to deliver the programme units.

Mathematics and Reasoning

Quantitative reasoning—understanding the relations between numbers and being able to use them to solve problems—formed the basis of the numeracy intervention. Four types of quantitative reasoning were explored: inversion; additive reasoning; multiplication and one-to-many correspondence; and division and sharing.

The numeracy intervention had two strands: understanding of number and problem solving. Lessons about understanding of numbers used additive composition and place value (both set in the context of money, such as counting on, exchanging coins, and comparing and composing amounts). Problem solving looked at inversion, and how different relations can be established (additive and multiplicative) and division. The problem solving activities were designed to focus attention on the need to reason about relations between quantities and to encourage pupils to build their own representations of one-to-many relationships.

Each lesson was designed to cover two different mathematical concepts (for example, money and negative numbers) and, like the literacy lessons, used whiteboard presentation, worksheets, board games and online tasks.

Literacy and Morphemes

The intervention was set out as a 10- to 12-week programme detailed in ten lesson plans with accompanying white board presentations and handouts. Each lesson started with whole-class work, with the aim of getting pupils to provide reasoned answers interactively. The whole-class work was followed by individual work (two levels of which were available). Each of the ten lessons was meant to be completed before moving on to the next.

There were two strands to the lessons: (1) composition and comprehension, with a focus at the sentence level with activities looking at sentence structure, and (2) spelling, with a focus at the word level, exploring morphemic structure.

In linguistics, a morpheme is the smallest grammatical/meaningful unit in a language. Sentence structure was explored by, for example, looking at subject, verb and object within sentences. Spelling tasks were designed to highlight how linking morphemes in different ways could change the meaning of words, such as by adding prefixes or suffixes, or by changing word endings. For example, the addition of the suffix 'ed' to the word 'walk' indicates that it happened in the past.

Online tasks were provided as part of the individual work to be carried out in lessons. These tasks were designed to further explore the areas covered in the lesson and, after the completion of a number of tasks, pupils were rewarded with access to a game.

Outcomes

The primary outcomes were test scores on the Progress in English (PiE) 7 (short form) test for the literacy intervention and Progress in Maths (PiM) 7 test for the numeracy intervention. Both tests were provided by GL Assessment. The tests were administered by NFER test administrators to ensure independence from the Oxford research team that delivered the intervention. Test administrators were not informed of the group the school had been allocated to, were advised not to discuss the interventions with teachers, and to conduct the administration as they normally would. Testing took place in classes between 22 April and 6 June 2014, with each school taking both tests on the same day.

The literacy test assessed pupils' ability to comprehend narrative and non-narrative texts and assessed spelling and grammar. The short form of the literacy tests was preferred to the longer form that includes a writing element to make the testing manageable alongside a mathematics assessment, and because the main focus is on pupils' development of ability with grammar, spelling and reading. The mathematics test was an orally administered assessment of pupils' mathematical skills and concepts in relation to number, data handling and shape, space and measures.

We used raw test scores (the number of correct answers) rather than age-standardised scores. We expected within-year age to be evenly distributed across the three groups because the schools were randomised. The primary interest is in average scores across schools, whereas age-standardised scores are useful for comparing pupils of different ages within a school. Processing the scores further would be of limited benefit and introduces ceiling or floor effects. The mean raw score for literacy in the trial was 24.1 and the standard deviation was 8.2, which compares to a mean raw score of 24.1 and standard deviation of 8.0 in the standardisation sample. The mean raw score for numeracy in the trial was 18.3 and the standard deviation was 5.2, which compares to a mean raw score of 19.3 and standard deviation of 5.4 in the standardisation sample.

We also tested some of the secondary hypotheses through a number of secondary outcomes that were identified in the protocol. These included:

- Key Stage 1 mathematics, and reading and writing points: points derived from Key Stage 1 levels in mathematics and average of levels in reading and writing. Because the underlying

levels are assessed by the teacher that participated in the intervention there is a risk that results from analysis using Key Stage 1 points could be biased.

- Sub-domains of Progress in English and Progress in Maths: the percentage of correct answers on the items of the Progress in English test that assess 'Grammar' and 'Reading non-narrative' and of the Progress in Maths test that assess 'Solving Routine Problems'. The sub-domains are groups of items that are pre-specified by the test provider GL assessment. The sub-domains in the protocol were identified by the Oxford research team as those where the interventions were particularly likely to have an effect.

Progress in English 6 (short form) and Progress in Maths 6 were administered by members of the University of Oxford team before randomisation. Test scores from the two pre-tests were used as covariates in analysis to explain attainment variance and increase the statistical power of the analysis.

Sample size

The aim at the outset of the study was to recruit 60 schools and allocate 20 to each group in the trial. Initial estimates of statistical power suggested that such a sample size would be sufficient to detect a standardised effect size of 0.22 with 80% power.¹ This effect size was smaller than might be expected of the programmes according to previous research, but was appropriately conservative given that previous research had been delivered by researchers rather than teachers, had employed tests that were closely aligned to the interventions and were delivered individually or in small groups rather than to whole classes.

The final sample of recruited schools was 55 (17 in the numeracy group, 19 in the literacy group, and 19 in the control group). We used the parameters from the analysed data to compare the initial estimates of the minimum detectable effect size (MDES) to the actual MDES yielded by the data. The number of schools and the number of pupils were both lower than expected, but the intra-cluster correlation was also slightly lower than expected, at 0.09 rather than 0.15. The actual MDES was 0.18 with 80% power.²

Randomisation

Schools were individually randomised to one of the three groups, with an equal allocation of schools to each group. Teachers in the literacy group schools received the *Literacy and Morphemes* training and resources, teachers in the numeracy group schools received the *Mathematics and Reasoning* training and resources; the control group were given the opportunity to receive training in the programme of their choice after the evaluation was completed.

It was intended at the beginning of the project that all schools would be randomised in one block and informed of their group allocation after every school had been tested at baseline. However, this was not possible for a number of reasons, so alterations were made that satisfied schools, while still preventing bias:

- Firstly, the training was the week after the end of the testing period. As schools understandably wanted to know their group allocation in advance of the training so they could arrange supply cover, schools were randomised by an NFER statistician before the testing period, and each school was notified of its group allocation shortly after it was tested. This ensured that neither the school staff nor the pre-test administrator knew the school's group allocation while the baseline testing was being carried out, but allowed schools to make preparations without waiting for every school to be tested.

¹ Assuming a two-tailed test with a 95% confidence level, the expected number of pupils per school = 45, intra-school correlation = 0.15, correlation between pre- and post-test scores = 0.8, and number of intervention and control schools = 20.

² Assuming a two-tailed test with a 95% confidence level, the expected number of pupils per school = 35, intra-school correlation = 0.09, correlation between pre- and post-test scores = 0.81, and number of intervention and control schools = 17(N), 19(L), 19(C).

- Secondly, school recruitment continued until the last week of the testing period. The 51 schools that had been recruited before the pre-testing period were randomised together, but five schools were recruited during the testing period and were randomised separately. The need to inform the school of its group allocation soon after testing meant that the first three were randomised in one block, and the final two in another block a week later.

The 51 schools that had agreed to participate at the beginning of the testing period were allocated, 17 to each group, using simple randomisation and informed of their group after testing (as explained above). One school decided to withdraw entirely from the research project after being randomised but before being tested at baseline. As the school staff did not know which group the school had been allocated to, this decision was made independently, so the drop-out could not have been biased. However, the randomisation outcomes that had been decided were retained meaning there is one less numeracy school because of the drop-out.

The 55 schools were allocated: 17 to the numeracy group, 19 to the literacy group, and 19 to the control group. The school that withdrew had been allocated to the numeracy group, which also happened to be the group not selected in the final randomisation block. Even groups would have been preferable for maximising statistical power, but each block was randomised independently to prevent bias.

The SPSS syntax used to randomise the school groups is shown in Appendix 3.

Analysis

The outcome measures are measured at pupil level, whereas schools were randomised into groups. Therefore, analysis of the difference in outcomes needed to take account of the fact that pupils were clustered within schools. We did this by using a multilevel model that allows the average outcome to be different across schools and has more statistical power than analysis of school-level averages. The multilevel models were estimated in the statistical software package 'R'. Multilevel multiple imputation of missing pre-test and post-test data was carried out using macros developed for software package 'MLwiN'.³

Each multilevel model had the outcome as the dependent variable, and the following covariates were included in every model:

- An indicator of whether the pupil's school was in the literacy group and a separate indicator of whether the pupil's school was in the numeracy group. The excluded group was the control group, so the coefficients of the group indicators measure the difference in (conditional) outcomes between that intervention group and the control group.
- An indicator of whether the pupil's school was one of the five schools that was recruited and randomised late (see 'Randomisation' section). The coefficient is incidental to the research, but controls for any differences in outcome due to underlying factors.
- The pupil's raw score on the pre-test, gender and age in completed months at post-test. For all analysis of literacy outcomes, pre-test score is the score on Progress in English 6, and for all analysis of numeracy outcomes pre-test score is the score on Progress in Maths 6. The coefficients are incidental to the research, but explain a large proportion of outcome variance, increasing the power of the analysis.

Other covariates which may explain additional outcome variance, such as free school meal (FSM) status, were not included in the models because including them would be at the expense of reduced sample size, and therefore statistical power. These variables were obtained from the National Pupil Database (NPD).⁴ The parents of 40 children opted out of having their child's NPD records matched

³ The macros were developed by www.missingdata.org.uk

⁴ The National Pupil Database contains detailed information about pupils in schools and colleges in England, including test and exam results, and background characteristics such as gender, ethnicity and eligibility for free school meals for pupils in the state sector.

with their test score, so including variables from NPD in the analysis reduces the sample size. We believe that the gain in terms of additional explanatory power from including extra variables would not be sufficient to outweigh the loss of pupils from the final analysis, given the covariates that are listed above.

Backwards selection was used to refine the covariates to those variables that explain a significant amount of outcome variance, to reduce overfitting. Backwards selection does mean that each model has a slightly different set of covariates, reducing the comparability across individual models, however pre-test score, which explained around two-thirds of the outcome variance, was included in each model. Both the literacy and numeracy group indicators were always included in final models.

The effect size was calculated as the coefficient on the intervention group indicator (the average difference in outcome between the intervention group and control group) divided by the pupil-level standard deviation. The pupil-level variance is most appropriate for a cluster-randomised trial because the impact of interest is that of the intervention on pupil performance and is the variance estimated for a pupil-randomised trial. Using the pupil-level rather than combined variance means the findings from pupil- and cluster-randomised trials are measured on the same terms. The pupil-level variance is estimated from a separately-run multilevel model with the outcome variable as the dependent variable and no covariates.

The analysis approach was specified in greater detail than in the protocol in light of how the randomisation, data collection and implementation had been carried out, but before the outcome data was received. The statistical analysis plan is detailed in Appendix 4.

Process evaluation methodology

The aims of the process evaluation were: (a) to examine the efficacy and fidelity of the implementation of the interventions, (b) to explore teacher perceptions of outcomes and scalability, and (c) to collect information on relevant activities in control schools. The methods included a questionnaire survey of all participating teachers and five case study visits (involving lesson observations and teacher/pupil interviews). Researchers from NFER also undertook observations of both the literacy and numeracy training days.

Of the five case studies, visits were made to three schools assigned to the numeracy intervention, and to two schools assigned the literacy intervention. The five case study schools were selected to include large and small schools, in urban and rural areas.

We conducted one lesson observation in each selected school; this was followed by teacher and pupil interviews. The aims were to gain a deeper understanding of how the interventions were being implemented, to gather perceptions of impact, and to identify any barriers that might exist for wider roll-out. We also sought views on the effectiveness of the training and guidance materials as preparation for delivery of the intervention, and on whether any improvements to these processes and documents would make a wider roll-out more likely to succeed.

We also used observations and feedback from the school visits to inform the construction of the survey questionnaires that were sent (in the summer term of 2014) to all teachers involved in the two interventions, and to those in the control group. The survey addressed participating teachers' perceptions of the programme. This included their views on training, implementation, pupil engagement and impact.

Impact evaluation

Timeline

Time	Activity
April–November 2013	School recruitment (University of Oxford)
October–November 2013	Pupil data collected from schools (NFER) Randomisation by statistician (NFER) Pre-test administered (University of Oxford) Group allocation revealed to schools
December 2013	Intervention (training sessions) delivered
January–April 2014	Teacher delivery of lessons to pupils Intervention school visits (University of Oxford, NFER process evaluation team)
April–June 2014	Post-testing (NFER test administrators)
June 2014	Training given to control group schools (University of Oxford)

Participants

The strategy for recruitment was to approach primary and infant schools in Oxfordshire and the surrounding counties so as to make the logistics easier for the Oxford-based team. The School Intervention Leader with Oxfordshire County Council supported recruitment by recommending participation to schools with which she worked. Approximately 270 schools across Oxfordshire, Buckinghamshire, Bedfordshire and Berkshire were directly invited to participate in the trial by letter or email.

Schools that had previously worked with the University of Oxford team, some from outside the surrounding counties, were also approached. An advert was placed on the Oxford University Press website, meaning the total number of schools approached indirectly is unknown and potentially quite large.

In total, 56 schools were recruited to participate, with one dropping out before being pre-tested and finding out its randomisation outcome. Of the 55 schools tested at baseline, 23 were in Oxfordshire, 12 were in the surrounding counties of Buckinghamshire, Bedfordshire and Berkshire, and 20 were from elsewhere in England. Recruitment began in April 2013 and continued until November 2013.

The main reason given by schools for not participating was that they wanted to choose the intervention that best fitted with their current school development plans, rather than be randomised. Another reason was that some schools were already involved in other literacy or numeracy research projects with other universities.

Table 1 shows that the characteristics of the 55 schools that participated in the research were broadly representative of all infant or primary schools in England. While lower FSM groups and higher Key Stage 1 groups of schools were slightly better represented among participating schools, chi-square tests found no significant differences for the four variables. However, we cannot be entirely confident that the results would be generalisable to all schools, since schools with a high proportion of pupils eligible for free school meals were underrepresented.

Table 1: Characteristics of participating schools

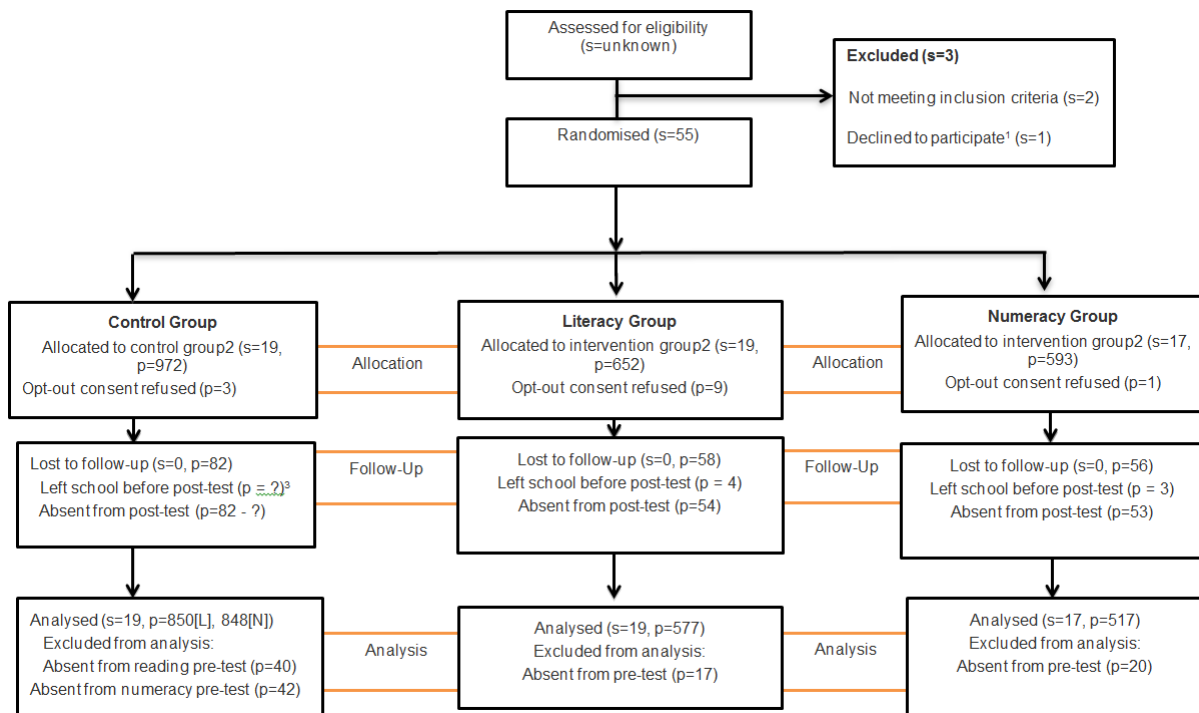
	Number of participating schools	Proportion of participating schools	Proportion of all English infant or primary schools
Free school meals band ($\chi^2=3.29$, $df=4$, $p=0.36$)			
Lowest 20%	15	27%	21%
Second lowest 20%	14	25%	21%
Middle 20%	12	22%	20%
Second highest 20%	9	16%	20%
Highest 20%	5	9%	18%
Key Stage 1 band ($\chi^2=4.18$, $df=4$, $p=0.38$)			
Lowest 20%	9	16%	19%
Second lowest 20%	11	20%	20%
Middle 20%	9	16%	20%
Second highest 20%	17	31%	20%
Highest 20%	9	16%	21%
Ofsted rating ($\chi^2=2.61$, $df=3$, $p=0.46$)			
Outstanding	11	20%	19%
Good	28	51%	56%
Satisfactory	16	29%	22%
Inadequate	0	0%	2%
Urban or rural ($\chi^2=2.59$, $df=7$, $p=0.92$)			
Urban	18	33%	33%
Rural	37	67%	67%

Figures may not sum to 100% due to rounding.

Figure 1 shows the flow of participants through the trial. The total number of eligible schools is unknown because indirect recruitment approaches were used to recruit schools. However, two were excluded after having been recruited: one was an independent school and the other was a special school. One eligible school dropped out of the research prior to being pre-tested and before finding out its group allocation, but as this school had been randomised the randomisation allocations were retained (see 'Randomisation' section). Two class teachers in multi-form entry schools excluded themselves from the trial as they were going on maternity leave during the evaluation period. This was before randomisation and pupils in those classes were not tested.

No schools that were pre-tested were lost to follow-up, so there was no school-level attrition. The pre-tests and post-tests were administered on a single day, which meant there was some pupil attrition due to absence. Some pupils were not tested because they had left the school. In total, 9% of all pupils were lost to follow-up due to absence on the day of testing or from leaving the school. A further 3.5% were excluded from the analysis because of absence from pre-testing. Overall pupil-level attrition was 12%; this was quite evenly distributed between the groups.

Figure 1: Participant flow diagram



Note: 's' indicates the number of schools, and 'p' indicates the number of pupils. The number of pupils sitting the reading and the numeracy post-tests were different: literacy group numbers relate to the reading test, numbers in the numeracy group to the numeracy test. Where they were different, numbers from each test are reported for the control group.

¹ One school dropped out of the trial after being randomised, but before discovering its group allocation. The randomisation outcome was retained, but the drop-out can be regarded as unbiased.

² Based on data uploaded by schools to NFER. Figures provided by the University of Oxford team were slightly different and they likely explain some of the pupil drop-out due to missing pre-test, post-test or both: 2,222 pupils were eligible; 21 (numeracy test) and 24 (literacy test) pupils were excluded from testing because of severe special educational needs; three (numeracy test) and nine (literacy test) pupils were excluded from testing because of English language problems.

³ Data on how many pupils had left the school before pre-test was only collected in intervention schools.

Pupil characteristics

Table 2 shows the average pre-test scores in the analysed groups. We estimated a multilevel model of between-group differences in pre-test score and found effect sizes of 0.08 (95% CI: -0.12–0.27) for literacy and 0.20 (95% CI: -0.03–0.42) for numeracy. Neither was statistically significant at the 5% level and the pre-test is included as a covariate in all analysis models to account for chance imbalance at baseline.

Table 2: Pre-test score

	Control group	Literacy group	Numeracy group
Literacy pre-test raw score			
Mean	23.4	24.0	23.8
Standard deviation	2.8	3.1	2.7
Number of schools	19	19	17
Numeracy pre-test raw score			
Mean	20.5	21.3	21.3
Standard deviation	1.4	1.9	1.5
Number of schools	19	19	17

ANOVA: literacy pre-test raw score ($F=0.44, p=0.65, n=55$), numeracy pre-test raw score ($F=1.70, p=0.19, n=55$)

Table 3 shows the school-level average baseline demographic characteristics for randomised pupils and analysed pupils. The proportion of pupils eligible for free school meals since starting school was highest in the control group. An analysis of variance (ANOVA) of school means shows no significant differences between the three groups in either the pre-test scores or the baseline demographic characteristics. Comparison between the randomised pupils and analysed pupils shows higher levels of FSM, SEN and EAL among randomised pupils, implying that pupils with these characteristics were more likely to drop out because of absence at testing or leaving the school. The extent of differential drop-out is similar across groups, suggesting it is unlikely to threaten internal validity.

Table 3: Baseline demographics

	Control group	Literacy group	Numeracy group
School-level averages (randomised pupils)¹			
Female (%)	47.5	49.5	47.5
Eligible for free school meals since starting school (%)	22.3	14.7	11.2
Special educational needs (%)	17.8	17.5	14.9
English as an additional language (%)	16.5	21.5	14.6
School-level averages (analysed pupils)			
Female (%)	47.9	49.7	46.3
Eligible for free school meals since starting school (%)	21.0	14.0	10.1
Special educational needs (%)	16.4	15.6	12.2
English as an additional language (%)	16.2	21.0	14.1
ANOVA Female ($F=0.52$, $p=0.60$, $n=55$), FSM ($F=2.26$, $p=0.12$, $n=54$), SEN ($F=0.75$, $p=0.48$, $n=54$), EAL ($F=0.45$, $p=0.64$, $n=54$).			
Number of schools	19 ²	19	17

¹ Gender data was collected for all pupils by NFER. The parents of 40 children opted out of having their child's National Pupil Database records matched with their test score, so do not have data relating to FSM, SEN or EAL status. These pupils were included in the primary outcome analysis.

² One school did not supply UPN data for its pupils, so data regarding FSM, SEN and EAL status was not available. The number of schools was 18 for those variables.

Outcomes and analysis

Table 4 summarises the school-level average post-test scores in the analysed groups.

Table 4: Post-test score

	Control group	Literacy group	Numeracy group
Literacy post-test raw score			
Mean	23.8	24.1	25.4
Standard deviation	2.5	3.2	2.4
Number of schools	19	19	17
Numeracy post-test raw score			
Mean	17.8	18.4	19.5
Standard deviation	1.0	2.3	2.3
Number of schools	19	19	17

Table 5 summarises the results of the primary and secondary impact analyses, and Table 6 the results of interaction effects, which explore differential impacts on particular sub-groups. Table 7 summarises the results of on-treatment analyses. A full set of statistical results from the impact analyses is provided in Appendix 5.

Primary outcome analysis

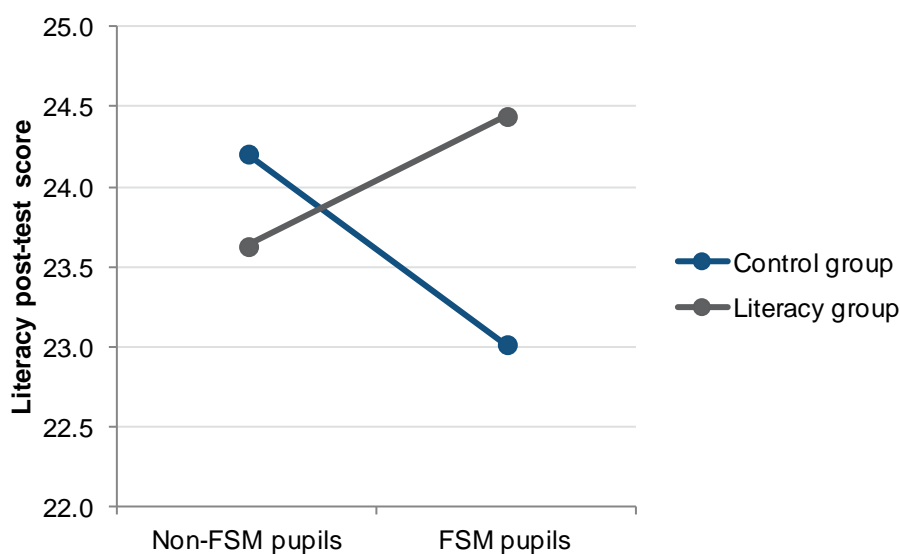
The numeracy programme had a significant positive impact on the primary numeracy outcome, with an **effect size of 0.20 (confidence interval 0.02–0.37)**. In contrast, the literacy programme had no significant impact on the primary literacy outcome, with an **effect size of -0.05 (-0.18–0.08)**. While this gives evidence that the numeracy intervention had an impact on pupils' development of numeracy ability, there is no evidence that the literacy intervention had an impact on the development of literacy ability.

We carried out multilevel multiple imputation to impute missing pre-test and post-test scores to see whether the effect size estimates were sensitive to those pupils' data being excluded from the primary analysis. The missing data analysis results were very similar, suggesting we can be confident that the differences observed were due to the intervention.

Secondary outcome analysis

The interventions did not have a significant impact on the sub-group of FSM pupils. The number of pupils eligible for FSM in each group was small, which limited the precision of this analysis.

We also tested the interaction between group allocation and FSM status: this compares the improvement that FSM pupils make relative to non-FSM pupils in both the intervention and control groups. The interaction between the literacy group and FSM status was positive and significant, which shows that while the literacy intervention showed a negative but insignificant effect for all pupils, the literacy intervention had a differential impact on FSM pupils. Figure 2 shows a graphical illustration of the positive interaction between the literacy group and FSM status. When other factors are held equal, non-FSM pupils, on average, achieved a lower than expected score in the literacy group compared to the control group, whereas FSM pupils, on average, achieved a higher than expected score in the literacy group compared to the control group.

Figure 2: Interaction between literacy group and FSM status

The points shown in the figure are predictions derived from the multilevel model reported in Appendix Table A8. They represent the predicted scores of a male pupil in an early-randomised school with the overall mean pre-test score (23.7 marks).

The interaction between the numeracy group and FSM status was not significant. There were also no significant interaction effects between the impact of the interventions and pupils' prior ability, and no significant interaction effects between the impact of the two interventions and pupils with English as an additional language.

There was a significant impact of the numeracy intervention on the 'solving routine problems' sub-domain of the numeracy test. As the magnitude of the effect was very similar to the overall effect, however, we cannot conclude that the intervention had a particularly strong impact on that sub-domain. The literacy intervention had no significant impacts on the 'grammar' and 'non-narrative reading' sub-domains of the literacy test.

The literacy intervention did not have a significant impact on numeracy scores, and vice versa, showing that the programmes had no significant transfer effects.

Neither the numeracy nor the literacy interventions had a significant impact on Key Stage 1 mathematics, or reading and writing attainment, respectively.

On-treatment analysis

Two on-treatment analyses were conducted to investigate whether greater programme fidelity and engagement were associated with a differential impact. We analysed the number of computer games played by each pupil, and the number of hours teachers spent teaching the intervention in class. Data on the former was gathered from usage data linked to each pupil's unique login, and data on the number of hours teachers spent teaching the intervention in class was derived from responses to the teacher survey.

Both were at the discretion of the pupil and/or teacher rather than randomly assigned, so we cannot confidently attribute a causal impact. However, differences may be indicative of greater impact from more engagement with the interventions. Analysis shows that pupils with a higher pre-test score played more games on average. A high proportion (40–50%) of the variance in the number of games played was between schools, which suggests that the extent to which pupils used the computer games was heavily influenced by whether or not the teacher chose to encourage them to play the games in class or at home. This was confirmed by the process evaluation school visits.

The number of games played by pupils in the numeracy intervention was significantly associated with higher performance on the numeracy test. The effect size was 0.19 (0.10–0.27) per 20 games played, where the interquartile range in the number of games played was 20 games per pupil, meaning the effect size relates to moving from the 25th percentile (9 games) to the 75th percentile (29 games played). The median number of games played was 16 per pupil. Some pupils played a very large number of games, for example five pupils played more than 100 games. Statistical analysis with linear regression can potentially be misleading with such extreme values because it is sensitive to outliers. The 8% of pupils that played more than 50 games were ‘top-coded’—that is, the number of games played was marked as 50.

There was no significant association between more games played and a greater impact of the literacy intervention. In the literacy group, the median number of games played was 10 per pupil; 2% of pupils were top-coded and the interquartile range was 15 games played.

For both interventions, there was also no significant association between more hours spent teaching the programme units in class and more impact of the intervention.

All the outcome measures analysed were pre-specified in the protocol and outlined in more detail in the analysis specification before the final data analysis was conducted. All analysis was carried out on an intention-to-treat basis. The two pieces of analysis that were additional to that specified in the protocol were:

- to include an interaction analysis of FSM pupils, in addition to the sub-group analysis (this analysis is encouraged by EEF); and
- on-treatment analysis of programme fidelity—this is intended to support judgements made about the mechanisms through which the interventions have, or have not, had an impact on pupil attainment.

Table 5: Summary of impact analysis results

Outcome description	Outcome measure	Effect size	95% confidence interval	Number of control group pupils	Number of literacy group pupils	Number of numeracy group pupils	Intra-cluster correlation
Primary	Literacy score (PiE)	-0.05	-0.18–0.08	850	577	513	0.09
Primary	Mathematics score (PiM)	0.20	0.02–0.37	848	577	517	0.11
Primary (FSM)	Literacy score (PiE)	0.10	-0.13–0.34	165	87	49	0.11
Primary (FSM)	Mathematics score (PiM)	0.14	-0.08–0.37	168	87	52	0.03
Secondary	Key Stage 1 reading and writing points	-0.06	-0.24–0.13	824	596	535	0.17
Secondary	Key Stage 1 mathematics points	-0.03	-0.23–0.16	825	601	539	0.15
Secondary	Grammar sub-score (PiE)	0.03	-0.19–0.25	850	577	513	0.13
Secondary	Literacy non-narrative sub-score (PiE)	-0.14	-0.31–0.02	850	577	513	0.07
Secondary	Solving routine problems sub-score (PiM)	0.24	0.08–0.41	848	577	517	0.08
Secondary	Literacy score (PiE): numeracy group	0.08	-0.05–0.22	850	577	513	0.09
Secondary	Mathematics score (PiM): literacy group	-0.02	-0.18–0.15	848	577	517	0.11

Note: Intra-cluster correlation shown is calculated after controlling for pre-test score and other pupil background factors.

Table 6: Summary of interaction effects

Interaction variable	Outcome measure	Interaction p-value	Number of control group pupils	Number of literacy group pupils	Number of numeracy group pupils	Intra-cluster correlation
FSM eligible	Literacy score (PiE)	0.00	765	558	498	0.07
FSM eligible	Mathematics score (PiM)	0.19	764	557	501	0.12
Literacy pre-test score	Literacy score (PiE)	0.43	850	577	513	0.09
Mathematics pre-test score	Mathematics score (PiM)	0.52	848	577	517	0.12
EAL	Literacy score (PiE)	0.84	850	577	513	0.07
EAL	Maths score (PiM)	0.77	848	577	517	0.12

Table 7: Summary of on-treatment analysis results

On-treatment variable	Outcome measure	Effect size	95% confidence interval	Number of control group pupils	Number of literacy group pupils	Number of numeracy group pupils	Intra-cluster correlation
Total number of computer games played	Literacy score (PiE)	0.05 per 15 games played ¹	-0.01–0.11	850	573	513	0.09
Total number of computer games played	Maths score (PiM)	0.19 per 20 games played ²	0.10–0.27	848	577	515	0.12
Total intervention classroom hours	Literacy score (PiE)	-0.005 per hour	-0.016–0.005	850	457	513	0.09
Total intervention classroom hours	Maths score (PiM)	-0.003 per hour	-0.010–0.003	848	577	436	0.14

¹ Fifteen games played was the interquartile range for literacy, so the effect size shows the effect of moving from the 25th percentile to the 75th percentile in terms of the number of games played.

² Twenty games played was the interquartile range for numeracy, so the effect size shows the effect of moving from the 25th percentile to the 75th percentile in terms of the number of games played.

Cost

The *Mathematics and Reasoning* and *Literacy and Morphemes* programmes involved a day's training to introduce the intervention to teachers, explain the concepts, and allow the teachers to explore the learning activities for themselves. Schools were provided with a set of teaching materials and each intervention school was also visited by a member of the University of Oxford team and given support with implementing the intervention.

The training materials were available for schools to access from the University of Oxford's Department of Education although they were restricted for the purpose of this evaluation so that teachers in control schools were unable to access them. The University of Oxford team view the theoretical training that underpins the teaching approaches as an important part of understanding the material in order to deliver it. The programmes may be developed in future so that training is delivered online or through intermediaries.

A form of delivery appropriate for a larger scale is likely to be lower cost, but the cost estimate presented here relates to the cost to schools of implementing the intervention in this trial form, and without EEF funding. Our estimate of the cost of a school participating in one of the programmes includes the cost of providing the training, the cost of a school visit from the University of Oxford team, and other costs associated with implementation.

Participating schools received an EEF bursary of £1,000 plus £250 per each additional class teacher beyond the first two for attending the training, which covered the cost of supply cover, transport and other costs associated with implementing the programme. The following sections show the components of these costs from the perspective of a school implementing the programmes and present the cost per class and per pupil for a typical school.⁵

Cost of providing the training

The estimated cost of training is the one-off cost for a teacher attending the training. Figures provided by the University of Oxford team indicate that the training cost £4,800 to provide, which equates to £112 per class. We do not include the cost of developing the interventions, as this is beyond the scope of the evaluation.

Table 8: Cost of providing the training

Cost	Total cost (£)	Cost per class (£)
Researcher time delivering the training	1,750	42
Printing and posting the teacher packs	1,265	29
Room booking and catering	1,785	42
Total	4,800	112

Source: figures provided by University of Oxford. Figures may not sum to the total due to rounding.

Teachers attended a day of training for each programme during term time, so required one day of supply cover for their class. Where this was not internally covered, schools incurred a financial cost: the average cost of supply cover on the training day (reported by schools in the teacher survey) was £160 for both numeracy and literacy teachers. Attending the training at the University of Oxford involved teachers incurring varying travel costs: we have estimated a typical cost of £40 per teacher, based on the University of Oxford figures relating to school visits (see Table 9).

⁵ There were 43 classes in total across the two intervention groups, and we have assumed an average class size of 25 pupils throughout.

Table 9: Cost to the school of attending the training

Cost	Unit	Cost per class (£)
Supply cover for attending training	1 day per class	-
Transport costs of attending the training	1 staff member per class	40
Total		40

Source: NFER survey of Improving Numeracy and Literacy teachers and figures provided by University of Oxford.

Cost of the school visit

School visits by members of the University of Oxford team—helping teachers to implement the teaching approaches—may be considered part of the investment that schools made. The cost presented below is the cost to an intervention school of one such visit. The University of Oxford team spent time arranging the visits, visiting the schools, and also incurred travel costs.

Table 10: Cost of the school visit

Cost	Unit	Cost per class (£)
Administration time for arranging school visit	Total 10 staff days	28
Researcher time spent visiting typical school	1 day per class	175
Travel costs for visiting a typical school	1 day per class	40
Total		243

Source: figures provided by University of Oxford. Figures may not sum to the total due to rounding.

Some teachers, during case study visits, mentioned that they also required supply cover during the school visit so that time could spend time with the visiting researcher. This, however, was unlikely to be for a long time so has not been formally included in the average cost here, but is an additional consideration for schools.

Cost of implementing the intervention

A significant cost for intervention schools—identified during case study visits—was the large amount of photocopying of teaching resources. Teachers were asked in NFER's teacher survey to estimate this cost, and that of purchasing other items during the training and implementation of the programmes such as number cards, counters, money or dice. Each group reported an average of around £125 for these costs.

The teacher survey found that around half of both literacy and numeracy teachers felt that the units took longer to prepare than their normal literacy or numeracy lessons, while half did not. The case study visits reflected this: some teachers felt that the modular approach actually reduced preparation time. This suggests the interventions do not place a significant additional time burden on teachers.

Table 11: Costs associated with implementing the programmes

Cost	Median cost per class (£)
Photocopying materials and purchasing other resources	125
Additional time planning lessons	Minimal

Source: NFER survey of Improving Numeracy and Literacy teachers, case study visits.

Other resources necessary for successfully implementing the programmes

Schools' fixed resources also affect how aspects of the interventions are best implemented.

The computer games that accompanied the literacy and numeracy interventions were regarded by the University of Oxford team as an important aspect of the programmes. The impact evaluation found evidence that greater use of the accompanying computer games in the numeracy programme was associated with greater impact on numeracy ability. Therefore, access to computers and/or tablets is an important factor in successful implementation and may require investment for a school with limited

IT resources. Pupils' access to computers at home may also be a factor, but is not within a school's control.

The costs of having a teaching assistant during the lessons, and the time to prepare the lesson materials, were mentioned by some teachers. The case study visits found that teaching assistants were most often used to carry out the large amount of necessary photocopying, and to support pupils with logging on to access the computer games.

Total costs of the intervention

Overall, the cost of the *Improving Numeracy and Literacy* programmes to schools is very low. Over the one year of the evaluation, schools spent £520 per class—£21 per pupil—to cover the costs of training, the school visit, and costs associated with implementing the programmes. The cost per pupil is considerably lower than the £131 per-pupil cost of the Mathematics Mastery programme and similar to the £20 per-pupil cost of the Grammar for Writing programme. Both are whole-class interventions aimed at improving the numeracy and literacy skills, respectively, of primary school pupils.

Table 12: Total cost associated with the programmes

Cost	Cost per class (£)	Cost per pupil (£)
One-off cost	395	16
Ongoing per-year cost	125	5
Total cost	520	21

Source: various, see Tables 8–11.

The training and the school visit may be regarded by schools as one-off investment costs that have lasting impact on teacher skills. As these up-front costs do not apply in subsequent years, the average cost per year reduces as the length of the interventions is increased. The only regular cost is the £125 per class for materials. This should be considered when budgeting..

Table 13 illustrates this. The programmes could be expected to have the same impact on pupil attainment every year, so the cost-effectiveness of the programme is higher if it is implemented for longer.

Table 13: Total cost per year associated with the programmes

Number of years using literacy or numeracy programme	Total cost per class (£)	Total cost per class per year (£)	Total cost per pupil per year (£)
1 year	520	520	21
2 years	645	322	13
3 years	770	257	10
4 years	895	224	9

For what length of time might it be appropriate to expect the intervention to be effective? The training is an investment in the teacher, but the teacher might leave the school. Teachers in England complete an average of six years at a school before either moving school, leaving teaching or retiring (Allen, Burgess & Mayo, 2010) meaning that the average teacher is likely to leave a school in three years.. Teachers might forget important aspects of the concepts underpinning the interventions over time, so refresher training every few years may be needed, incurring additional cost. Three years, therefore, seems a reasonable length of time over which to assess the cost of the intervention: the estimated cost over 3 years is £257 per class per year (£10 per pupil per year).

Process evaluation

The aims of the process evaluation were to examine the efficacy and fidelity of the implementation of the two interventions, to explore teacher perceptions of outcomes, and to identify any issues that might have a bearing on a larger scale roll-out. We also collected information on relevant activities in control schools.

NFER researchers observed the training days for both interventions, and carried out in-school session observations of both the numeracy and literacy interventions in practice. Teacher and pupil interviews were also conducted during the visits to five schools, with a view to gaining a deeper understanding of participants' perceptions of the intervention's impact, and to explore any barriers to implementation or scalability.

In addition, we conducted a questionnaire survey of all participating teachers in both intervention groups as well as in the control group at the time of the post-test. The questionnaire took approximately ten minutes to complete. For intervention teachers this included questions relating to the training and support provided, their views on programme implementation, the impact on their classroom practice, and the outcomes for their pupils. For all teachers, including the control group, the survey included questions designed to identify 'business as usual' patterns of teaching numeracy and literacy, and any unusual events or initiatives which might have affected pupil performance in numeracy or literacy within a school.

The evidence from the survey and from the observations and interviews was intended to contribute directly to the scalability evaluation and inform any recommendations for sustainability and replication.

Survey responses are summarised in Table 14. The overall response rate was 82%, which is very high and gives us confidence in the representativeness of the responses. The response rate was so high because the questionnaires were given to teachers and collected again by the test administrators. Due to the small number of teachers in each group we have reported numbers of teacher responses rather than percentages throughout.

Table 14: Teacher survey response rates

	Mathematics and Reasoning	Literacy and Morphemes	Control group
Number of teachers who completed questionnaires	22	21	32
Number of teachers invited to complete a survey	25	24	42
Response rate	88%	88%	76%

The sections that follow summarise the key findings from the combined process evaluation data-gathering exercises.

Training observations (December 2013)

The training for the interventions, *Literacy and Morphemes* and *Mathematics and Reasoning*, each consisted of a one-day course, and both followed the same basic structure. Each day began with an introduction to the intervention, which included background theory and evidence, and was followed by a description of the specific concepts addressed in the programme and an explanation of the structure and detail of the teaching units. The afternoon sessions involved logging on to the online materials, checking usernames and passwords, and provided opportunities for teachers to try out some of the activities related to the literacy or numeracy programmes.

There was no assessment in order to pass the training. Follow-up visits were arranged by the project team to monitor and support implementation of the interventions.

Both interventions, each of 10 to 12 weeks' duration, were to be delivered during the first half of the Spring term of 2014.

The training days were led by Professor Peter Bryant (numeracy) and Professor Terezinha Nunes (literacy)—the creators of the *Mathematics and Reasoning* and *Literacy and Morphemes* programmes—and were supported by other members of the Oxford team. Both training days were held at the Department of Education, University of Oxford. The training rooms used were appropriate for both the introductory sessions and for the online practice. Where teachers were unable to attend, the project team arranged to visit the schools to deliver training in school.

All delegates received a folder containing the relevant resource materials and detailed plans for each of the teaching units (not available separately from the training). Each programme (literacy or numeracy) provided lesson plans for ten units to be delivered in sessions of one hour per week. For both programmes, teachers were required to adhere to the unit plans as presented but, if they could not be completed within the allocated one-hour session, they were asked to complete the unit at the beginning of the next session before moving on to deliver the next unit. No units were to be omitted. It was therefore possible that the full course of delivery could take more than ten weeks to complete.

All units involved some whole-class teaching in the form of a series of PowerPoint presentations and whole-class activities. These were then followed by differentiated activities (as required): Level 1 activities required more teacher support whereas Level 2 activities were designed for more independent learners. The pupil activities for each unit consisted of a series of photocopiable worksheets and booklets. Pupils were to be encouraged to access and play the online games throughout, both in school and at home.

Teachers in both groups were also invited to provide feedback for the University of Oxford team in terms of how the programme worked in their own classrooms, and to collect examples of children's work resulting from the lessons, such as writing scripts and details of discussions/reasoning. The Oxford team also asked for volunteers to collate portfolios of work, to provide feedback on 'surprises' and to make suggestions for improvement.

Teacher engagement during the training appeared to be good, and most delegates were enthusiastic about using the materials and techniques provided.

Teachers' views on the training

Teacher survey

The training sessions for the interventions, held in December 2013, were attended by almost all participating teachers. Only four from the numeracy programme and four from the literacy programme did not attend. This was due to school inspections on the day, for example, or to staff changes, meaning that teachers in the programme had not been in post on the day of the training. Most of these teachers were either represented by colleagues or had the training delivered in school at a later date by a member of the University of Oxford team.

The vast majority of teachers (16 numeracy, 18 literacy) reported that the training covered everything that they needed to know to deliver the programme. Six numeracy teachers and two literacy teachers felt that the training was not necessary.

As well as the practical implementation of the programme, the training outlined the underlying theoretical background and previous research in the area. Only three numeracy and two literacy teachers said they did not find this aspect of the training valuable.

Following the training, 18 numeracy and 18 literacy teachers reported that they were visited by a member of the Oxford University team. Two teachers in each group had had supplementary training initiated by Oxford.

Where teachers had requested support or information from the Oxford team, the topics were: advice on differentiating activities for pupils of different ability; how to replace the classroom presentations which contained errors; and advice on delivering the programme in the time available.

All participating teachers thought that the level of detail was appropriate and believed the follow-up visits to schools from the Oxford University team had been very helpful: *'They explained how we could be flexible with the materials, and were encouraging when we felt like giving up'*. The teachers believed it would be important for every school to have such a visit about two to three weeks into the programme.

Teachers interviewed as part of this trial generally agreed that the training was very useful and essential for understanding some parts of the programme. One teacher felt that the training was good, but that, with so much to cover, a lot of it was not fully explained.

Implementation

Overall, the programme design—lesson plans provided for one hour per week to replace the normal numeracy or literacy lesson—was very straightforward to implement. The necessary conditions for success were that teachers were trained and familiar with the programmes, and that pupils had access to the associated games and, therefore, computers.

The teacher survey asked teachers about their experiences of implementing the two programmes and their views on the programmes' impact and value. As noted, the main concern was the amount of material to be covered in the limited time available. Survey data and case study observations indicated that different schools dealt with this challenge in different ways. Some schools worked through everything methodically, spreading the programmes over more than ten weeks, while others simply decided to cut out some of the intended teaching material in order to make the programmes fit into the time available. Similarly, while some schools encouraged pupils to play the computer games—allowing extra time in school and sending log-on details and passwords home—others allowed only what could be fitted into the lesson time, and some did not provide pupils with log-in information and passwords so that they could play at home if they wished.

On these and other issues, teachers were fairly evenly divided in their views and attitudes, and practice varied between schools. How much pupils used the computer games, which was logged automatically, differed as much between schools as it did between pupils. The findings on programme implementation are presented in more detail below.

The materials

- Almost all participating teachers agreed that the lesson plans and guidance materials were easy to use. Only two numeracy teachers and one literacy teacher felt they were not.
- In terms of preparation time, around half of both literacy and numeracy teachers felt that the units took longer to prepare than their normal numeracy/literacy lessons, while half did not. Some teachers even said that the modular lesson structure to the teaching materials reduced planning time.
- While around two thirds (14) of the numeracy teachers and half (10) of the literacy teachers felt the slides and handouts provided for use in lessons were of high quality, eight teachers in each group thought that they were not.

Many teachers commented on errors in the PowerPoint slides, or mismatches between the PowerPoint presentations and the pupils' worksheets. Although the errors were amended as the

online resources were updated, this was still an issue for some teachers, particularly those who were using the CD-ROM. In the teacher interviews, one teacher said, *'In the end we made a game of "Spot the deliberate mistake"*. One school decided not to send the pupil logins home: *'At first it was because we didn't want them to run too far ahead... but then we decided not to send the materials/passwords home—we have some parents who would pounce on all the errors in the materials'*. In this case, the pupils did not have the opportunity to reinforce their learning by playing the online games at home and were therefore dependent on time being made available in school, which, the teacher admitted, was 'a bit hit and miss'.

- Most teachers thought that the materials were age-appropriate or level-appropriate for their pupils, but five teachers in each group felt that they were not.
- Just over half the teachers felt the differentiated activities in each unit worked well, but just under half the teachers felt they did not (10N; 8L).
- Most teachers (18N; 16L) said they had made use of the extension activities in the units.
- Around half of the numeracy teachers and more than half of the literacy teachers felt that each unit provided learning opportunities for pupils of all abilities, although 11 numeracy and seven literacy teachers did not agree.

Teachers in the numeracy group generally felt the units catered better for less able pupils than for their more able pupils.

Of the 22 numeracy teachers, most (17) felt that the units did not provide work at the appropriate level for the more able pupils, commenting that such pupils were often left waiting or not challenged by units, and one teacher said that this led to disruption from some. Around two thirds (14) of numeracy teachers felt that the units did provide work at the appropriate level for less able pupils, although one commented that without a teaching assistant they would have found it hard to move from one concept to another. One numeracy teacher commented: *'The materials were level-appropriate for approximately half the class. The rest found it too easy'*. Another said: *'Children enjoyed activities, but there was not enough pace, challenge and independent work for the majority'*. However, a different teacher commented that *'Once extension worksheets were provided, they really challenged the high ability pupils'*.

Teachers in the literacy group were more evenly split in their opinions. Of the 21 literacy teachers, 12 felt the units did provide work at the appropriate level for the more able pupils. However, almost half (nine) of the literacy teachers felt the units did not provide work at the appropriate level for the less able pupils. Some teachers commented that they sent their lower ability children out of the class when teaching the later *Literacy and Morphemes* units, and this was observed in two of the case study visits.

Although teachers' perceptions varied, the impact evaluation found no differential effect of either intervention on pupils of different ability levels.

Teachers were asked to indicate the extent to which the *Mathematics and Reasoning* and *Literacy and Morphemes* interventions fitted with what they would normally be teaching in these lessons.

- Numeracy: all but five of the numeracy teachers felt that the content of the units was different from what they would normally teach and almost all (except one) felt that the units supported other aspects of numeracy that their pupils were learning.
- Literacy: 12 teachers said the units covered content that was different from what they would normally teach while nine said it was not different. All but three felt the units supported other aspects of literacy that pupils were learning.

Some teachers made comments—either in the questionnaire survey or during site visits—about specific aspects of the materials. These are outlined below.

Teachers in the numeracy group said:

'It would be better if we spent time on a few good things in isolation, before moving on to the next thing. Some topics have only come up once and I am not sure if that is sufficient time for the children to grasp the idea. If I did it again, I would do more on money alone before diversifying.'

'More thought should be given to extension activities; they should be challenging enough for able mathematicians but should not involve too much reading.'

Regarding the literacy intervention, one teacher commented that there were many positives about the *Literacy and Morphemes* programme. She felt it was very good for reinforcing basic sentence structure, and that it was excellent for pupils with English as an additional language (EAL). This observation was not confirmed by the impact evaluation, however the sample size was small.

Another teacher thought the 'books' were rather 'stilted and old-fashioned,' and not like the books her pupils normally read. Also, in relation to the literacy programme, teachers said:

'The first five [literacy] sessions provided very focused evidence for the "big write" initiative we do across the school, but the second half seemed to take a very big leap and were sometimes a bit random. There were no stepping stones, for example, about root words. It wasn't very coherent and [was] very prescriptive—we are used to being more exploratory—but it was easy enough to work from.'

'I would make sure parts of speech are covered, and use them throughout the week, not just one lesson a week. I would cut down on worksheets.'

Delivery

The process evaluation indicated that teachers found both interventions straightforward to implement. All teachers were able to use the teaching units to deliver the numeracy or literacy content in weekly sessions.

- Most teachers taught the units in the programme to all pupils in the class, as intended. In a few cases younger pupils (for example Year 1 pupils in mixed age classes) or SEN pupils were removed for the session.
- Around half the teachers in each group felt that the range of activities in each session was manageable, while the other half did not.
- The majority of teachers in both groups felt that there was too much content to cover in each lesson (13N; 17L), and 12 teachers in each group agreed that having fewer concepts in each unit would improve the programmes.

One literacy teacher said that, in her opinion, *'...really, there's too much to cover in each session. We could do with less stuff!'* The same teacher reported discussing the lessons with a colleague beforehand and, for some units, they agreed which parts to leave out when they thought the lessons were too long.

There were, however, some positive comments about the numeracy programme, for example: *'The concepts were excellent concepts to focus on—just often too many'*; and *'The reasoning aspect of the programme was its core strength'*.

- The vast majority of teachers felt it was necessary to have a teaching assistant present during the sessions, most or all of the time. Only three numeracy teachers reported that they hardly ever needed a teaching assistant.

Teaching assistants were involved in a lot of preparatory photocopying of the worksheets, but most teachers also reported their value in supporting pupils as they worked through the individual and group activities and the computer games. One teacher said:

'This programme is so reliant on paper and technology. I think it could be done without all the worksheets. There is an awful lot of sitting and writing—we don't usually work like that. It would be impossible without the help of a teaching assistant.'

- Availability of computers was reported as a problem by some teachers, such as having to wait for the next timetabled slot to continue with the tasks.

Only 7 literacy classes and 13 numeracy classes generally did computer work in their classrooms. The others used a computer suite, and a few schools did both. In two schools, children had to be escorted to a computer suite in a different building, and two other teachers reported that teaching assistants were needed to support pupils logging in and using computers.

In one school, the teachers reported that incorporating the computer games into each lesson had posed problems as each class had to book their sessions in the computer suite. This meant that they could not always do the follow-up computer games during, or immediately after, the lessons. They usually arranged follow-up sessions, on another day, to play the games.

Barriers

The findings indicate that the main barriers to implementing the programmes as planned concerned excessive content: teachers felt there was too much content in each unit to fit into one hour a week. Teachers found different ways of dealing with this issue: some cut out some of the content (but not systematically), and a few spent longer than intended delivering the intervention.

Secondly, the developers of the intervention had requested that teachers promote the use of the reinforcing computer games, but the opportunities for pupils to play the games varied considerably. In some schools these were incorporated into the lessons, and in others additional time was allocated in school. Most pupils had the opportunity to log on to the games at home, but some did not.

As reported above, some teachers described difficulties concerning the use of computers due to limited time in the computer suite, or the excessive time needed to get everyone logged on. This appeared to be less of an issue where pupils were using tablets in the classroom.

Other barriers mentioned by some teachers were preparation time, too many worksheets, lack of flexibility, and insufficient differentiation. Other school features that negatively affected implementation included teaching in mixed year-group classes, and not having the support of teaching assistants.

Fidelity

Overall, the interventions were delivered as intended in that all teachers taught one literacy or numeracy lesson per week using the teaching units provided. However, there was considerable variation between schools in the way they dealt with content overload, and the extent to which they supported and encouraged the playing of computer games. These issues are set out in more detail below.

- Fifteen out of 22 numeracy teachers and 18 out of 21 literacy teachers reported that they had completed all the units in the programmes.
- The vast majority of teachers reported that they felt it was not possible to deliver the units in sessions of one hour per week. Only three numeracy and three literacy teachers said they were able to do this.

- The majority of teachers reported that the intervention units had replaced one of their normal numeracy/literacy lessons, as the design of the trial intended, but three numeracy and four literacy teachers reported that they had taught these in addition to their usual lessons.
- On average, most teachers had spent around 10 hours delivering the programme during the term, although the times reported ranged from 9 to 25 hours.

Teachers developed a number of strategies to deal with the issue of excessive lesson content. Some (8N; 7L) began each session from where they had left off in the previous lesson, as advised. However, 11 numeracy and 7 literacy teachers reported that they began a new unit each session, even if they had not completed the previous session. This was a deviation from the training instructions. In other cases, teachers reported that they had deliberately omitted some sections of the unit in order to make sure they did not do more than one hour per week as this had been stressed during the training. Variability of this kind, between schools, was also observed during some lesson observations.

- Opportunities to play the computer games varied considerably between schools.

Over half the teachers in both groups reported that they gave additional time each week to play computer games in school. Only seven numeracy and eight literacy teachers reported that their pupils had played on the computer games at home 'most or all of the time'.

- One third of the numeracy teachers and half of the literacy teachers reported adapting the classroom materials provided for the programme.

For numeracy, this involved using additional or larger objects than those suggested, or allowing/encouraging children to draw rather than use physical materials. These were permissible adaptations. One teacher said she had to provide supplementary materials for the higher ability pupils. Another said, '*The activities were too hard and boring for my lower ability pupils and too easy for my higher ability so I added things here and there*'.

For literacy, teachers reported cutting out some of the content and making some adaptations to the materials for less able and SEN pupils. Removing content from the unit in order to make the lesson fit within a one hour lesson was not a permissible adaptation. Training instructions clearly stated that teachers should work through the materials, completing one unit before starting the next. Some teachers said that they *would* have made adaptations, but did not think this was permitted. During case study visits, one teacher reported that she had felt reassured during the school visit that the materials could be used flexibly and could be adapted to fit better within her normal teaching methods. Others had stuck rigidly to one hour per week delivering the programme and no more.

- In terms of teaching processes, seven numeracy and eight literacy teachers reported 'adaptations to the lesson plans'.

For numeracy, this was mostly to allow for differentiated activities for higher and lower ability pupils or for different ways of working (discussion, pairs or small groups). For literacy, teachers mainly reported adaptations to deliver the content in line with their normal classroom practice, often reducing the amount of 'teacher talk' and encouraging more independent work. These were permissible adaptations and were not, therefore, a threat to the fidelity of programme delivery.

Outcomes

Perceived impact on pupil learning

- All teachers from both groups (with the exception of one) felt that the *Mathematics and Reasoning* and *Literacy and Morphemes* units effectively developed pupils' understanding of the concepts presented.

Seventeen numeracy teachers felt the *Mathematics and Reasoning* programme had helped improve their pupils' numeracy in general, and the same number believed it had improved their reasoning skills in particular. In the literacy group, most teachers (14) thought the *Literacy and Morphemes* programme had helped to improve their pupils' literacy skills and the same number thought it had helped expand their vocabulary. Around half the teachers thought it had helped improve their pupils' spelling, although seven said they did not know.

- The vast majority of teachers (20 N; 19 L) reported that they believed that their pupils benefitted from the discussions they had in each session.

Teachers were asked what other impact they had observed in their pupils.

The majority of comments about the numeracy programme were positive. Three teachers mentioned specific areas of mathematics (counting on, money and division/multiplication using dots to represent numbers), and another that pupils were using more practical methods for working out problems. Two numeracy teachers said that some pupils had shown greater confidence with mathematical problems: *'I particularly noticed that children of all abilities felt comfortable sharing their reasoning with the class'*. One numeracy teacher said that there had been a negative impact: *'When the children are given similar activities, now they won't have a go at them—as they found the numeracy materials so hard to access—especially the lower ability pupils'*.

For literacy, several teachers perceived that their pupils had developed a better understanding of parts of speech and grammatical rules as a result of the programme. One teacher noticed that pupils were correcting each other's oral errors, while another reported that her pupils had a greater enthusiasm for literacy. Two teachers reported specifically that they did not detect any improvement in spelling.

The headteacher in one school, who also taught literacy to Year 2 and had delivered the *Literacy and Morphemes* programme, said: *'It had most impact on my average and more able pupils who were able to get more out of it. Some less able pupils sometimes showed that some of it had gone in but not in a structured way'*. She felt it had probably improved her pupils' literacy and spelling skills, but had no evidence. She felt the programme was good for dyslexic pupils, but believed her lowest achievers felt demoralised by the games because of the amount of reading involved and, therefore, the need for support to get through them.

These observations were not reflected in the impact evaluation findings which found no differential impact on pupils of any ability level.

Impact on classroom practice

- Nine numeracy teachers and six literacy teachers said they had made improvements to their day-to-day practice as a result of the programmes, but the majority felt they had not.
- The majority of teachers said they would retain some elements of the intervention units in their future teaching.

Twelve of the 22 numeracy teachers said they would continue to use the *Mathematics and Reasoning* materials, but 20 said they would use only selected elements in the future. Sixteen said they would recommend the programme to other teachers.

For literacy, 14 teachers said they would continue to use the *Literacy and Morphemes* materials, with 18 of the 21 teachers saying they would use selected elements. Fourteen said they would recommend their use to colleagues.

While the units were largely popular with teachers, there were elements that they would change and adapt if working with them as part of their normal practice. Some teachers had felt they should not make adaptations because of the requirements of participating in the trial, and several teachers commented that the worksheet approach was not how they normally taught.

In one school, the teachers felt the *Literacy and Morphemes* programme had reinforced pupils' understanding of sentence structure and word structure, particularly in lower achievers. Both teachers agreed that pupils enjoyed the lessons and looked forward to the next one. However, one teacher said that the lesson format would not be appropriate for an Ofsted inspection: '*They don't want us to teach like that.*'

- No teachers reported any negative effects or unintended consequences, although some teachers felt the structure of the units meant that too much was delivered in the first teacher-led whole-class activity, which meant all children had to work at the pace of the slowest pupil.

Pupil engagement

- The majority of teachers (19 N; 16 L) reported that their pupils found the units stimulating and enjoyable.

This was confirmed during the case study visits. In pupil interviews, all generally agreed that they had enjoyed working through the programme units and some reported that *Literacy and Morphemes* had made them think about words in a different way. One boy said '*I had never thought about those kind of rules before—but once you know them it all makes sense!*'. Two numeracy teachers reported that children enjoyed the variety and that repetition of activities across units allowed children to become familiar with the concepts. One teacher commented: '*Some sessions did have a vast range of activities for the children but most were manageable in our extended session time.*'

- Some teachers (7N; 11L) thought that the unit plans provided more opportunities for class discussion than in their normal lessons, but large proportions in each group disagreed.

One numeracy teacher commented, '*The materials were often very worksheet based—they did not encourage discussion and were very closed tasks if used as set.*'

The issue of numerous worksheets was also raised during the school visits where a number of teachers reported that using so many worksheets did not reflect their normal classroom practice. Some, however, reported that (sometimes to their surprise) pupils appeared to enjoy the worksheet tasks and looked forward to the 'project' sessions.

Other factors

Teachers were asked if anything had happened since the pre-test that might affect their pupils' numeracy or literacy skills, for example, starting a new numeracy or literacy scheme in addition to the Oxford University programme.

Four numeracy teachers reported that they had started to use a new numeracy scheme. Two named Abacus (Active Learn), one Numicon and one mentioned 'a new maths framework and objectives'. One teacher reported there was an ongoing focus on improving mathematics across the school.

Eight literacy teachers reported that, since the pre-test, new literacy schemes or approaches had been introduced in the classroom. Two teachers reported that a new literacy scheme had been introduced and six teachers reported that something had happened since the pre-test that might impact on their pupils' literacy skills. These included Talk for Writing (Pie Corbett), oral storytelling, the introduction of weekly spelling tests, and a new teacher taking over the class.

Further details are reported in the control group comparisons section below.

Formative findings

Teachers were generally positive about the content of the units. In an attempt to ensure consistency of implementation, the trialling design required teachers to restrict their delivery of the intervention to one hour per week. In fact, because of the quantity of material provided, most teachers were unable to complete the units in the time and different solutions were adopted. This actually led to more inconsistency in terms of what content was left out or whether extra time was spent, and some teachers found it difficult to fit the computer game practice into the teaching hour allotted. We recommend that, in future, teachers should be allowed to deliver the units in whatever way is most suitable for their own particular situations—in other words, one unit per week, without the time restriction. The key elements should be teacher familiarity with the understanding and skills promoted by the programme, and ensuring that pupils have the opportunity to practise those skills.

In the questionnaire and interviews, numeracy and literacy teachers were asked for suggestions for improvements to the materials. Those raised by a number of teachers are summarised below.

- The main suggestion for both numeracy and literacy programmes was to reduce the number of worksheets and the amount of photocopying.
- Some literacy teachers suggested that it would be better to provide workbooks in future, and that higher quality texts and stories should be included.
- Teachers in both groups felt that the whole-class activities should be reduced and the amount of active learning increased, for example, using more games, investigations, practical activities, group and paired work.
- Many teachers reported that the units did not reflect their normal teaching practice; in particular, the use of worksheets is no longer common in primary classrooms.
- Some teachers felt that a way of monitoring and controlling pupils' use of the computer activities, and feedback from them, would be very helpful.

A few specific suggestions raised by individual teachers were: to cover only one new concept per session; to improve the quality of the computer activities; and to break up activities and make them shorter.

Scalability

A further aim of the process evaluation was to consider the extent to which the interventions could be scaled up and rolled out to a large number of schools.

Essentially, both interventions involved one day of teacher training and a support visit from the delivery team. Responses from the teacher questionnaire showed that the training was considered valuable, and teacher interviews suggested that the school visit (two to three weeks in) had proved helpful. While the school support visit may be feasible for an effectiveness trial, it may not be sustainable for national roll-out.

The delivery team should therefore consider how the training and ongoing support could be delivered in future. If face-to-face training remains the preferred method then more trainers, with sufficient experience and understanding of the interventions, would need to be available.

The University of Oxford delivery team have indicated that they would be happy to make the *Improving Numeracy and Literacy* materials and games freely available online. It is not clear, at this stage, whether there would be a charge for the training and support sessions if the face-to-face sessions were to be maintained. Case study interviews suggested that teachers valued a question and answer type of conversation with the delivery team after they had delivered two or three sessions, with teachers reporting that they had welcomed the opportunity for discussion afforded by the school visits. They were able to check specific aspects of their approach to the lessons, and were reassured that it was permissible to work flexibly with the materials. In some cases, it was the feedback from the visiting University of Oxford researcher that had proved most useful, after they had observed the teacher delivering one of the teaching units. If this practice is to be maintained it would add to the scalability costs.

There are a number of ways in which cost savings could be realised by adapting the model of training and support used during this trial:

- The introduction to the principles and theory could be delivered by an online training module using video and online practice.
- A set of 'frequently asked questions' (FAQs) could be prepared and made available on the website.
- E-mail support or a telephone helpline could be offered via the website (especially two to three weeks into the programme).
- Webinar support sessions could be offered at regular intervals.
- Schools could be offered the opportunity to buy in-school training for staff, or for cluster groups.

When teachers are familiar with the principles and processes of the interventions, the programmes can be implemented using the downloadable unit materials as provided.

If a way can be found to maintain the type of support provided through the school visit, we do not envisage any major issues associated with rolling out these interventions at scale. If face-to-face training and school support visits are considered essential, this would have substantial implications for scalability.

Control group activity

Control group comparisons

Teachers in the control group were asked questions that were also included in questionnaires for the intervention teachers. The results for the three groups are shown in the tables that follow.

Table 15: Changes introduced since pre-tests

	Control group (N=32)		Numeracy group (N=22)		Literacy group (N=21)	
New literacy scheme	4	13%	1	5%	2	10%
New numeracy scheme	3	9%	4	18%	2	10%
Other change that may affect literacy	5	16%	3	14%	6	29%
Other change that may affect numeracy	6	19%	1	5%	3	14%

Note: new literacy/numeracy schemes do not include the Improving Literacy and Numeracy programmes.

The new schemes and approaches reported by the control group for numeracy included Abacus Maths, Connections Method, and Numicon. Other changes that may have affected pupils' numeracy skills included specific in-school focuses on developing mental skills and visualisation, problem solving and ability grouping.

For literacy, newly introduced schemes included Talk for Writing and Story Telling (Pie Corbett), Big Write (Ros Wilson), and updated Hamilton Trust materials (for the new National Curriculum). Some teachers reported other changes such as an in-school focus to improve writing/literacy, revised phonics, and ability grouping which may have impacted on pupils' literacy skills.

On average, teachers in the control group spent 5 hours teaching numeracy and 5.5 hours teaching literacy each week. The intervention groups spent slightly more time, as shown in Table 16.

Table 16: Numeracy and literacy teaching time

	Control group (N=32)	Numeracy group (N=22)	Literacy group (N=21)
Numeracy teaching time (hours)	5.0	6.5	5.5
Literacy teaching time (hours)	5.5	6.5	7.0

In each group, the average time spent by pupils on computer-based literacy and numeracy tasks each week was around 30 minutes for each subject.

Teachers were asked whether their pupils were familiar with taking tests at school, whether the pre-test and post-test had been carried out under the same conditions, and if they believed the tests would give a true reflection of the ability of their pupils. Table 17 shows the results. The majority of teachers in each group felt that the tests did not give a true reflection of their pupils' ability. This proportion was largest among teachers in the control group, who also reported that their pupils were less familiar with test-taking than did teachers in the intervention groups. This may reflect the fact that, in some schools, assessment policy is specifically designed to minimise summative testing with instead more focus on formative approaches. Teachers were also asked if the Progress in English or Progress in Maths tests used for evaluation were used in their schools; only one teacher (in the numeracy group) reported that they were.

Table 17 - Use of tests in schools

	Control (N=32)	group	Numeracy (N=22)	group	Literacy (N=21)	group
Familiar with test taking	21	66%	18	82%	15	71%
Pre- and post-tests carried out in the same conditions	24	75%	18	82%	15	71%
Do test results give a true reflection of their pupils' ability?	9	28%	10	45%	10	45%

Conclusion

Key Conclusions

1. This evaluation provided evidence that the *Mathematics and Reasoning* programme had a positive impact on pupils' numeracy ability equating to three additional months' progress.
2. There was no evidence to suggest that the *Literacy and Morphemes* programme had an impact on pupils' literacy ability overall.
3. There was an association between greater use of the accompanying computer games and greater impact in the numeracy intervention, suggesting the computer games were important to successful implementation.
4. All teachers were able to implement the programmes, but most agreed there was too much content to deliver in one hour per week and so made various adaptations to their delivery of the programme. In future trials of the programmes, teachers should be permitted to use and integrate the materials in their own way, as they would in a normal teaching situation.
5. A future trial could evaluate the programmes at scale in more than one location. When drawing up plans for bringing the programmes to scale, the Oxford team should consider whether training and ongoing technical support could be delivered remotely, rather than in person.

Limitations

As there were no schools—and only a small proportion of pupils—that were lost to follow-up, and the trial protocol was adhered to, there is little threat to the internal validity of the trial. Therefore, we can be confident that the observed differences between the literacy and numeracy groups and the control group are evidence of the impact the programmes had on pupils' development of literacy and numeracy ability.

The schools that participated in the trial are broadly representative of schools nationally, based on their observed characteristics such as FSM eligibility, attainment, Ofsted rating, and urban or rural setting, suggesting that the results are generalisable to schools with Year 2 pupils. However, the schools that opted in to the trial may still be fundamentally different if their decision to participate in the trial was influenced by unobserved traits, for example engagement with research. We cannot be entirely confident that the results would be generalisable to all schools, since schools with a high proportion of pupils eligible for free school meals were under-represented. We are not confident that the results are generalisable to independent or special schools as they were not eligible to participate in the trial.

Interpretation

This study was designed to address the following research questions:

- What is the impact of *Mathematics and Reasoning* on pupil development of numeracy ability?
- What is the impact of *Literacy and Morphemes* on pupil development of literacy ability?

This independent evaluation was designed to ascertain whether pupils randomly assignment to either of the two intervention groups would achieve gains in numeracy or literacy scores that were significantly greater than the gains made by pupils who had been assigned to the control group.

Numeracy

The impact analysis found that pupils in the numeracy group made greater progress in numeracy than the control group.

Teacher feedback on the numeracy programme was generally positive, and many teachers indicated that they believed the programme's focus on developing pupils' reasoning skills, and the opportunities for discussion that the teaching units provided, supported their pupils' understanding and their ability to use and manipulate numbers and mathematical concepts. These teacher perceptions are consistent with the findings that pupils demonstrated a significant improvement in their numeracy test scores compared with pupils in the control group.

The *Mathematics and Reasoning* programme specifically requires pupils to discuss, explore, reflect on, and explain the logical relationships in a range of numerical operations in order to develop and practise their mathematical reasoning skills. Two strands—number sense and quantitative reasoning—explore concepts such as additive composition and place value representation, inverse relations and part-whole (additive) reasoning and one-to-many correspondence (multiplicative) reasoning. Pupils are encouraged to use objects (rather than symbols) and to unpick and 'play with' mathematical ideas so that they develop an understanding of what addition, subtraction, multiplication, division, and so on, actually mean and how they relate to real life scenarios—as opposed to the rote learning of rules.

The impact of the *Mathematics and Reasoning* programme is reflected in the fact that pupils in this trial showed significantly greater progress in the sub-domain of 'solving routine problems'. These elements of the assessment specifically require reasoning skills, and the results of the impact evaluation suggest that the programme was effective in supporting the development of these reasoning and problem solving skills in the participating pupils.

This progress, however, was not reflected in Key Stage 1 mathematics points. This could be for a couple of reasons. First, the Key Stage 1 results are based on teacher assessment. Teachers are required to build a picture of what children can do from their observations in every day teaching and learning, use a test or task to support their judgment, and finally make an end of Key Stage level judgment based on a range of information. Whereas the primary outcome measure was administered by an independent test administrator, Key Stage 1 was assessed by the teacher who had participated in the intervention, so could be biased. However, we might have expected any bias from non-blinding to be in favour of finding an effect, whereas no effect was found.

Second, the teacher judgements are made on the four National Curriculum attainment targets: *using and applying mathematics*; *number, shape, space and measures*; and *handling data*. Teachers are advised to give greatest emphasis to a child's performance in *number* and *handling data*—these account for more than half of the overall performance in mathematics. The skills developed by the *Mathematics and Reasoning* programme would, most likely, be demonstrated in the *using and applying* elements of numeracy learning which only accounts for one fifth of the overall teacher assessment. It is possible that the fine distinctions in skills development detected by the Progress in Maths test were not reflected in this broad end of key stage judgement.

The effect size calculated in this independent evaluation (0.2) is considerably lower than the effect size found in the previous randomised intervention study by the University of Oxford research team (effect size 1.2). There are three main reasons why this may be the case. First, the programmes in the previous intervention studies were delivered to individual pupils on a one to one basis, and delivered by a member of the research team who was familiar with the programme. In contrast, all teachers in this trial delivered the intervention to a whole class at a time, and most were unfamiliar with the programme before training for trial delivery. It is not surprising, therefore, that the effectiveness might be diluted.

Second, the tests in multiplicative reasoning used in previous research were closely aligned to the concepts taught through the programme; hence larger effect sizes might be expected. The tests were also administered by the researchers so might have suffered from bias because the testing was not blind to group allocation.

Third, another factor likely to contribute to the higher impact detected in previous research was that the programme was delivered to pupils of a different age group who had little or no previous experience of multiplication or division (only addition and subtraction). Therefore, all the learning from the programme was new to them and a greater impact might be expected.

Literacy

This trial has found no evidence that *Literacy and Morphemes* produced significant improvements in literacy ability over and above those in the control group. In addition, no significant differences were found in the Key Stage 1 reading and writing results.

The *Literacy and Morphemes* programme focuses on helping pupils to understand how the meanings of words can be constructed and altered by manipulating the smallest units of meaning—morphemes. This approach develops pupils' understanding of the grammatical and semantic aspects of word reading as opposed to the phonics-based approach that focuses on how the sounds of written letters can be blended and segmented to support decoding and spelling. Phonics is universally used to teach reading in early learning settings in England and forms part of the statutory assessment programme. Morphemic awareness training develops quite different, but complementary, language and literacy skills, and the *Literacy and Morphemes* programme can be viewed as providing an additional tool to support pupil's literacy learning and understanding.

The purpose of this study was to evaluate the impact of the *Literacy and Morphemes* programme on pupils' overall literacy performance when delivered by the class teacher to the whole class. In previous intervention studies conducted by members of the University of Oxford team, the interventions were delivered by researchers, who were very familiar with the programme, to small groups of pupils over 12 weeks. The assessments used in the previous randomised intervention studies were more closely aligned to the content of the morphological training lessons, with more focus on spelling, than the more general literacy test used for this impact evaluation, and for Key Stage 1 teacher assessment for reading and writing.

Pupils who are eligible for free school meals made slightly more progress if they participated in the literacy programme, but we are not able to conclude that this observed effect was caused by the programme rather than occurring by chance. However, comparing the relative progress of FSM pupils in the intervention and control groups and non-FSM pupils in the intervention and control groups, we can see a significant positive effect. While the literacy programme had no overall impact on the development of literacy ability, there is evidence that the literacy programme had a differential positive impact on FSM pupils. This might suggest that the programme would be more effective if it was targeted at FSM pupils rather than delivered to the whole class.

The process of reading, and of comprehending, requires receptive language skills whereas writing (and speaking) use expressive language skills. It seems likely that *Literacy and Morphemes* may be more effective in supporting the latter (writing and spelling skills), but the assessment used to measure the impact of the programme focused mainly on reading comprehension. Sub-domains of non-narrative reading and grammar were examined more closely and showed no significant differences from the control group. It is possible that future research into the effects of the programme on pupil writing might produce more favourable results.

In terms of the teaching approach advocated by the *Literacy and Morphemes* programme, a number of teachers stressed that the use of PowerPoint and individual worksheets did not reflect the prevailing practice in primary classrooms which is generally more discursive and collaborative. Some felt the units promoted outdated teaching methods. If teachers were not fully engaged with the teaching and learning style embodied in the programme, the effectiveness of their delivery may have been impaired.

As with the numeracy programme, further observations on the method of delivery and pupil learning would enhance our understanding of which aspects of the programme are more or less effective in improving performance. It would also be useful to examine more closely the development of the writing skills of the intervention group and follow up the long term impact of the programme.

Computer games

The opportunity to practise the skills taught in class by playing relevant computer games is an important and integral part of both of the *Improving Numeracy and Literacy* programmes.

As described in the process evaluation—and from the responses to the questionnaire—there would appear to be some variance in how the programmes were delivered in different schools and, particularly, the extent to which pupils played the supporting computer games. The impact analysis did find a positive association between the number of numeracy games played and impact, although it is not clear whether these differences had a causal impact on how pupils developed their skills in numeracy or literacy.

Future research and publications

While there is evidence that the *Mathematics and Reasoning* programme had a significant impact on pupil development of numeracy ability, it is questionable whether the programme would have the same impact if it was scaled up to a larger number of schools. Given that the p-value of the effect size is close to the conventional threshold of 0.05, the case for whether an effectiveness trial is warranted should be considered very carefully, given that the likelihood of the result being a false positive is fairly high (Colquhoun, 2014).

The delivery model used a training day delivered by the researchers and a school visit to help with implementation, which may not be feasible using the existing team at scale. An effectiveness trial of the teaching materials using a delivery form that is more appropriate for a large number of schools would determine whether there is an impact without direct support from the research team. If an effectiveness trial were to take place in the future, we would recommend that detailed classroom observations were conducted, in both intervention and control schools, to explore the extent to which learning gains are associated (simply) with delivery of the programme, or whether other factors—such as the method of delivery, teaching style adopted, or the amount of discussion or student engagement—are more strongly associated with improved performance. Further observational assessment and exploration of the pupils' understanding and application of the concepts taught through the programme would enhance our understanding of how and why the programme can improve their performance in mathematics overall.

Long-term follow-up of the participants in this trial to Key Stage 2 pupils would provide a test of whether the impact of the numeracy intervention has been sustained. However, this should be attempted cautiously as the interventions may have been used with control group pupils because of the waitlist design. Teachers in the control group were given training in June 2014, opening the possibility that the resources were used in classes with control group pupils in July 2014, used remedially on those pupils in the next academic year, or of wider knowledge-sharing within the school. Also, some teachers in literacy group schools had training in the numeracy programme (and vice versa) at the same time as the control group schools.

References

- Allen, R., Burgess, S. and Mayo, J. (2012) 'The teacher labour market, teacher turnover and disadvantaged schools: new evidence for England', *CMPO Working paper*, 12/294.
- Bradley, L. and Bryant, P.E. (1983) 'Categorising sounds and learning to read – A causal connection', *Nature*, 301, 419–421.
- Bryant, P., Nunes, T. and Barros, R. (2014) 'The connection between children's knowledge and the use of grapho-phonetic and morphemic units in written text and their learning at school', *British Journal of Educational Psychology*, 84, 211–225.
- Colquhoun D. (2014) 'An investigation of the false discovery rate and the misinterpretation of p-values', *R. Soc. open sci.* 1: 140216. <http://dx.doi.org/10.1098/rsos.140216>
- Nunes, T., Bryant, P. and Olsen, J. (2003) 'Learning Morphological and Phonological Spelling Rules: An Intervention Study', *Scientific Studies of Reading*, 7(3): 289–307.
doi: 10.1207/S1532799XSSR0703_6
- Nunes, T., Bryant, P., Evans, D., Bell, D., Gardner, S., Gardner, A. and Carraher, J. (2007) 'The contribution of logical reasoning to the learning of mathematics in primary school', *British Journal of Developmental Psychology*, 25: 147–166. doi: 10.1348/026151006X153127
- Nunes, T., Bryant, P., Burman, D., Bell, D., Evans, D. and Hallett, D. (2009) 'Deaf children's informal knowledge of multiplicative reasoning', *Journal of Deaf Studies and Deaf Education*, 14(2): 260–77.
- Nunes, T., Bryant, P., Burman, D., Evans, D., and Bell, D. (2010) 'The scheme of correspondence and its role in children's mathematics', *British Journal of Educational Psychology: Monograph Series, II Understanding Number Development and Difficulties*, 7, 83–99.
- Nunes, T., Bryant, P. and Barros, R. (2012) 'The development of word recognition and its significance for later reading skills', *Journal of Educational Psychology*, Online First Publication, March 19, 2012.
doi: 10.1037/a0027412
- Nunes, T., Bryant, P., Barros, R. and Sylva, K. (2012) 'The relative importance of two different mathematical abilities to mathematical achievement', *British Journal of Educational Psychology*, 82, 136–156.
- Torgerson, D., Torgerson, C., Mitchell, N., Buckley, H., Ainsworth, H., Heaps, C. and Jefferson, L. (2014) 'Grammar for Writing: Evaluation Report and Executive Summary', Education Endowment Foundation.
- Vignoles, A., Jerrim, J. and Cowan, R. (2015) 'Mathematics Mastery: Primary Evaluation Report', Education Endowment Foundation.

Appendix 1: Covering letter to headteacher

Independent Evaluation of Two Initiatives to Improve

Literacy and Mathematics Achievement in KS1

Dear [headteacher],

We are very pleased and grateful that your school will take part in our research study “An independent evaluation of two initiatives to improve Literacy and Mathematics Achievement in KS1”. We are writing to you now to tell you more about our plans for the project.

As you already know, two teams of researchers are involved in carrying out the research. One of these teams, which is based in the Department of Education at Oxford University, has designed the teaching programmes that are to be investigated and will provide the material and the advice needed to carry out the teaching. The Oxford team will also carry out tests (pre-tests) of the children’s literacy and mathematical skills at the start of the project. The other team who work at the National Foundation for Educational Research (NFER) and at GL Assessment will be responsible for evaluating the effectiveness of the teaching programmes. This team will test the children’s reading and mathematical abilities at the end of the project (post-tests) to find out whether these are affected by the teaching programmes.

Data sharing¹

We would like to assure you that all the information that we gather about individual pupils, teachers and schools will be kept completely confidential in accordance with the Data Protection Act. No information about individual children will be made available to anyone outside of GL Assessment or the research teams within the University of Oxford Department of Education, NFER, Education Endowment Foundation (who fund the work) and the UK Data Archive. We will not use pupils’ or teachers’ names or the names of any of the schools involved in the project when we write about the results of the research. Our accounts of the effectiveness of the teaching programmes will be presented in the form of aggregated or averaged data.

The NFER team also plans to make a request for further information about the children in the project that is held in the National Pupil Database (NPD). The team will make a request for Unique Pupil Numbers (UPNs) which will make it possible for them to access data about individual children in the NPD. As soon as this additional information is included in the project’s data set, the data will be anonymised and no one will be able to identify individual children in our data files.

Process evaluation

As part of its evaluation, the NFER team would very much like one of its members to visit some schools to attend and observe one of the teaching sessions if that is possible, and your school could be randomly selected for a visit. The NFER group would also like to conduct a survey of all the teachers involved in the project. When they come to your school to carry out the post-tests, they will bring a copy of the survey and the teachers would be able to fill this in while the NFER assessor is working with the class. It would also be of great value if they could interview over the telephone a sample of head teachers of the schools involved, and perhaps other SMT members who might know about the programme.

We do hope that you will be happy about these plans. If you have any concerns about them, please contact us. You can telephone or email Professor Terezinha Nunes (01865 284892:

terezinha.nunes@education.ox.ac.uk) or Deborah Evans (01865 284893: deborah.evans@education.ox.ac.uk).

Once again, we would like to express our gratitude to you for joining us in this research. We believe that it will be a thoroughly worthwhile project and that it will produce some valuable results.

[someone from Oxford team]

1 Applies to pupils whose parents have not opted out of the study. There are two levels of opting out; parents can either opt for their child(ren) not to take part at all, or they can opt for their child(ren) to take part in the study but not for the research team to have access to their UPN. It should be noted that all pupils also have the right to withdraw themselves from the study at any time or opt for their UPN not to be shared with the research team.

Appendix 2: Opt-out consent letter to parents

Dear Parent,

I am writing to inform you of a project that aims to assess materials designed to improve children's attainment in literacy and mathematics in Key Stage 1. These materials were found effective in previous studies with a small number of children. In this study, a large number of children are participating in order to see whether the materials can help children who have different levels of attainment at the start. The project is funded by the Education Endowment Foundation (EEF). Your child's school has kindly agreed to cooperate with the project and now we are informing you of the study and giving you the opportunity to withdraw your child if you so wish.

In line with the EEF guidelines, all the schools will have the opportunity to use the materials; some schools will use them during this school year and others will use them in the next academic year. The assignment to this or next year will be done randomly by the independent evaluators for both the literacy and mathematics materials. There are no expenses to be incurred from participation.

In order to evaluate the programmes, all children will complete some classroom based assessments at the beginning and the end of the academic year, which take in total no more than two hours. These assessments do not influence your child's placement in school. They are necessary only for the research. Pupil data and test responses will be collected by GL Assessment and accessed by NFER. No information about individual children will be made available to anyone outside the research teams within the different bodies listed below. The data will be kept confidential, in accordance with the Data Protection Act. We will only publicise aggregated results of the programme evaluation. We will not use your child's name or the name of the school in any report arising from the research.

The programme involves a series of activities, which the teachers will be using in the classroom, as part of teaching literacy or mathematics. The children will not be taken out of the classroom if they are participating in the programme and will not be missing any lessons.

The research will be carried out by staff from the Department of Education, University of Oxford, in collaboration with an independent evaluator team from the National Foundation for Educational Research (NFER). The ethics committee of Oxford University has approved this study and it falls within NFER's Code of Practice.

We will also obtain your child's UPN (Unique Pupil Number) to complement the assessment of the programmes. This will allow the assessors to link test results to the NPD (National Pupil Database) and share data with the EEF and UK data archive for research purposes. Once this information is included in the data set, the data will be anonymised and no one will be able to identify individual children.

Your child can still participate in the learning activities if you do not consent for this information to be released, so we have two separate boxes in the consent form. If you have any questions you would like to ask before replying, please do not hesitate to contact one of the researchers by phone on (01865) 284 893 or by emailing Professor Terezinha Nunes (terezinha.nunes@education.ox.ac.uk). We expect that your child will enjoy doing the tests and being part of the programme. You may withdraw your child from the project at any time by advising the teacher or the researchers of your decision.

If you DO NOT wish your child to participate in the project and/or for the research team to have access to your child's UPN, please complete and return the attached form to your child's class teacher by *(date to be filled in depending on date school joins)*.

Professor Terezinha Nunes

Department of Education

University of Oxford

Title of Project: An evaluation of a programme designed to improve literacy and mathematics achievement in Key Stage 1

If you DO NOT wish your child to participate in this project, return this form to your child's class teacher.

Please note that your child can still participate in the teaching programme even if you do not wish to release the UPN. In this case, tick only the second box.

Child's name:Date of birth:

Child's class Teacher:

School:.....

I DO NOT wish my child to participate in the teaching activities.

I DO NOT consent to my child's Unique Pupil Number to be released to the research team.

Parent name (BLOCK CAPITALS)

Parent signature:

Date

Appendix 3: Randomisation syntax

Randomisation syntax—Block 1

Title 'Randomisation for Oxford trial (EEOL).'

subtitle 'Block 1 - 22nd October 2013'.

```
GET DATA /TYPE=XLSX
```

```
  /FILE="\milesan1\projects\EEOL\Randomisation\Schools in sample 22.10.13 EXCEL.xlsx'.
```

*Check for duplicates.

```
sort cases by nfer_no.
```

```
match files file=*/first=f/last=l/by nfer_no.
```

```
cross f by l.
```

```
temp.
```

```
select if any(0, f, l).
```

```
list vars=nfer_no description post_code.
```

*Randomise pupils.

```
set rng=mt, mtindex=22102013.
```

```
compute random=rv.uniform(0,1).
```

```
sort cases by random.
```

*Now allocate schools into groups on the basis of where they appear in the randomised order.

*If the number of schools is odd, decide which group the remainder will fall into before randomisation.

```
if $casenum le 17 Group=1.
```

```
if $casenum gt 17 AND $casenum le 34 Group=2.
```

```
if $casenum gt 34 Group=3.
```

```
ADD VALUE LABELS Group 1 'Literacy intervention' 2 'Numeracy intervention' 3 'Control'.
```

```
freq group.
```

Randomisation syntax—Block 2

```
GET DATA /TYPE=XLSX
```

```
  /FILE="\milesan1\projects\EEOL\Randomisation\Schools in sample Block 2.xlsx'.
```

*Check for duplicates.

```
sort cases by nfer_no.
```

```
match files file=*/first=f/last=l/by nfer_no.
```

```
cross f by l.
```

```
temp.
```

```
select if any(0, f, l).
```

```
list vars=nfer_no Schoolname.
```

*Randomise pupils.

```
set rng=mt, mtindex=20112013.
```

```
compute random=rv.uniform(0,1).
```

```
sort cases by random.
```

*Now allocate schools into groups on the basis of where they appear in the randomised order.

*If the number of schools is odd, decide which group the remainder will fall into before randomisation.

```
if $casenum eq 1 Group=1.
```

```
if $casenum eq 2 Group=2.
```

```
if $casenum eq 3 Group=3.
```

```
ADD VALUE LABELS Group 1 'Literacy intervention' 2 'Numeracy intervention' 3 'Control'.
```

```
freq group.
```

Randomisation syntax—Block 3

Title 'Randomisation for Oxford trial (EEOL).'

subtitle 'Block 3 - 29th November 2013'.

```
GET DATA /TYPE=XLSX
```

```
  /FILE="\milesan1\projects\EEOL\Randomisation\Schools in sample Block 3.xlsx'.
```

*Check for duplicates.

```
sort cases by nfer_no.
```

```
match files file=*/first=f/last=l/by nfer_no.
```

```
cross f by l.
```

```
temp.
```

```
select if any(0, f, l).
```

```
list vars=nfer_no Schoolname.
```

*Randomise pupils.

```
set rng=mt, mtindex=29112013.
```

```
compute random=rv.uniform(0,1).
```

```
sort cases by random.
```

*Now allocate schools into groups on the basis of where they appear in the randomised order.

*If the number of schools is odd, decide which group the remainder will fall into before randomisation.

```
if $casenum eq 1 Group=1.
```

```
if $casenum eq 2 Group=2.
```

```
if $casenum eq 3 Group=3.
```

```
ADD VALUE LABELS Group 1 'Literacy intervention' 2 'Numeracy intervention' 3 'Control'.
```

```
freq group.
```

* There is a blank entry to make up the numbers. Delete it.

```
select if nfer_no ne 0.
```

Appendix 4: Statistical Analysis Plan

University of Oxford 'Improving Numeracy and Literacy'

Evaluation by National Foundation for Educational Research

18th July 2014

Introduction

This project will test the impact of two different initiatives on student outcomes: 'Mathematics and Reasoning' and 'Literacy and Morphemes'. The former is a numeracy intervention that develops children's understanding of the logical principles underlying mathematics. The latter is a literacy intervention that sees children being taught about morphemic spelling rules. The aim of the evaluation is to answer two primary research questions:

- what is the impact of 'Mathematics and Reasoning' on student development of numeracy ability?
- what is the impact of 'Literacy and Morphemes' on student development of reading ability?
- The research team aims to answer these questions by recruiting sixty primary or infant schools to participate in the study and randomly allocating each to one of the three groups:
- a literacy group: year 2 teachers and support staff to receive 'Literacy and Morphemes' training and materials
- a numeracy group: year 2 teachers and support staff to receive 'Mathematics and Reasoning' training and materials
- a control group: on a wait-list to receive training and materials in a program of their choice after the study is completed.

The Oxford research team administered tests in literacy (Progress in English 6) and numeracy (Progress in Maths 6) to all year 2 pupils in all schools to measure literacy and numeracy ability at baseline. The NFER research team will administer post-tests in literacy (Progress in English 7) and numeracy (Progress in Maths 7) to all the original pupils to measure the development of literacy and numeracy ability over time.

The aim of the analysis is to measure the differential progress in numeracy and literacy development over time between pupils in each group, to measure the effect of the training programmes and teaching materials.

Sample size

The aim of the study was to recruit 60 schools and allocate 20 to each group in the trial. Initial estimates of statistical power suggested that such a sample size would be sufficient to detect a standardised effect size of 0.22 with 80% power.⁶ This effect size was smaller than might be expected of the programs according to previous research, but was conservative given that previous research had not included a cluster-randomised trial.

However, the research team was only able to recruit 55 schools, which were all tested at baseline and allocated to a group. This reduced the expected power, increasing the minimum detectable effect size to 0.23.

⁶ Assuming a two-tailed test with a 95% confidence level, expected number of pupils per school=45, intra-school correlation=0.15, correlation between pre- and post-test scores=0.8, number of intervention and control schools=20.

Randomisation

It was intended at the start of the project that all schools would be randomised in one block and informed of their group allocation after every school had been tested at baseline, to prevent bias at the baseline test. However, this was not possible for a number of reasons, so alterations were made that satisfied schools, while still preventing bias.

Firstly, the training was the week after the end of the testing period. Schools understandably wanted to know their group allocation in advance of the training so they could arrange supply cover. Therefore, schools were randomised by an NFER statistician before the testing period and each school was notified of its group allocation shortly after being tested. This ensured that no school or test administrator knew the school's allocation at baseline testing, but allowed schools to make preparations.

Secondly, school recruitment continued until the last week of the testing period. The 51 schools that had been recruited before the pre-testing period were randomised together, but five schools were recruited during the testing period and were randomised separately. The need to inform the school of its group allocation soon after testing meant that the first three were randomised in one block and the final two in another block a week later.

Failing to include a variable that captures the potential difference in outcomes between those randomised earlier and those randomised later could overestimate the variance of estimates. On the other hand, controlling for randomisation block as well as the school's group in statistical modelling reduces the degrees of freedom, perhaps needlessly if there is no reason for thinking that the groups would differ. It seems very unlikely that there would be differences between the schools randomised in the second and third blocks substantial enough to justify an extra control variable in the modelling. Therefore, all analysis will include a dummy variable to indicate the five schools that were recruited after the first randomisation block to allow for any underlying differences between them and the schools that were recruited before testing.

The 51 schools that had agreed to participate at the beginning of the testing period were allocated 17 to each group using simple randomisation and informed of their group after testing (as explained above). One school decided to withdraw entirely from the research project after being randomised but before being tested at baseline. As the school staff did not know which group the school had been allocated to, this decision was made independently, so the dropout could not have been biased. However, the randomisation outcomes that had been decided were retained.

The 55 schools were allocated 17 to the numeracy group and 19 to the literacy group and 19 to the control group. The school that withdrew had been allocated to the numeracy group, which also happened to be the group not selected in the final randomisation block. Even groups would be preferable for maximising statistical power, but each block was randomised independently to prevent bias.

Outcome measures

Primary outcomes

The primary outcomes in the research are test scores on the Progress in English 7 for the literacy intervention and Progress in Maths 7 for the numeracy intervention. The tests have been administered by NFER test administrators to ensure independence from the research team at Oxford that delivered the intervention. Test administrators were not informed of which group the school had been allocated to and were advised to not discuss the interventions with teachers and to conduct the administration as they normally would. Testing took place in classes between 22nd April and 6th June 2014, with each school taking both tests on the same day.

The scores used will be raw test scores (i.e. number of correct answers) rather than age standardised scores. We expect within-year age to be evenly distributed across the three groups since they have been randomised. The primary interest is in average scores across schools, whereas age-standardised scores are useful for comparing pupils of different ages within a school; processing the scores further would be of limited benefit for this research and might introduce ceiling or floor effects.

Secondary outcomes

A number of secondary outcomes that were identified in the protocol will be used to test some of the secondary hypotheses.

Key Stage 1 maths, and reading & writing points – point score derived from Key Stage 1 levels in maths and average of levels in reading and writing. The data will be released by the Department for Education as ‘version 1’ data in September 2014 and as ‘final’ data in November 2014. Initial analysis will be conducted using ‘version 1’ data so that the provisional findings can be considered alongside the results of other analysis. While it is not expected that the two data sets will differ very much, the analysis reported in the final report will be based on ‘final’ data and any differences between the two sources noted. Because the underlying levels are assessed by the teacher that participated in the intervention there is a risk that results from analysis using Key Stage 1 points could be biased.

Sub-domains of Progress in English and Progress in Maths – the raw scores on the items of the Progress in English test that assess ‘Grammar’ and ‘Reading non-narrative’ and of the Progress in Maths test that assess ‘Solving Routine Problems’. The sub-domains are groups of items that are pre-specified by the test provider GL assessment. The sub-domains were identified by the Oxford research team in the protocol as being where the interventions are particularly likely to have an effect.

Analysis: overview and definitions

Multilevel modelling

Schools rather than pupils were randomised into groups, whereas the outcome measures are measures of pupil performance, so analysis of the difference in outcomes will need to take into account the fact that pupils are clustered in schools. The clustering of pupils within schools will be accounted for by using a multilevel model. Multilevel models estimate a school-level variance and a pupil-level variance, allowing the average outcome to be different across schools. Multilevel models have more statistical power than school-level analysis of averages, so is the main model for the primary analysis. The multilevel models will be estimated in the statistical software package ‘R’.

Each multilevel model will have the outcome of interest as the dependent variable and the following covariates will be included in every model:

- an indicator of whether the pupil’s school is in the literacy group and a separate indicator of whether the pupil’s school is in the numeracy group. The excluded group is the control group, so the coefficients of the group indicators measure the difference in (conditional) outcomes between that intervention group and the control group
- an indicator of whether the pupil’s school was one of the five schools that was recruited and randomised late (see ‘Randomisation’ section). The coefficient is incidental to the research, but controls for any differences in outcome due to underlying factors
- the pupil’s raw score on the pre-test, gender and age in completed months at post-test. For all analysis of literacy outcomes pre-test score will be the score on Progress in English 6, and for all analysis of numeracy outcomes pre-test score will be the score on Progress in Maths 6. The coefficients are incidental to the research, but explain a large proportion of outcome variance, increasing the power of the analysis.

Other covariates which may explain additional outcome variance, such as FSM status will not be included in the models because they are not costless to include. Such variables are obtained from the National Pupil Database (NPD). Some parents opted out of having their child's NPD records matched with their test score, so including variables from NPD in the analysis reduces the sample size. We believe that the gain in terms of additional explanatory power from including extra variables would not be sufficient to outweigh the loss of pupils from the final analysis, given the covariates that are listed above.

Standardised effect size

The coefficients of the statistical models are measured in terms of raw test score. As is standard for EEF evaluations and other research, a standardised effect size will be calculated which has a wider comparability with other research. The effect size will be calculated as the coefficient on the intervention group indicator divided by the pupil-level standard deviation.

The pupil-level variance is most appropriate for a cluster-randomised trial because the impact of interest is of the intervention on pupil performance and is the variance estimated for a pupil-randomised trial. Using the pupil-level rather than combined variance, means that the findings from pupil- and cluster-randomised are measured on the same terms.

The variance will be estimated from a separately-run multilevel model with the outcome variable as the dependent variable and no covariates. The pupil-level variance estimated by the multilevel model is a weighted-average sample standard deviation, so the effect size calculated is equivalent to Cohen's *d*. Since there are around 2,000 pupils in the trial there is no need to make Hedges' adjustment for small sample bias of estimated variance (except for the analysis of the FSM sub-sample: see below).

Confidence intervals

We will estimate a 95% confidence interval alongside the standardised effect size to give the precision with which the effect size has been estimated. The upper and lower bounds of the confidence interval will be calculated as the effect size plus/ minus the product of the critical value of the normal distribution (≈ 1.96) and the standard error of the group indicator coefficient estimated from the multilevel model.

Primary analysis

1. Primary analysis: the primary analysis of the impact of the interventions will be multilevel models of literacy outcome in the literacy group vs control and the numeracy outcome in the numeracy group vs control. The model will be estimated with all pupils in all schools that completed a pre-test and a post-test and include the covariates described above. The sample size and the rate of school and pupil attrition (i.e. the number of schools and pupils analysed compared to the number of schools and pupils that were randomised) will be reported.

2. Missing data analysis: missing data presents a problem for analysis, whether a pupil is missing a value for an outcome variable (post-test score) or for covariates (e.g. pre-test score). If outcome data is 'missing at random' given a set of covariates then the analysis has reduced power to detect an effect; if data is 'missing not at random' (for example, differential dropout in the intervention and control groups for unobserved reasons) then omitting these pupils (as in the primary 'completers' analysis) could bias the results. Imputing missing data could improve the robustness of the analysis and examine how sensitive the results are to alternative assumptions.

Every school that was randomised completed both a pre-test and a post-test. However, some individual pupils within those schools were not present for one or both tests because of absence or, for the post-test, because they left the school. For the pre-test, 130 pupils missed the literacy test and

119 pupils missed the numeracy test (116 missed both). Because the pre-test was conducted before the school knew the group allocation it is highly likely that pupils are missing at random. We will impute the missing pre-test scores for those pupils using multiple imputation, conditional on NPD data. NPD data includes Early Years Foundation Stage Profile (EYFSP) points, a measure of attainment at the end of reception. The correlation with pre-test score is high (0.63 for literacy, 0.62 for numeracy), so the imputed values will be estimated with a high level of precision. We will use the MLwiN software package to implement multilevel multiple imputation for pre-test scores, using EYFSP CLL7 (for literacy) or PSRN8 (for numeracy) points, gender, age and FSM as explanatory variables.

A discussion of the results in the context of missing follow-up data will be presented. If follow-up data is missing at random given covariates, and these covariates are included in the model, the results will be unbiased. It may be that the results of the multiple imputation do not differ appreciatively from the completers analysis. If this is the case and we are reasonably confident that covariates explain any missingness then this will complete the primary analysis. Otherwise, some sensitivity analysis (e.g. using extreme values) may be necessary.

Pre-specified secondary analysis

1. Analysis of differential impact on FSM pupils: the primary analysis models will be estimated for the sub-sample of pupils that have been FSM since starting school (the indicator of Pupil Premium status) at the time of the 2013 Spring School Census. The effect size will be measured in the same way except for the application of Hedges' correction for bias in small samples to the estimated sample variance. The pupil-level standard deviation will be re-estimated as described above for the FSM-only sample to use for standardising the estimated effect size.
2. Analysis of differential impact by pupil ability: the primary analysis models will each be run with the addition of two interaction covariates: pre-test score multiplied by the literacy group and numeracy group indicators.
3. Analysis of differential impact for EAL pupils: the primary analysis models will be run with the addition of two interaction covariates: EAL status (English as a first language/ English as an additional language) multiplied by the literacy group and numeracy group indicators.
4. Impact on Key Stage 1 points: the primary analysis models will be run with Key Stage 1 reading and writing points as the outcome variable for literacy and Key Stage 1 maths points as the outcome variable for numeracy. As the sample will be restricted to those with matched NPD data anyway because of the outcome variable, FSM and EAL will be included as additional covariates.
5. Impact on test sub-domains: the primary analysis models will be run with 'Grammar' and 'Reading non-narrative' raw scores as outcome variables for literacy and 'Solving routine problems' raw score as the outcome variable for numeracy.
6. Transfer effects: the impact of the intervention on the other outcome. The coefficients estimated as part of main primary analysis will be reported as evidence of the impact of the literacy intervention on the development of numeracy ability and the impact of the numeracy intervention on the development of literacy ability.

Exploratory analysis

Analysis will be conducted that is additional to the pre-specified analysis set out in the protocol. This set of analyses will focus on the implementation of the interventions to attempt to understand how the interventions worked rather than whether they worked. Differences in the way the intervention was implemented across intervention schools were at the discretion of school staff and researchers on the

⁷ Communication, Language and Literacy scale.

⁸ Problem Solving, Reasoning and Numeracy scale.

Oxford team, so this analysis cannot be definitively interpreted as causal. However, the analysis may give clues as to what drives any effect that may be identified by the impact analysis. It will investigate whether different approaches to implementation led to different amounts of progress made between schools within each intervention. Two analyses will be conducted:

1. Use of computer games: an integral part of both the literacy and numeracy interventions was the use of computer games to reinforce the class teaching. Pupils accessed the games with an individual login and the number of games played was automatically logged. The usage data for each pupil will be matched to the pupil baseline, characteristics and outcome data. We will run the primary analysis models again with the number of literacy (numeracy) games played replacing the literacy (numeracy) group indicator to see whether increased use of computer games by pupils was associated with more progress made in the tests. The variable will be coded as zero for the control and numeracy (literacy) groups as the games were only accessible for those schools allocated to each intervention.

2. Time spent delivering the intervention: the teacher questionnaire asked teachers how many hours they spent teaching the intervention and how much time pupils spent playing the computer games in and out of class. We will run the primary analysis models again with the number of hours spent doing activities related to the literacy (numeracy) intervention replacing the literacy (numeracy) group indicator to see whether more time spent on the intervention was associated with more progress made in the tests. The variable will be coded as zero for the control and numeracy (literacy) groups.

Appendix 5: Primary analysis

Table A1: Primary analysis (numeracy)

	Coefficient	Standard error	p-value	95% confidence interval
Numeracy pre-test raw score	0.81	0.02	0.00	0.78 – 0.84
Literacy group	-0.07	0.42	0.86	-0.91 – 0.76
Numeracy group	0.97	0.44	0.03	0.11 – 1.83
Late randomised	-1.62	0.65	0.02	-2.89 – -0.35
Female	Excluded as non-significant by backward selection			
Age in months at post-test	Excluded as non-significant by backward selection			
Intercept	1.45	0.45	0.00	0.57 – 2.32

Number of cases = 1,942.

	No covariate model	With covariates
School-level variance	3.16	1.35
Pupil-level variance	24.00	10.36
Intra-cluster correlation	0.12	0.11
Pupil-level standard deviation	4.90*	3.22

* Used for calculating effect size.

Table A2: Primary analysis (literacy)

	Coefficient	Standard error	p-value	95% confidence interval
Literacy pre-test raw score	0.73	0.01	0.00	0.71 – 0.75
Literacy group	-0.40	0.52	0.45	-1.42 – 0.62
Numeracy group	0.66	0.54	0.23	-0.40 – 1.71
Late randomised	-2.12	0.81	0.01	-3.70 – -0.54
Female	0.46	0.21	0.03	0.05 – 0.86
Age in months at post-test	Excluded as non-significant by backward selection			
Intercept	6.97	0.46	0.00	6.06 – 7.88

Number of cases = 1,940.

	No covariate model	With covariates
School-level variance	4.65	1.88
Pupil-level variance	60.86	19.74
Intra-cluster correlation	0.07	0.09
Pupil-level standard deviation	7.80*	4.44

* Used for calculating effect size.

Table A3: Missing data analysis (numeracy)

	Coefficient	Standard error	p-value	95% confidence interval
Numeracy pre-test raw score	0.81	0.02	0.00	0.78 – 0.84
Literacy group	-0.06	0.45	0.89	-0.94 – 0.81
Numeracy group	0.96	0.46	0.04	0.05 – 1.86
Late randomised	-1.65	0.68	0.02	-2.97 – -0.32
Female	Excluded as non-significant by backward selection			
Age in months at post-test	Excluded as non-significant by backward selection			
Intercept	1.33	0.46	0.00	0.43 – 2.22

Number of cases = 2,217.

	No covariate model	With covariates
School-level variance	3.16	1.53
Pupil-level variance	24.00	10.42
Intra-cluster correlation	0.12	0.13
Pupil-level standard deviation	4.90*	3.23

* Used for calculating effect size.

Table A4: Missing data analysis (literacy)

	Coefficient	Standard error	p-value	95% confidence interval
Literacy pre-test raw score	0.73	0.01	0.00	0.71 – 0.76
Literacy group	-0.40	0.55	0.47	-1.48 – 0.67
Numeracy group	0.63	0.57	0.27	-0.48 – 1.74
Late randomised	-2.05	0.85	0.02	-3.70 – -0.39
Female	0.46	0.21	0.03	0.05 – 0.87
Age in months at post-test	Excluded as non-significant by backward selection			
Intercept	6.82	0.48	0.00	5.88 – 7.76

Number of cases = 2,217.

	No covariate model	With covariates
School-level variance	4.65	2.18
Pupil-level variance	60.86	19.91
Intra-cluster correlation	0.07	0.10
Pupil-level standard deviation	7.80*	4.46

* Used for calculating effect size.

Table A5: FSM sub-group analysis (numeracy)

	Coefficient	Standard error	p-value	95% confidence interval
Numeracy pre-test raw score	0.87	0.04	0.00	0.79 – 0.96
Literacy group	-0.44	0.55	0.43	-1.52 – 0.64
Numeracy group	0.77	0.63	0.23	-0.46 – 2.00
Late randomised	Excluded as non-significant by backward selection			
Female	Excluded as non-significant by backward selection			
Age in months at post-test	Excluded as non-significant by backward selection			
Intercept	-0.56	0.87	0.52	-2.28 – 1.15

Number of cases = 307.

	No covariate model	With covariates
School-level variance	0.10	0.43
Pupil-level variance	29.77	11.98
Intra-cluster correlation	0.00	0.03
Pupil-level standard deviation	5.46*	3.46

* Used for calculating effect size.

Table A6: FSM sub-group analysis (literacy)

	Coefficient	Standard error	p-value	95% confidence interval
Literacy pre-test raw score	0.71	0.03	0.00	0.65 – 0.77
Literacy group	0.86	0.97	0.38	-1.04 – 2.76
Numeracy group	0.68	1.07	0.53	-1.41 – 2.77
Late randomised	Excluded as non-significant by backward selection			
Female	Excluded as non-significant by backward selection			
Age in months at post-test	Excluded as non-significant by backward selection			
Intercept	6.27	0.87	0.00	4.56 – 7.97

Number of cases = 301.

	No covariate model	With covariates
School-level variance	6.71	2.81
Pupil-level variance	66.70	22.15
Intra-cluster correlation	0.09	0.11
Pupil-level standard deviation	8.17*	4.71

* Used for calculating effect size.

Table A7: FSM interaction analysis (numeracy)

	Coefficient	Standard error	p-value	95% confidence interval
Numeracy pre-test raw score	0.80	0.02	0.00	0.76 – 0.83
Literacy group	-0.14	0.45	0.76	-1.02 – 0.74
Numeracy group	0.77	0.46	0.10	-0.14 – 1.68
Literacy group * FSM	-0.05	0.50	0.92	-1.03 – 0.93
Numeracy group * FSM	0.75	0.58	0.19	-0.38 – 1.89
FSM	-1.07	0.30	0.00	-1.66 – -0.48
Late randomised	-1.66	0.67	0.02	-2.97 – -0.36
Female	Excluded as non-significant by backward selection			
Age in months at post-test	0.05	0.02	0.03	0.00 – 0.09
Intercept	-2.14	1.84	0.25	-5.76 – 1.47

Number of cases = 1,822.

	No covariate model	With covariates
School-level variance	3.15	1.41
Pupil-level variance	23.93	10.12
Intra-cluster correlation	0.12	0.12
Pupil-level standard deviation	4.89*	3.18

* Used for calculating effect size.

Table A8: FSM interaction analysis (literacy)

	Coefficient	Standard error	p-value	95% confidence interval
Literacy pre-test raw score	0.74	0.01	0.00	0.71 – 0.76
Literacy group	-0.58	0.50	0.25	-1.55 – 0.40
Numeracy group	0.63	0.52	0.22	-0.38 – 1.65
Literacy group * FSM	2.00	0.69	0.00	0.65 – 3.35
Numeracy group * FSM	0.60	0.81	0.46	-0.99 – 2.19
FSM	-1.19	0.42	0.00	-2.00 – -0.37
Late randomised	-2.20	0.76	0.01	-3.69 – -0.72
Female	0.45	0.21	0.03	0.03 – 0.86
Age in months at post-test	Excluded as non-significant by backward selection			
Intercept	6.76	0.47	0.00	5.84 – 7.68

Number of cases = 1,821.

	No covariate model	With covariates
School-level variance	4.54	1.47
Pupil-level variance	61.50	19.42
Intra-cluster correlation	0.07	0.07
Pupil-level standard deviation	7.84*	4.41

* Used for calculating effect size.

Table A9: Prior ability interaction analysis (numeracy)

	Coefficient	Standard error	p-value	95% confidence interval
Numeracy pre-test raw score	0.81	0.02	0.00	0.76 – 0.86
Literacy group	-0.02	0.92	0.99	-1.82 – 1.78
Numeracy group	1.50	0.93	0.11	-0.32 – 3.31
Numeracy group * Numeracy pre-test raw score	-0.03	0.04	0.52	-0.10 – 0.05
Literacy group * Numeracy pre-test raw score	0.00	0.04	0.94	-0.08 – 0.07
Late randomised	-1.61	0.65	0.02	-2.89 – -0.33
Female	Excluded as non-significant by backward selection			
Age in months at post-test	0.04	0.02	0.04	0.00 – 0.08
Intercept	-2.31	1.82	0.20	-5.88 – 1.26

Number of cases = 1,942.

	No covariate model	With covariates
School-level variance	3.16	1.35
Pupil-level variance	24.00	10.33
Intra-cluster correlation	0.12	0.12
Pupil-level standard deviation	4.90*	3.21

* Used for calculating effect size.

Table A10: Prior ability interaction analysis (literacy)

	Coefficient	Standard error	p-value	95% confidence interval
Literacy pre-test raw score	0.72	0.02	0.00	0.69 – 0.75
Literacy group	-0.92	0.84	0.28	-2.57 – 0.73
Numeracy group	0.47	0.86	0.58	-1.20 – 2.15
Literacy group * Literacy pre-test raw score	0.02	0.03	0.43	-0.03 – 0.08
Numeracy group * Literacy pre-test raw score	0.01	0.03	0.78	-0.05 – 0.06
Late randomised	-2.09	0.81	0.01	-3.67 – -0.52
Female	0.45	0.21	0.03	0.05 – 0.86
Age in months at post-test	Excluded as non-significant by backward selection			
Intercept	7.16	0.55	0.00	6.08 – 8.25

Number of cases = 1,940.

	No covariate model	With covariates
School-level variance	4.65	1.87
Pupil-level variance	60.86	19.74
Intra-cluster correlation	0.07	0.09
Pupil-level standard deviation	7.80*	4.44

* Used for calculating effect size.

Table A11: EAL interaction analysis (numeracy)

	Coefficient	Standard error	p-value	95% confidence interval
Numeracy pre-test raw score	0.80	0.02	0.00	0.77 – 0.83
Literacy group	-0.10	0.44	0.82	-0.97 – 0.77
Numeracy group	1.01	0.46	0.03	0.11 – 1.90
Numeracy group * EAL	-0.17	0.57	0.77	-1.28 – 0.95
Literacy group * EAL	0.04	0.49	0.93	-0.92 – 1.01
Late randomised	-1.71	0.66	0.01	-3.00 – -0.42
Female	Excluded as non-significant by backward selection			
Age in months at post-test	0.05	0.02	0.03	0.00 – 0.09
EAL	0.28	0.34	0.41	-0.39 – 0.94
Intercept	-2.41	1.84	0.19	-6.01 – 1.19

Number of cases = 1,848.

	No covariate model	With covariates
School-level variance	3.23	1.35
Pupil-level variance	23.86	10.25
Intra-cluster correlation	0.12	0.12
Pupil-level standard deviation	4.88*	3.20

* Used for calculating effect size.

Table A12: EAL interaction analysis (literacy)

	Coefficient	Standard error	p-value	95% confidence interval
Literacy pre-test raw score	0.73	0.01	0.00	0.71 – 0.76
Literacy group	-0.29	0.50	0.56	-1.27 – 0.69
Numeracy group	0.94	0.51	0.07	-0.07 – 1.94
Literacy group * EAL	0.13	0.67	0.84	-1.18 – 1.45
Numeracy group * EAL	-0.49	0.78	0.53	-2.01 – 1.03
Late randomised	-2.10	0.75	0.01	-3.58 – -0.63
Female	0.51	0.21	0.02	0.09 – 0.92
Age in months at post-test	Excluded as non-significant by backward selection			
EAL	0.30	0.46	0.51	-0.59 – 1.19
Intercept	6.98	0.46	0.00	6.08 – 7.87

Number of cases = 1,845.

	No covariate model	With covariates
School-level variance	4.68	1.43
Pupil-level variance	61.49	19.74
Intra-cluster correlation	0.07	0.07
Pupil-level standard deviation	7.84*	4.44

* Used for calculating effect size.

Table A13: Key Stage 1 mathematics points

	Coefficient	Standard error	p-value	95% confidence interval
Numeracy pre-test raw score	0.50	0.01	0.00	0.48 – 0.52
Literacy group	-0.19	0.30	0.54	-0.78 – 0.40
Numeracy group	-0.10	0.31	0.74	-0.71 – 0.51
Late randomised	-0.96	0.45	0.04	-1.85 – -0.08
Female	Excluded as non-significant by backward selection			
Age in months at post-test	0.04	0.01	0.00	0.02 – 0.07
Intercept	2.86	1.14	0.01	0.63 – 5.08

Number of cases = 1,965.

	No covariate model	With covariates
School-level variance	0.58	0.69
Pupil-level variance	9.81	4.04
Intra-cluster correlation	0.06	0.15
Pooled standard deviation	3.13*	2.01

* Used for calculating effect size.

Table A14: Key Stage 1 reading and writing points

	Coefficient	Standard error	p-value	95% confidence interval
Literacy pre-test raw score	0.30	0.00	0.00	0.29 – 0.31
Literacy group	-0.18	0.31	0.55	-0.78 – 0.41
Numeracy group	0.05	0.32	0.88	-0.57 – 0.67
Late randomised	-0.95	0.46	0.04	-1.84 – -0.06
Female	0.37	0.09	0.00	0.20 – 0.54
Age in months at post-test	0.03	0.01	0.04	0.00 – 0.05
Intercept	6.98	1.07	0.00	4.88 – 9.08

Number of cases = 1,955.

	No covariate model	With covariates
School-level variance	0.67	0.74
Pupil-level variance	10.74	3.50
Intra-cluster correlation	0.06	0.17
Pooled standard deviation	3.28*	1.87

* Used for calculating effect size.

Table A15: Solving routine problems sub-domain score (numeracy)

	Coefficient	Standard error	p-value	95% confidence interval
Numeracy pre-test raw score	3.34	0.08	0.00	3.18 – 3.49
Literacy group	0.79	1.83	0.67	-2.79 – 4.38
Numeracy group	5.39	1.89	0.01	1.68 – 9.10
Late randomised	-7.37	2.85	0.01	-12.97 – -1.78
Female	-1.63	0.75	0.03	-3.10 – -0.15
Age in months at post-test	Excluded as non-significant by backward selection			
Intercept	-14.16	2.13	0.00	-18.33 – -9.98

Number of cases = 1,942.

	No covariate model	With covariates
School-level variance	66.11	22.57
Pupil-level variance	499.19	267.94
Intra-cluster correlation	0.12	0.08
Pooled standard deviation	22.34*	16.37

* Used for calculating effect size.

Table A16: Grammar sub-domain score (literacy)

	Coefficient	Standard error	p-value	95% confidence interval
Literacy pre-test raw score	1.98	0.06	0.00	1.86 – 2.10
Literacy group	0.82	3.23	0.80	-5.51 – 7.15
Numeracy group	2.83	3.34	0.40	-3.71 – 9.36
Late randomised	Excluded as non-significant by backward selection			
Female	Excluded as non-significant by backward selection			
Age in months at post-test	Excluded as non-significant by backward selection			
Intercept	14.25	2.68	0.00	9.00 – 19.50

Number of cases = 1,940.

	No covariate model	With covariates
School-level variance	87.10	79.98
Pupil-level variance	843.86	541.28
Intra-cluster correlation	0.09	0.13
Pooled standard deviation	29.05*	23.27

* Used for calculating effect size.

Table A17: Reading non-narrative sub-domain score (literacy)

	Coefficient	Standard error	p-value	95% confidence interval
Literacy pre-test raw score	2.17	0.06	0.00	2.06 – 2.29
Literacy group	-4.16	2.50	0.10	-9.07 – 0.74
Numeracy group	3.12	2.59	0.23	-1.95 – 8.19
Late randomised	Excluded as non-significant by backward selection			
Female	Excluded as non-significant by backward selection			
Age in months at post-test	Excluded as non-significant by backward selection			
Intercept	13.55	2.23	0.00	9.17 – 17.92

Number of cases = 1,940.

	No covariate model	With covariates
School-level variance	79.84	41.76
Pupil-level variance	882.20	522.35
Intra-cluster correlation	0.08	0.07
Pooled standard deviation	29.70*	22.86

* Used for calculating effect size.

Table A18: On-treatment analysis—number of games played (numeracy)

	Coefficient	Standard error	p-value	95% confidence interval
Numeracy pre-test raw score	0.80	0.02	0.00	0.77 – 0.83
Literacy group	-0.07	0.39	0.85	-0.84 – 0.69
Numeracy group * number of games played	0.05	0.01	0.00	0.02 – 0.07
Late randomised	-1.79	0.65	0.01	-3.07 – -0.51
Female	Excluded as non-significant by backward selection			
Age in months at post-test	Excluded as non-significant by backward selection			
Intercept	-14.16	2.13	0.00	-18.33 – -9.98

Number of cases = 1,940.

	No covariate model	With covariates
School-level variance	3.13	1.38
Pupil-level variance	24.02	10.29
Intra-cluster correlation	0.12	0.12
Pooled standard deviation	4.90*	3.21

* Used for calculating effect size.

Table A19: On-treatment analysis—number of games played (literacy)

	Coefficient	Standard error	p-value	95% confidence interval
Literacy pre-test raw score	0.73	0.01	0.00	0.70 – 0.75
Literacy group * number of games played	0.02	0.02	0.13	-0.01 – 0.06
Numeracy group	1.02	0.49	0.04	0.07 – 1.97
Late randomised	-2.18	0.80	0.01	-3.75 – -0.60
Female	0.45	0.21	0.03	0.05 – 0.86
Age in months at post-test	Excluded as non-significant by backward selection			
Intercept	6.67	0.40	0.00	5.89 – 7.45

Number of cases = 1,936.

	No covariate model	With covariates
School-level variance	4.54	1.86
Pupil-level variance	60.86	19.74
Intra-cluster correlation	0.07	0.09
Pooled standard deviation	7.80*	4.44

* Used for calculating effect size.

Table A20: On-treatment analysis—number of classroom hours (numeracy)

	Coefficient	Standard error	p-value	95% confidence interval
Numeracy pre-test raw score	0.80	0.02	0.00	0.77 – 0.83
Literacy group	-0.63	0.42	0.14	-1.46 – 0.20
Numeracy group * number of classroom hours	-0.02	0.02	0.31	-0.05 – 0.02
Late randomised	-1.79	0.70	0.01	-3.15 – -0.42
Female	Excluded as non-significant by backward selection			
Age in months at post-test	Excluded as non-significant by backward selection			
Intercept	2.11	0.44	0.00	1.25 – 2.98

Number of cases = 1,861.

	No covariate model	With covariates
School-level variance	3.28	1.63
Pupil-level variance	23.61	10.40
Intra-cluster correlation	0.12	0.14
Pooled standard deviation	4.86*	3.23

* Used for calculating effect size.

Table A21: On-treatment analysis—number of classroom hours (literacy)

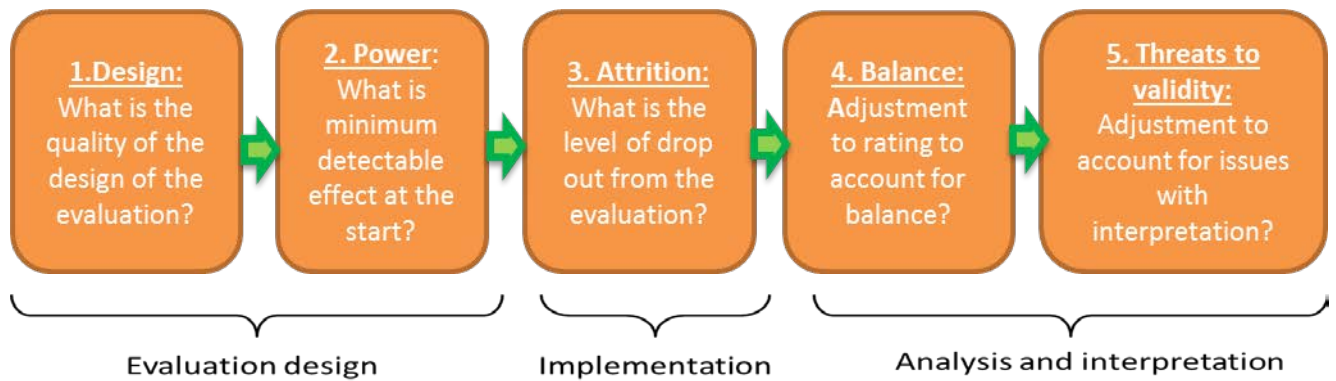
	Coefficient	Standard error	p-value	95% confidence interval
Literacy pre-test raw score	0.74	0.01	0.00	0.71 – 0.76
Literacy group * number of classroom hours	-0.04	0.04	0.32	-0.13 – 0.04
Numeracy group	0.80	0.53	0.13	-0.23 – 1.84
Late randomised	-1.74	0.83	0.04	-3.37 – -0.11
Female	0.53	0.21	0.01	0.12 – 0.94
Age in months at post-test	Excluded as non-significant by backward selection			
Intercept	6.60	0.45	0.00	5.71 – 7.49

Number of cases = 1,820.

	No covariate model	With covariates
School-level variance	5.01	1.88
Pupil-level variance	60.79	19.25
Intra-cluster correlation	0.08	0.09
Pooled standard deviation	7.80*	4.39

* Used for calculating effect size.

Appendix 6: Security classification of trial findings



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

Light : MDES 0.18; 0% attrition; no imbalance; NFER administered tests = good;

The final security rating for this trial is 5 . This means that the conclusions have high security.

This evaluation was designed as a randomised controlled trail. The sample size was designed to detect a MDES of less than 0.2. There was little attrition of pupils, and zero attrition of the schools – the units of randomisation. The post-tests were administered by the independent evaluators. Balance at baseline was high, and there were no substantial threats to validity. Therefore, the final security rating is 5 .

Appendix 7: Cost rating

Cost ratings are based on the approximate cost per pupil per year of implementing the intervention over three years. More information about the EEF's approach to cost evaluation can be found on the EEF website. Cost ratings are awarded as follows:

Cost rating	Description
£	<i>Very low:</i> less than £80 per pupil per year.
£ £	<i>Low:</i> up to about £200 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.

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