

E
E
F
ducation
ndowment
oundation

Butterfly Phonics

Evaluation Report and Executive Summary

February 2015

Independent evaluators:

Christine Merrell

Adetayo Kasim



The Education Endowment Foundation (EEF)



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

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

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

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

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



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

Peter Henderson

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

p: 0207802 1923

e: peter.henderson@eefoundation.org.uk

w: www.educationendowmentfoundation.org.uk

About the evaluator

The project was independently evaluated by a team from the Centre for Evaluation & Monitoring (CEM), Durham University and the Wolfson Institute for Health and Wellbeing, Durham University.

Contact details:

Centre for Evaluation and Monitoring,

Rowan House,
Durham University,
Stockton Road,
Durham,
DH1 3UZ.

p: 0191 3344226

e: Christine.Merrell@cem.dur.ac.uk

Contents

| | |
|--------------------------|----|
| Executive Summary | 4 |
| Introduction | 6 |
| Methodology..... | 9 |
| Impact evaluation | 16 |
| Process evaluation | 26 |
| Conclusion | 34 |
| References..... | 38 |

Executive summary

The project

Butterfly Phonics aims to improve the reading of struggling pupils through phonics instruction and a formal teaching style where pupils sit at desks in rows facing the teacher. The teacher directs questions to the pupils throughout the lesson in order to check their understanding. It is based on a course book created by Irina Tyk, and was delivered in this evaluation by Real Action, a charity based in London.

Real Action staff recruited and trained practitioners to deliver the intervention. These practitioners worked with trained teaching assistants to teach classes of six to eight pupils, although some groups were larger. Pupils were eligible for participation in the trial if they did not reach level 4 in their Key Stage 2 SATs or their reading skills were at least a year behind their chronological age. In most schools, lessons were taught over a period of ten to twelve weeks, typically with two one-hour lessons each week. One school delivered the intervention over just four weeks, and the implications of this variation are discussed in the main body of the report.

The evaluation was set up as a randomised controlled trial, which compared the progress of pupils who received Butterfly Phonics to a “business-as-usual” control group. It did not test the delivery model at scale, and should therefore be considered an efficacy trial. The study was funded by the Education Endowment Foundation as one of 23 projects focused on literacy catch-up at the primary-secondary transition. It is one of seven literacy catch-up projects with a focus on phonics.

Key conclusions

1. This evaluation provided evidence of promise; there was a positive, statistically significant effect on the primary outcome measure of reading comprehension. However, this effect size was lower than the minimum detectable effect size of the trial, so we cannot confidently conclude that the effect was due to the intervention and did not occur by chance.
2. The secondary outcome measures indicated positive impacts on children’s literacy skills, but these were not statistically significant.
3. This intervention is recommended to take place during the school day, when it is easier to secure sustained co-operation and support from school staff. Where that support was present, the intervention was able to progress more satisfactorily than in schools where it was lacking.
4. Schools should ensure that people delivering the intervention receive training in the Butterfly method so that it is implemented as intended.
5. Further research could investigate the intervention’s impact on early readers. Its emphasis on larger word units and comprehension skills might enable a more rapid progression in early reading than a pure phonics course.

What impact did it have?

The evaluation found that, on average, the reading comprehension skills of pupils who received the intervention improved at a faster rate than those in the control group. This improvement is equivalent to the pupils who received the intervention making an additional five months’ progress over the course of the school year. This estimate is statistically significant, but the observed effect size is lower than the minimum detectable effect size that was specified at the beginning of the study. This means that,

although this evaluation provides evidence of promise, we are unable to confidently conclude that the observed effect is real and did not occur by chance.

The evaluation also considered the impact of the intervention on two secondary outcome measures of literacy skills, but did not find statistically significant impacts.

How secure is this finding?

The primary analysis in this evaluation is judged to be of weak security and was awarded a security rating of 0 padlocks. The main cause of the low security rating awarded to this evaluation is that the differences between schools' post-test results were greater than envisaged at the beginning of the trial. This meant that the trial was not large enough to confidently detect an effect as small as the one that was ultimately observed. We cannot therefore confidently conclude that the observed effect is real and did not occur by chance.

A further limitation of the study is that the test administrators reported that the post-tests in two schools were disrupted by the poor behaviour of the pupils involved. It was judged reasonable for the pupils to re-sit the test in one of these schools and the primary analysis, which was performed on an 'intention to treat' basis, used this re-sit data. It was not judged reasonable for the pupils in the other school to re-sit the post-test and therefore the data from the disrupted post-test session was included in the primary analysis. A report about the compromised administration conditions of the post-test in this school is included in Appendix 2. These disrupted post-tests and their inclusion in the final analysis should be considered when interpreting the security of the findings.

It is possible that the positive impact was caused by a 'confounding' factor. Most of the participating schools scheduled the intervention to take place outside English lessons and some of the effect was possibly due to the pupils in the intervention group receiving additional English teaching, not the nature of the Butterfly Phonics intervention itself. Also, pupils received the intervention in small groups, while pupils in the control group continued with normal classroom teaching.

The schools involved were located within a small geographical area of London, and care should be taken when applying these findings to schools in different contexts.

How much does it cost?

The cost of Butterfly Phonics as it was delivered in this evaluation is estimated at £108.50 per pupil. This estimate is based on the assumption that there are eight pupils in each Butterfly Phonics class, and includes the salaries of the Butterfly Phonics teaching staff, training costs and the cost of course books (about £10 each). The cost of the Butterfly teaching staff for a one hour class was £35-40 and the specialist training from Irina Tyk cost £600 for half a day.

| Group | No. of pupils | Effect size (95% confidence interval) | Estimated months' progress | Evidence strength* | Cost** |
|---|---------------|---------------------------------------|----------------------------|---|--------|
| Intervention vs control (all pupils) | 310 | 0.43 (0.03, 0.84) | +5 months |  | ££ |
| Intervention vs control (FSM) | 140 | 0.16 (-0.18, 0.49) | +2 months |  | ££ |

**For more information about evidence ratings, see Appendix 3 in the main evaluation report. Evidence ratings are not provided for sub-group analyses, which will always be less secure than overall findings.*

***For more information about cost ratings, see Appendix 4 in the main evaluation report.*

Introduction

Intervention

The study was funded as part of a £10 million grant awarded from the Department for Education to the EEF for projects dedicated to literacy catch-up for pupils at the primary–secondary transition who do not achieve level 4 in English by the end of Key Stage 2.

The course book formed the basis of the Butterfly intervention. Each lesson was a chapter in the course book and was intended to take one hour to deliver. *The Butterfly Book* (Tyk, 2007) was written for beginning readers and the least skilled catch-up readers. This was used to teach the weakest readers in the intervention. The *Advanced Butterfly Reader* (unpublished) was aimed at more skilled, yet struggling, readers with content suitable for older children including teenagers or even adults. The strongest readers in the sample began their studies with this book. Pupils were taught with the *Advanced Reader* when they reached the appropriate level to benefit from it.

Pupils were taught in small groups, typically of six to eight, by a trained Butterfly practitioner and assisted by a trained teaching assistant. The recommended class size for Butterfly is less than 15. Pupils were withdrawn from their normal lessons (which were a variety of subjects but usually not English) to receive the intervention in all but one of the schools. In the remaining school, the intervention was implemented outside of school hours. The teaching style was formal, whereby the pupils sat at desks in rows facing the teacher, who directed questions to the pupils at random throughout the lesson in order to check their understanding.

Background evidence

The EEF Toolkit lists phonics as an effective strategy for early reading and for catch-up programmes although impact declines with age and there is a weaker effect when used as an intervention for struggling readers at and beyond the end of primary school. The Toolkit suggested that a gain of three months could be achieved but that the strength of the evidence on which this was based was moderate. Butterfly Phonics was taught as an intervention to pupils in Year 7 and so its effect might be expected to be weak.

There is also the question of which literacy skill a phonics intervention would address. The National Reading Panel (2000) in the United States reported that phonics programmes benefited typical and struggling readers between the ages of 5 and 12, but that the improvement was in decoding, with very little improvement in comprehension for older struggling readers (non-significant, effect size (d) = 0.12). The primary outcome of this report is that of reading comprehension, and decoding was measured as a secondary outcome. However, there is the question of how much of a ‘pure’ phonics course Butterfly actually is, incorporating aspects of comprehension instruction. Mixed instructional techniques have been found to be linked with bigger effect sizes than phonics courses in older children (Suggate, 2010).

The main source of background evidence for the effectiveness of Butterfly is the document written by Alister Wedderburn on behalf of The Educational Trust in 2011, which can be downloaded from the Real Action website: <http://www.realaction.org.uk>. The main conclusion reached by Wedderburn was that Butterfly is an effective way to teach reading. The research took place between October 2010 and July 2011 and studied children between the ages of 5 and 12 years in the area of London where Real Action delivers the Butterfly course. Among its findings, the study reported that children who attended the Saturday morning Butterfly school were 35 times more likely to attain National Curriculum level 4 or above in their Key Stage 2 (KS2) English SATs than their local counterparts who did not attend the school, and 90% more likely to gain level 4 or above in both maths and English in KS2 SATs. It was also found that the children who attended Butterfly lessons had a better school

attendance record compared to the local children who did not attend. However, it is possible that although attendance was not an intended outcome of the Butterfly intervention, those who attended Butterfly were more likely to attend school.

The present randomised controlled trial employs a control group, which has been randomly selected, in order to compare with the treatment group's results. A randomised controlled trial within a full-time school environment avoids the inconsistencies that can emerge from studying children who attend a voluntary out-of-school course which would have a self-selecting sample and where children's progress may vary more because of missing sessions than they would in a school situation. The geographical area in which the research was conducted was very narrow.

Butterfly Phonics is presently confined to northwest London, where the Real Action charity is situated. Butterfly is used in their Saturday morning classes, which are available to the public and are popular with children of all ages who are struggling with their literacy skills. Real Action has worked with some secondary schools in the area and they currently have invitations from headteachers in other parts of London to deliver the intervention in those schools too.

Evaluation objectives

The main question that the evaluation set out to answer is: Does Butterfly Phonics positively impact on the reading comprehension scores of struggling readers in Year 7 compared with a randomly selected control group of pupils who followed the usual school curriculum?

Owing to the particular nature of the relationship between reading comprehension and decoding, two subsidiary questions were included, which were tested by two further outcome measures: Does the intervention have a positive impact on the treatment group compared to a control group in terms of the ability to read:

- Single real words, both with regular relationships between letters and sounds (graphemes and phonemes) and irregular grapheme–phoneme correspondences?
- Non-words (made-up words), indicating the use of the sublexical phonological route of grapheme–phoneme conversion and phoneme blending (Coltheart, 1978; Coltheart et al., 2001)

The secondary outcomes were included to investigate an improvement in reading skills which could be considered as intermediary towards improvement in comprehension, which is a higher-order skill.

The process evaluation focused on the fidelity of the intervention's implementation. Thus, the primary aim was to monitor whether the intervention was being delivered in accordance with the training received by the delivery team. The training should accurately represent how the author intended the intervention to be delivered. The evaluation was not intended to identify areas of improvement in delivery, but the interviews and surveys with the staff delivering the intervention may include such points.

Project team

The Butterfly project was headed by Katie Ivens, the Education Director of Real Action. Jemma Carvajal-Pym, Project Director for Butterfly, was responsible for the overall management of the intervention and was the chief liaison agent with the EEF and the evaluator. Viviane Peressini, Project Manager for Butterfly, had responsibilities in much of the day-to-day running of the project. Almaz Ohene and Sam Grolimund were Project Assistants. Seventeen Butterfly Practitioners, recruited by Real Action, taught the intervention lessons, accompanied by seven trained teaching assistants. These individuals were trained and observed lessons without payment prior to doing any teaching. They were paid £20 per hour when they were teaching. Among these Butterfly practitioners were

some local people and others from further afield, but all had experience in working with young people; most were postgraduate students attending courses at London University (UCL, Kings and Goldsmiths, LSE, and other London colleges), and some were fully trained teachers.

Thirteen people were recruited separately by Real Action to administer and mark the tests. They were intended to be independent of the implementation of the project and blind to the allocation of each pupil in order to ensure that the testing procedures and marking were fair. The EEF regulations on blinded testing were therefore carried out.

Ethical review

The Durham University School of Education Ethics Committee gave ethical approval for the evaluation in December 2012.

The EEF guidelines for parental consent were followed and parents were given the opportunity, before the start of the evaluation, to opt their children out of the trial. They were informed that the data would be stored by the EEF for longitudinal research purposes. The right of the children to withdraw at any time, and the preservation of anonymity in reports, were among the points covered in the letters and information about the trial was also shared with the parents. See Appendix 1 for the parental consent letter.

Methodology

Trial design

This randomised controlled trial consisted of a treatment group, which received the intervention, and a control group who continued their schooling as usual. The intervention took place in school time for five out of the six participating schools and in a variety of lessons, detailed later in the report. No control task was involved so that comparison was 'business as usual'. The unit of randomisation was the individual pupil.

The project team chose this design to conform to the EEF's rigorous design specifications, whereby an intervention can be assessed by comparing a treatment group to a group that does not receive the treatment. Critically, individual students are randomly assigned to either the treatment or the control group so as to avoid bias. They were randomised within each school.

The control group was actually a waitlist condition which continued with 'business as usual' school activities while the treatment group received the intervention. Their reading was assessed at the same time as their counterparts in the intervention group. At the end of the intervention, the pupils in the control group were given the opportunity to receive the intervention by the Butterfly practitioners.

The trial was originally intended to be conducted with the one cohort of Year 7 pupils within the same academic year. After establishing the eligibility of pupils, as described below, the eligible pupils were assessed with the pre-tests, which are described in detail later. The scores of the pre-tests were used within the randomisation procedure. The intervention was implemented and the pupils then completed the post-tests. The pre-tests were also included in the analysis of the impact of the intervention.

The withdrawal of the biggest school from the study on the day before the pre-test necessitated an alteration in the structure of the trial in that it was permitted to join with their new cohort of Year 7 pupils in the next academic year, referred to as Phase Two. A sixth school was also recruited to Phase Two. The advantages of this were considered to outweigh the possible variability this might introduce by including children born in successive academic years. The increased numbers would increase the statistical power, thus improving the likelihood of any uncontrolled variables being randomly distributed between the groups and reducing the chance of bias.

Eligibility

The participating schools were situated in London and recruited by Real Action through opportunity sampling.

The initial eligibility criteria for the inclusion of pupils stipulated that they should be struggling readers in Year 7 who, typically, had attained less than level 4 in their KS2 English SATs. They were recruited according to the following eligibility criteria:

- First, pupils in each school with KS2 score below level 4.
- Second, pupils without KS2 score but with a Year 7 teacher assessment score for English of below level 4.
- Third, pupils with KS2 score of level 4 but with Year 7 teacher assessment of below level 4.
- Last, in cases where no Year 7 National Curriculum teacher assessments of English were available, a reading age on a standardised reading test in Year 7 of at least one year lower than chronological age was accepted as evidence of eligibility.

Parental consent was sought before pre-testing and the subsequent randomisation into treatment and control groups (see Appendix 1: Parental Consent Letter).

Intervention

Central to the Butterfly scheme are the Butterfly books. Two books were used in this trial: The Butterfly Book (Tyk, 2007), published by Civitas, and the Advanced Butterfly Reader (Tyk, unpublished). Worksheets are deliberately avoided as Irina Tyk regards them as giving the impression of being a temporary resource to be discarded at the end of a lesson. Instead, she prefers the presence of a more permanent record of progress, in the form of an exercise book, which allows both teacher and student to review the headway achieved and aids the consolidation of knowledge. Each chapter of the book is designed for an hour's teaching, the chapters building and revisiting literacy skills according to the author's view of how to advance children through the acquisition of literacy skills in the quickest possible time.

There was a teacher and a teaching assistant in each class and there were usually six to eight pupils present. The teacher and teaching assistant, recruited and paid by Real Action, had received training, initially from senior Real Action staff and later attending teacher training sessions given by the author of Butterfly. Attendance at the Saturday morning class, which is Real Action's model school, was required in order to observe experienced teachers of the Butterfly method as part of the training. Thereafter, regular Butterfly practitioner and Butterfly staff meetings served as ongoing training. The Butterfly practitioners were observed by experienced Real Action staff from time to time and received feedback on their teaching. Each lesson was intended to last an hour but this varied over the course of the intervention because some school lesson times were shorter than this. The mean time for the whole intervention delivery was 20 hours during this evaluation.

In the first lessons of the intervention, short phrases and sentences constitute the focus of the lesson, with an emphasis on phoneme blending. Longer passages are central to later lessons, especially in the Advanced Butterfly Reader. The excerpt forming the main topic of the lesson is read aloud by the teacher towards the beginning of the lesson so that the pupils appreciate the context, which is usually historical, and gain the gist of the passage. This aspect of Butterfly instruction sets it apart from many purely phonics schemes which may not spend as much time as Butterfly in overtly teaching an understanding of the global aspects of a text. A class discussion of the gist and meaning of the passage is encouraged by the teacher, who picks on individuals to ask for their contributions, in addition to pupils volunteering their own contributions. This direct questioning of pupils at random is intended to engage everyone and to check on the level of understanding of individuals. Misconceptions can be identified by the teacher and there is scope to discuss them and revisit them later in the lesson or in another lesson.

Pupils are then asked to read a section of the text and their mistakes are corrected by the teacher, who reminds the class of the spelling, pronunciation, and grammatical rules they have encountered previously as well as introducing new ones. This allows the pupils to question the teacher about literacy rules that they are not clear about, or meanings that require clarification. A period of writing, using the new vocabulary learnt, is included. This may be dictation from the passage or answering questions on it or making up some sentences of their own. Homework may be given in connection with the subject of the lesson.

An interesting feature of the Butterfly programme is the continuation of the phonics approaches of early reading into the essentially comprehension-based Advanced Butterfly Reader lessons. There is an emphasis on the new vocabulary encountered in each lesson which concentrates on the sound patterns in those new words, for example, 'imminent', 'eminent', 'persistent'. This feature is thought by the project to be unique to the scheme. These words are introduced at the start of the lesson; some of them can be identified in the text but others are not and their meanings may not be consolidated but their common sound patterns are pointed out.

Spelling is important to the programme and spelling progress is monitored. Pupils only move on to higher levels of tuition when their spelling and reading are up to levels where they can cope with the

more demanding course. The Butterfly reading programme is divided into four classes as shown in Table 2.

Table 2: Points of entry to the scheme according to reading age

| Reading age (years) | Butterfly class | <i>Butterfly Book</i> , starting at Chapter | <i>Advanced Butterfly Reader</i> , starting at Chapter |
|--|-----------------|---|--|
| 6–7 | 1 | 20–30 | |
| 7–8 | 2 | 40 | |
| 8–9 | 3 | 52 | |
| > 9 and with confirmation from informal assessments that the pupil was ready | 4 | | 1 |

Table 2 shows the approximate entry points to the Butterfly programme. The reading ages of the pupils are determined by an in-house test of single word reading. The exact entry lesson is tailored for a particular group of pupils because groupings and starting points will vary according to the needs of the pupils within a particular school. At the end of the *Butterfly Book*, a consolidation test checks that the pupils are sufficiently skilled to take on the demands of the *Advanced Butterfly Reader*. In the study, all the pupils except two remained in the class in which they started, and progressed through the books together; two pupils began in Butterfly Class 2 and found the work very easy so were moved up to Class 3.

Each lesson covered one chapter as the programme is class-orientated and didactic. A very small number of pupils who could not cope in class were taught on a one to one basis.

Pupils in the control group continued all their normal lessons as usual while pupils in the treatment group were withdrawn from lessons but not from English unless that was unavoidable. This was so that pupils in the treatment group would not fall behind in literacy through missing English lessons. However, this does introduce the possibility that the effect could be due to these pupils receiving more English teaching, as opposed to it being due to the Butterfly programme itself. Table 3 shows which lessons were missed at the various schools.

Table 3: Lessons from which children were withdrawn to attend intervention

| Phase One schools | Lessons missed |
|-------------------|---|
| School 1 | Learning for life |
| School 2 | Maths and science |
| School 3 | Various |
| School 4 | Maths, science, English |
| Phase Two schools | |
| School 5* | PE, languages, learning for life, science |
| School 6 | None, as the intervention was out of school hours |

**Phase Two school for which re-test NGRT data were reported*

Outcomes

Table 4 summarises the three reading outcomes used.

Table 4: Summary of outcome measures

| Outcome measure | Aspect of reading tested by measure | Primary or secondary outcome |
|---|---|------------------------------|
| <i>New Group Reading Test of reading comprehension</i> | Understanding the meaning of written sentences and texts | Primary |
| <i>PhAB non-word reading test</i> | Letter–sound (grapheme–phoneme) correspondences, phoneme blending, mainly regular spellings | Secondary |
| <i>Single Word Reading Test</i> | Regular and irregular word reading without a context | Secondary |

Thirteen individuals were especially recruited by Real Action to administer and mark all of the assessments used in the trial. These administrators were unknown to the children and had no knowledge of which treatment group the children belonged to. Most were postgraduate students from several London colleges.

A random selection of test papers was checked by the evaluator to make sure that they had been correctly marked.

Primary outcome measure: NGRT

The primary outcome measure was the age-standardised score from the New Group Reading Test (NGRT; Burge et al., 2010). These standardised scores had a mean of 100 and standard deviation of 15. The NGRT was chosen by the EEF with the intention of it being a common measure to evaluate the effectiveness of several catch-up interventions. It is a test of reading comprehension that includes completing stand-alone sentences by filling in the missing word, choosing the correct word to complete sentences in the first paragraph of a passage of text, and selecting the correct answer to questions about passages of increasing difficulty. It was originally published as a paper test but has since been converted into a computer adaptive test. The pre-test for the Phase One schools was the computer version. It was decided to opt for the paper version for post-testing as the digital test had thrown up unexpected inconsistencies in the way that the pupils interacted with it. For example, some pupils were observed to misunderstand or failed to remember instructions, some clicked through to later sections without completing earlier ones, and there were reported problems with some of the headphones and the computers. An advantage of using the paper version as the post-test was that it was more closely related to the EEF's stipulated outcome measure of reading comprehension as it exclusively assessed sentence and passage comprehension, whereas the computer test contained a section on phonology, which although related to reading was not considered by the evaluator as a measure of reading per se and certainly was not a measure of reading comprehension.

The computer test begins with a sentence completion question and adapts according to the response of the test taker to the question, a correct answer eliciting a more difficult question and an incorrect response prompting an easier question. This process continues until a convergence is reached between the difficulty level of the questions and the ability of the pupils to answer them. Providing that the pupil's reading age on sentence completion is greater than six-and-a-half years, they then answer the passage comprehension section. The passages are of different difficulty levels and pupils are presented with texts of greater or lesser difficulty according to their previous answers. However,

should the pupil's reading age from the sentence completion fall below the reading age threshold for advancement to the passage comprehension, they are then presented with a phonology section which requires them to recognise sounds in words. Consequently, the pupils are given scores for their sentence completion, and either the phonology section or the passage comprehension section, but not both. The Rasch ability scores for the items in the sections completed by the pupils are processed by the test proprietor, GL Assessment Ltd, creating a scale score for the section: i.e. a 'sentence comprehension' score for the sentence completion section and a 'passage comprehension scale' score for the comprehension passages. An overall common scale, the 'overall scale score' is derived for the two completed sections, which is out of a maximum of 550. This is adjusted to allow for the pupil's chronological age and the Standard Age Score (SAS) is derived. That score was used in the analysis of the present trial for the pre-test measure.

In September 2012, GL conducted a trial comparing the digital and paper versions of the NGRT on a sample of approximately 12,000 students (GL Assessment, 2012). GL concluded that the two tests were in line with each other. The paper test's mean of 100 and standard deviation of 15 was closely matched by the digital test mean of 100 and standard deviation of 16. This comparability between the computer and paper NGRTs gave some confidence to change to the paper NGRT for the post-test of the Phase One schools when the drawbacks of the digital NGRT were noticed in the pre-testing at these schools. A change to paper was also possible because the paper NGRT is a parallel form test: 3A intended as the pre-test and 3B as the post-test. The passages and items that appear in the 'A' version of the paper test are utilised in the computerised pre-test while those from the 'B' version of the paper test form the content of the computerised post-test. All the NGRTs, whether digital or paper, are multiple-choice whereby the test-taker must choose one of five options to answer each question.

In the light of the experiences described above with administering the digital NGRT in the pre-tests for the Phase One schools, the pupils at the Phase Two schools were administered the paper tests only, NGRT 3A and 3B. Each paper consists of 20 questions about what is the correct missing word from a sentence, and the test-taker must select one of five options. Afterwards, there are four comprehension passages of increasing difficulty, which are a mixture of factual and fictional texts. The first items in each passage continue this missing word format, then change to stand-alone questions about the text. The number of responses required for each passage ranges from 7 to 9, making a total of 32 answers to the passages. The maximum raw score is therefore 52. Standardisation is reported to have been conducted in 2010 (Burge et al., 2010, p.138). Raw scores were converted to age-standardised scores for statistical analysis in this study.

However, when interpreting the findings and assessing the security of the findings, it should be noted that the two versions are different assessments, conducted under different administration procedures.

Secondary Outcome Measures

The two secondary outcome measures chosen were intended to assess skills which support reading comprehension, which can be thought of as the composite skill at the apex of the pyramid of reading abilities.

PhAB

The PhAB non-word reading test was administered pre- and post-testing. Although it is recognised that Butterfly tuition departs in many ways from most contemporary phonics programmes, its early teaching, which targets letter-sound associations and sound blending, and its continued emphasis on the comparison of sounds in complex polysyllabic words, suggests that better phonological skills will accompany improved word reading and reading comprehension. In order to test this, a nonword (pseudo-word) reading assessment was taken from the Phonological Assessment Battery (PhAB; Frederickson, Frith & Reason, 1997). Non-word reading is regarded as a measure of reading via the phonological route (for example, Colheart, 2001) when letters are converted to sounds which are

blended together to form the spoken word. This test has a maximum score of 20 and ranges from single syllable non-words such as 'pim' to two-syllable non-words like 'plutskirl'. The child reads aloud the list to the independent tester. The raw test scores were converted to standardised scores by the markers for analysis.

SWRT

The reading of individual words in English involves more than straightforward regular relationships between letters and sounds: there are many words which include letter groups whose pronunciations are not as expected from these simple conversion rules. Indeed, many of the most commonly read words fall into this category, such as 'was', which is pronounced as though it ends in the letter z. It is important, then, to assess ability in the reading of single words. The test chosen for this was the Single Word Reading Test (SWRT6-16; Foster, 2007). The pupils read the list of single real words out loud to the independent tester. There are 60 words in total, ranging from high-frequency monosyllabic words such as 'play' to low-frequency words such as 'colloquial'. The raw test scores were converted to standardised age scores for analysis. SWRT has two parallel forms: SWRT1 was administered at pre-test and SWRT2 at the post-test.

Sample size

The target sample size was 400 pupils from six secondary schools, which was the number of pupils published in the initial information published by the EEF. A subsequent power calculation suggested that this was sample size was appropriate: $N=370$ ($J=6$), $ICC=0.08$, pre-post correlation = 0.71 ($R^2=0.51$), cluster size = 66 (average cluster size) at 0.05 significant level and 80% power.

The Minimum Detectable Effect Size (MDES) at randomisation for the primary outcome (New Group Reading Test) at 80% power was estimated as 0.22 based on $N=310$ (sample used for analysis) and $R^2 = 0.51$. The reported MDES was underestimated because it ignored intra-cluster correlation. The reported MDES of 0.22 was calculated under the assumption that the data was independent. However, accounting for the nested nature of pupils within schools resulted in a bigger MDES of 0.63, partly due to heterogeneity between schools.

Randomisation

A statistician at Durham University, independent of the intervention team, conducted the randomisation of the participants into groups. Randomisation for Phase One schools took place after the pre-tests in March 2013. A total of 191 eligible pupils were recruited from the four schools. Using a permuted block randomisation with a mixture of block sizes of 3, 4, and 5, 96 and 95 pupils were randomised into the control and intervention groups, respectively. The same randomisation procedure was carried out following pre-testing in the Phase Two schools (in November 2013 for School Six and in February 2014 for School Five) where 179 pupils were randomised: 89 into the treatment group and 90 to the control group. Therefore, the total number randomised in the trial was 370, comprising 184 in the treatment group and 186 in the control group.

The randomisation was carried out with the schools, as strata, with further stratification based on gender, NGRT standardised scores, PhAB standardised scores, and SWRT standardised scores.

The randomisation protocol attempted to control for variability in eligibility criteria and schools by treating them as stratification factors in the randomisation process such that both the intervention and the control groups contained approximately the same number of pupils recruited using the same eligibility criteria and from the same school. This pragmatic approach for recruiting pupils implies that the number of pupils recruited into the trial may differ between schools, but since each school will have equal representation in both the experimental and control groups, the difference in the number

of pupils recruited from the schools should have little or no effect on investigating the impact of the Butterfly Phonics programme on struggling Year 7 readers.

Analysis

The intervention effect was calculated on the basis of 'intention to treat' (ITT) in that the data were analysed in accordance with the treatment groups to which the children were originally randomly allocated. This was done for primary and secondary outcomes. The degree of adherence to ITT is debated in the literature. In this evaluation, data from all pupils was included in the analysis. There were differences in pre-tests between schools (some used a computer-delivered version and some used a pencil and paper version) differences in the intervention, and reported problems with the administration of the post-tests in the two Phase Two schools. The post-tests in School Five were re-administered and these re-test results were used instead of the problematic post-tests. The pupils in School Six, where problems were also reported with the administration of the post-tests, were not re-tested. The results from their problematic post-test were included in the ITT analysis. These difficulties are explained in more detail later in this report. All of these factors should be considered as threats to the validity of the findings. However, since the pupils were randomised to intervention or control groups within schools, issues such as reported disruption in the post-test sessions could be assumed to apply equally to the pupils in both groups. There is no evidence to suggest otherwise. Following the analysis on the basis of ITT, further analyses are reported which exclude the data from School Six.

The data were analysed using multilevel modelling to account for the clustering of pupils within schools. Although the study was such that randomisation was conducted at the level of the individual participant, it was important to account for school effects in order to obtain a robust standard error for testing the intervention effects. A zero intra-cluster correlation was assumed to work with a minimum sample size possible because of difficulties encountered in recruitment. This was corrected in the analyses to avoid false positives, but with a corresponding reduction in power. The effect sizes were calculated as recommended by Hedges (2007) for clustered data.

The primary outcome measure was the New Group Reading Test of comprehension, and the secondary measures were the Single Word Reading Test and the Phonological Assessment Battery test of non-word reading. Subgroups of pupils in receipt of free school meals (FSM) formed part of the analysis. These data were supplied by the schools.

Further sensitivity analyses were then conducted.

Process evaluation methodology

Four lessons in two of the Phase One schools were observed by the evaluator. The evaluator chose these because they provided the best opportunity of seeing lessons covering Butterfly stages 1 to 4.

A survey was sent by the evaluator to the Butterfly practitioners by email asking them a range of questions about how they thought the pupils were responding to the intervention and about the organisation and content of the intervention. The evaluator also interviewed six teachers.

Impact evaluation

Timeline

Table 5 shows the timetable for the evaluation.

Table 5: Evaluation timeline

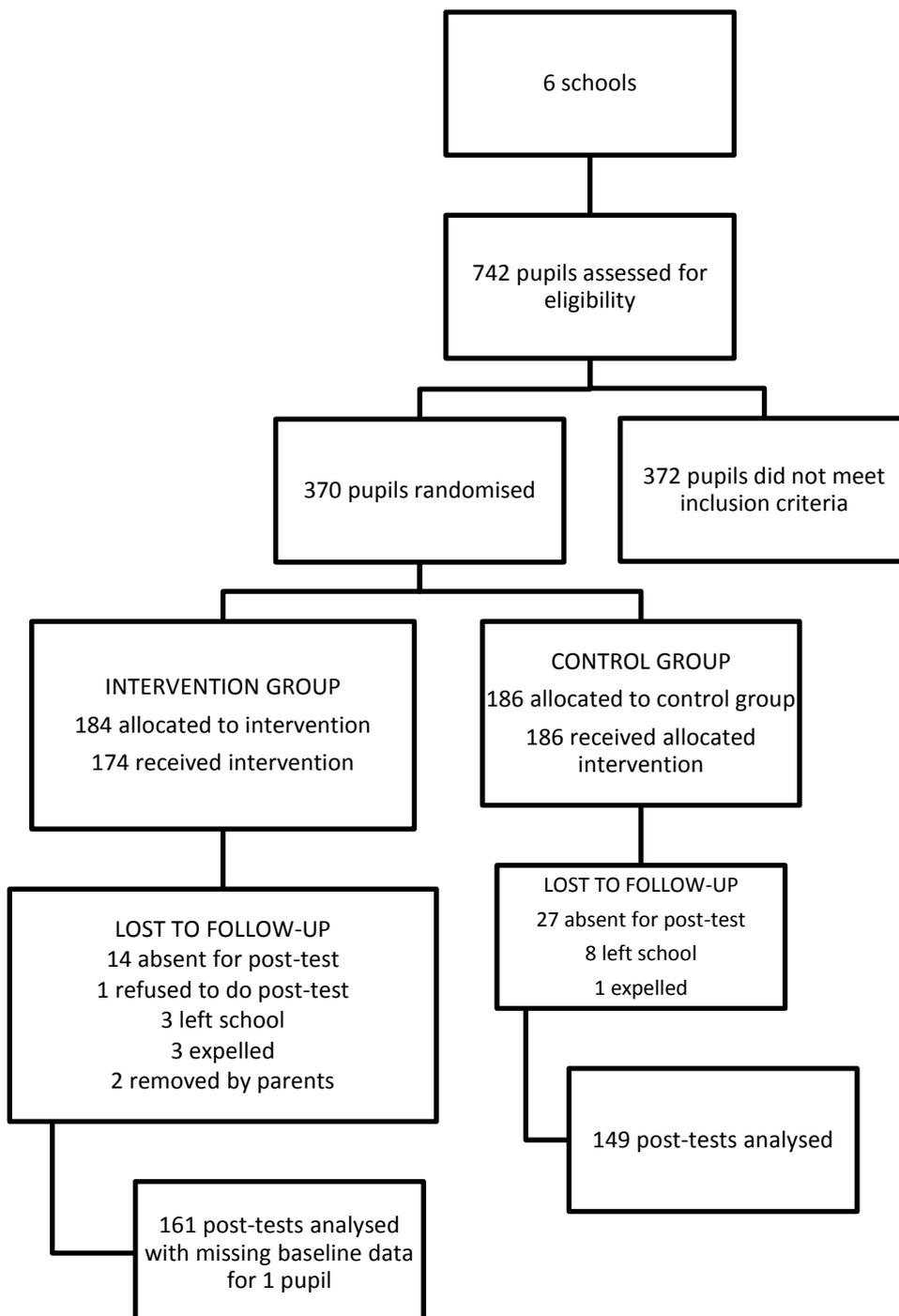
| PHASE ONE | | | |
|--|---|---------------------|------------------------------|
| Activity | Detail | Responsibility | Timescale |
| Recruitment of schools | Real Action recruited the secondary schools to be involved in the project | Real Action | December 2012 |
| Pre-intervention testing of children selected for project | New Group Reading Test A, 1 st SWRT, and PhAB | Real Action | February 2013 |
| Random allocation to intervention or control group | Allocation to intervention or control group | Durham statistician | March 2013 |
| Implementation of intervention | | Real Action | March–July 2013 |
| Post intervention assessment of pupils | NGRT B, 2 nd SWRT, and PhAB | Real Action* | July 2013 |
| PHASE TWO | | | |
| Pre-intervention testing of children selected for project in School Six | New Group Reading Test A, 1 st SWRT, and PhAB | Real Action | October/November 2013 |
| Random allocation to intervention or control group for School Six | Allocation to intervention or control group | Durham statistician | November 2013 |
| Implementation of intervention in School Six | | Real Action | November 2013–February 2014. |
| Post intervention assessment of pupils in School Six | NGRT B, 2 nd SWRT, and PhAB | Real Action* | February 2014 |
| Pre-tests of pupils in School Five | NGRT A, 1 st SWRT, and PhAB | Real Action | February 2014 |
| Random allocation to intervention or control group School Five | Allocation to intervention or control group | Durham statistician | March 2014 |
| Implementation of intervention in School Five | | Real Action | March 2014–April 2014 |
| Post intervention assessment of pupils in School Five | NGRT B, 2 nd SWRT, and PhAB | Real Action* | April 2014 |

* Real Action was responsible for recruiting a team of administrators to conduct the post-tests. Although they were blind to which children were allocated to intervention and control groups, they were not entirely independent of the intervention delivery organisation.

Participants

Figure 1 shows the numbers of children assessed for eligibility, randomised, missing from post-testing, lost from the trial for other reasons such as leaving the school, and the number of the post-tests from the NGRT.

Figure 1: Participant chart



Pupil characteristics

Table 6 presents the demographic characteristics of the pupils in terms of gender, English as a second language (EAL), pupil premium status, entitlement to free school meals (FSM), special educational needs (SEN), and ethnicity. These are based on statistics provided by the schools. The proportion of males and females are comparable between the experimental groups, which is expected because gender was used as a stratification variable during randomisation. EAL is disproportionate between the intervention and control groups with 65% (28) of pupils with a different language as their first language in the control group. However, the percentage of missing values for EAL is approximately 50% in each intervention group. Other factors except SEN are also comparable between the experimental groups.

Table 6: Baseline characteristics of the pupils in the study

| Gender | Intervention | Control | EAL | Intervention | Control |
|---------------|--------------|----------|---------------|--------------|----------|
| Female | 71(39%) | 73(39%) | No | 15(8%) | 28(15%) |
| Male | 113(61%) | 113(61%) | Yes | 123(67%) | 113(61%) |
| Missing | 0% | 0% | Missing | 46(25%) | 45(24%) |
| Pupil Premium | Intervention | Control | FSM | Intervention | Control |
| No | 55(30%) | 59(32%) | No | 89(48%) | 83(45%) |
| Yes | 67(36%) | 63(34%) | Yes | 90(49%) | 99(53%) |
| Missing | 62(34%) | 64(34%) | Missing | 5(3%) | 4(2%) |
| SEN | Intervention | Control | Ethnicity | Intervention | Control |
| No | 95(51%) | 112(60%) | White British | 141(77%) | 146(78%) |
| Yes | 71(39%) | 58(31%) | Others | 5(3%) | 3(2%) |
| Missing | 18(10%) | 16(9%) | Missing | 38(20%) | 37(20%) |

The randomisation process is expected to generate comparable average scores between the control and the intervention group whenever the sample is large enough. However, there is a possibility for differences between the control and intervention group to occur by chance.

Table 7 shows the descriptive statistics of the baseline scores by intervention groups. As expected, all the outcomes are comparable between the groups. The baseline scores would be included as a predictor variable in their respective models.

Table 7: Baseline characteristics of the pupils in the study

| Outcomes | Group | Randomised | | Final analysis | |
|----------|--------------|------------|---------------|----------------|--------------|
| | | N | Mean(SD) | N | Mean(SD) |
| NGRT | Intervention | 184 | 85.12(10.58) | 161 | 85.57(10.72) |
| | Control | 186 | 85.29(10.96) | 149 | 85.52(11.30) |
| SWRT | Intervention | 184 | 94.12(12.62) | 170 | 94.25(12.29) |
| | Control | 186 | 94.63(13.63) | 167 | 94.31(13.65) |
| PhAB | Intervention | 184 | 99.73(11.21) | 170 | 99.71(11.26) |
| | Control | 186 | 100.20(11.82) | 167 | 99.98(12.07) |

The baseline effect size at randomisation for the NGRT is -0.02 (-0.22, 0.19).

The NGRT scores were age standardised with a mean of 100 and standard deviation of 15. The mean baseline score for the pupils in this study was almost one standard deviation below what would be expected for their age.

School characteristics

Table 8 provides information about the latest Ofsted results and the urban environments of all the schools recruited. Variation is observed in certain characteristics, for example, the percentage of children entitled to free school meals and the Ofsted ratings. However, the dates of the most recent Ofsted reports varied and for School Six, where there were reported difficulties with the implementation and post-test conditions, the most recent report was three years prior to the study taking place. Changes in leadership and staff over two or three years have the potential to change the characteristics of a school. The Ofsted reports from three out of four schools in Phase One were recent, and all were positive.

Table 8: Latest Ofsted results and environmental information for all six schools recruited

| Phase | School | Percentage of children in sample with free school meals | Latest Ofsted result | Latest Ofsted result | Year of latest Ofsted report | Urban/rural classification |
|-------|--------|---|----------------------|----------------------|------------------------------|----------------------------|
| 1 | 1 | 58% | 1 | outstanding | 2013 | Urban: >10K- less sparse |
| 1 | 2 | 59% | 1 | outstanding | 2013 | Urban: >10K- less sparse |
| 1 | 3 | 44% | 2 | good | 2013 | Urban: >10K- less sparse |
| 1 | 4 | 10% | 2 | good | 2011 | Urban: >10K- less sparse |
| 2 | 5 | 30% | 4 | inadequate | 2014 | Urban: >10K- less sparse |
| 2 | 6 | 59% | 1 | outstanding | 2011 | Urban: >10K- less sparse |

Outcomes and analysis

Intervention effect for the primary outcome: NGRT

The data was analysed as intention-to-treat using all the recruited schools and the randomised pupils in each school with post-intervention scores. There was a significant effect for the intervention on the NGRT scores, with an effect size of 0.43 (0.03, 0.84) (see Table 9). Pupils in the intervention group had, on average, higher NGRT scores than those in the control group after adjusting for baseline scores. The significant effect size is equivalent to five months of reading progress. This is of educational as well as statistical significance, and larger than the effect sizes found for reading comprehension from other phonics programmes used with older struggling readers. However, given the caveats with regard to the MDES discussed in the earlier section about the sample size, although promising, this is not a secure finding.

There was no significant effect of the intervention on the SWRT scores with an effect size of 0.38 (-0.14, 0.90) or on the PhAB scores with an effect size of 0.23 (-0.03, 0.49). While the effects were positive in favour of the intervention group, they did not reach statistical significance. It might have

been expected that since Butterfly was a phonics programme, greater improvement would have been seen in scores on the SWRT and the PhAB for the intervention group.

The results from the multilevel models indicated that most of the variability in the outcomes was between pupils rather than between schools, with 8%, 16%, and 2% of the total variability respectively for the NGRT, SWRT, and PhAB scores explained by heterogeneity between schools. In other words, the attainment of the pupils within each school varied much more than the attainment of pupils in one school compared with another.

Table 9: Intention-to-treat analysis of all outcomes: intervention effect (95% CI)

| Outcomes | Group | N | Mean(SD) | Effect size (g)* | Estimate** | MDES |
|----------|--------------|-----|---------------|-------------------|-----------------|------|
| NGRT | Intervention | 161 | 87.58(10.14) | | ICC=0.08 | |
| | Control | 149 | 84.21(11.20) | 0.43(0.03, 0.84) | 3.55(1.92,5.19) | 0.63 |
| SWRT | Intervention | 170 | 99.37(12.00) | | ICC=0.16 | |
| | Control | 167 | 95.66(13.59) | 0.38(-0.14, 0.90) | 3.87(2.20,5.54) | 0.24 |
| PhAB | Intervention | 170 | 104.80(12.62) | | ICC=0.02 | |
| | Control | 167 | 102.50(12.92) | 0.23(-0.03, 0.49) | 2.46(0.23,4.69) | 0.46 |

*Calculated based on Hedges (2007).

**Estimates based on multilevel model to account for School effects; N= number of participants; Confidence Intervals (CIs) are shown in brackets in the Effect Size column. CIs represent the possible range of the effect size. The MDES was calculated as suggested by Hutchinson and Styles (2010) for cluster randomised trials. Baseline effect sizes: NGRT = -0.01(-0.59, 0.58); SWRT = -0.01(-0.45, 0.43); PhAB = -0.03 (-0.34, 0.29).

Subgroup analysis for children entitled to free school meals

A separate analysis was conducted for pupils entitled to free school meals. This group is of particular interest to the EEF, which aims to reduce the gap between pupils from disadvantaged backgrounds and those from more affluent backgrounds.

The results from the subgroup analyses for pupils entitled to FSM are presented in Table 10. There were no statistically significant intervention effects, with effect sizes of 0.16 (-0.18, 0.49) for the NGRT, 0.42 (-0.06, 0.91) for the SWRT, and 0.18 (-0.14, 0.51) for the PhAB scores. The higher effect size for the SWRT suggests that these pupils gained from some aspects of the intervention that could be considered as lower-level skills but which nevertheless comprise an important requirement of comprehension. However, pupils entitled to FSM form a smaller sample, which may account for the non-significance.

Table 10: Subgroup analysis of outcomes for children entitled to FSM only: intervention (95% CI)

| Type | Group | N | Mean(SD) | Effect size (g)* | Estimate** |
|------|--------------|----|---------------|-------------------|-------------------|
| NGRT | Intervention | 81 | 85.32(9.20) | | ICC=0.0 |
| | Control | 59 | 84.59(13.32) | 0.16(-0.18, 0.49) | 1.18(-1.35, 3.71) |
| SWRT | Intervention | 87 | 98.17(10.58) | | ICC=0.09 |
| | Control | 70 | 93.49(13.26) | 0.42(-0.06, 0.91) | 3.66(1.24, 6.09) |
| PhAB | Intervention | 87 | 104.84(12.40) | | ICC=0.00 |
| | Control | 70 | 100.96(13.19) | 0.18(-0.14, 0.51) | 1.92(-1.43, 5.28) |

*Calculated based on Hedges (2007).

**Estimates based on multilevel model to account for school effects; N = number of participants; confidence intervals (CIs) are shown in brackets in the effect size column. CIs represent the possible range of the effect size.

Sensitivity analysis I: Independent data analysis

A recommendation from the EEF is to analyse trials at the level of randomisation. Table 11 presents the results, which ignores the clustered nature of pupils within schools. The intervention was assigned at the level of the pupil but it was actually delivered to pupils who sat together in a class. There may be a teacher and class effect that would not be taken into account in an independent analysis of the outcomes. All the outcomes resulted in significant effect sizes, which are expected, given smaller standard errors from the independent data analyses compared to clustered data analyses. The effect size for NGRT was slightly higher than that of clustered data analysis.

The analysis of the SWRT scores showed why one needs to be careful with the independent data analysis for clustered data. It is an important principle that analysis should be driven by study design. The study design for this trial is equivalent to block design in experimental designs and it is important to always account for block effects. A multilevel model for Gaussian data is flexible and robust enough to reduce to independent data analysis when intra-cluster correlation is approximately zero. Lastly, there are sometimes concerns about Hedges' effect size. Table 11 also shows that Hedges' and Cohen's effect sizes are very similar from the independent data analysis. The Hedges effect size adjusts for small sample size. However, both Hedges' and Cohen's effect sizes are equivalent for large sample sizes.

Table 11: Results from independent analyses of outcomes

| Outcomes | Group | N | Mean(SD) | Effect size: Hedges | Effect size: Cohen |
|----------|--------------|-----|---------------|------------------------|-----------------------|
| NGRT | Intervention | 161 | 87.58(10.14) | | |
| | Control | 149 | 84.21(11.20) | 0.46(0.23, 0.68) | 0.46(0.23, 0.66) |
| SWRT | Intervention | 170 | 99.37(12.00) | | |
| | Control | 167 | 95.66(13.59) | 0.44(0.23,0.66) | 0.44(0.23,0.66) |
| PhAB | Intervention | 170 | 104.80(12.62) | | |
| | Control | 165 | 102.40(12.80) | 0.23(0.02,0.45) | 0.23(0.02,0.45) |

Sensitivity analysis II: On treatment analysis of primary outcome

One of the challenges and limitations of this evaluation is that the recruitment of schools was done in two stages: Phase One with four schools and Phase Two with two schools. This was a consequence of situations outside the control of the research teams; for example, in one school there was a change of headteacher between the initial recruitment of schools and the commencement of the intervention. Although all six schools were located within three London Boroughs, the schools in Phase Two experienced severe challenges during the evaluation which resulted in less support for the project. The intervention delivery and evaluation in Phase One was rigorous, but there were reported problems with the computer version of the NGRT used for the pre-test. The intervention delivery and administration of the NGRT post-tests in Phase Two were reported to be problematic: One of the schools delivered the intervention outside school hours and one of the issues was that a group of pupils could not get to the sessions on time because of their school bus transportation arrangements. The administration of the NGRT post-tests was reported to be extremely problematic due to significant behavioural difficulties displayed by the pupils. Table 12 presents the results for additional analyses for (1) Phase One schools only and (2) Phase One schools plus the Phase Two school in which the pupils re-sat the test. The results from School Six in Phase Two, where the collection of post-test NGRT data was reported to be significantly problematic, was excluded.

The effect size from the analysis of Phase One schools only is very similar to the results based on intention-to-treat analysis reported in Table 9, but with wider confidence intervals and consequently a non-significant effect size. Analysis of the sample which includes the Phase One schools and the Phase Two school with the re-sit data also produced a similar effect size but the confidence intervals were narrower than for the analysis of the data from Phase One schools only. It is clear from these analyses that intention-to-treat analysis of the study benefited from increased sample size and power but the effect sizes for the NGRT primary outcome are consistent across analyses. It is possible that the impact of problematic intervention delivery and evaluation in Phase Two cancelled out between the intervention and the control groups. However, it is not clear whether the significant effect of the intervention is robust to external influences since including Phase Two schools resulted in the same effect size as analysing Phase One schools only.

Table 12: Sensitivity analysis for impact of delivery and evaluation problems in Phase Two schools

| Outcome | Group | N | Mean(SD) | Effect size (g)* | Estimate** |
|------------------------------|--------------|-----|--------------|------------------|-----------------|
| Phase One only | Intervention | 87 | 86.08(9.30) | | ICC=0.13 |
| | Control | 86 | 82.40(11.93) | 0.43(-0.16,1.03) | 3.72(1.63,5.82) |
| Phase One + Phase Two | Intervention | 129 | 88.15(10.38) | | ICC=0.10 |
| | Control | 122 | 84.47(11.60) | 0.43(-0.04,0.90) | 3.62(1.78,5.45) |

*Calculated based on Hedges (2007).

**Estimates based on multilevel model to account for school effects; N= number of participants; confidence intervals (CIs) are shown in brackets in the effect size column. CIs represent the possible range of the effect size.

Sensitivity analysis III: Descriptive analysis of missing data

A total of 15.95% (59 out of 370) pupils had missing post-intervention scores for the NGRT. Out of the 59 pupils, 38.98% (22) pupils were from the intervention group and 61.03% (36) pupils were from the control group. The differential missing data between the intervention and the control groups may bias the significant results of the intervention if the pupils who dropped out of the control group were from the brighter pupils in the group. Likewise 7.75% (28 out of 370) pupils had missing post-intervention scores for SWRT and PhAB. Table 13 describes missing data for NGRT by baseline factors. There were more boys with missing data than girls. Pupils with EAL had higher percentages of missing data than those without EAL. There were also differential missing data for Pupil Premium status, FSM, SEN, and ethnicity. The patterns of missing data suggest that assuming missing completely at random seems unrealistic for this study. Note that the data were analysed as available cases.

Table 13: Description of attrition for NGRT by baseline characteristics (columns percentages)

| | Post-intervention | | | Post-intervention | |
|--|--------------------------|---------------------------|--|--------------------------|---------------------------|
| Gender | Missing Post-test pupils | Observed Post-test pupils | EAL | Missing Post-test pupils | Observed Post-test pupils |
| Pupils with missing gender information | 0% | 0% | Pupils with missing EAL information | 13.56%(8) | 26.69%(83) |
| Female | 30.51%(18) | 40.51%(126) | No | 15.25%(9) | 10.93%(34) |
| Male | 69.49%(41) | 59.49%(185) | Yes | 71.19%(42) | 62.38%(194) |
| Pupil Premium | | | FSM | | |
| Pupils with missing pupil premium information | 18.64%(11) | 36.98%(115) | Pupils with missing FSM information | 3.39%(2) | 2.25%(7) |
| No | 35.59%(21) | 29.90%(93) | No | 49.15%(29) | 51.45%(160) |
| Yes | 45.76%(27) | 33.12%(103) | Yes | 47.46%(28) | 46.30%(144) |
| SEN | | | Ethnicity | | |
| Pupils with missing SEN information | 6.78%(4) | 9.65%(29) | Pupils with missing ethnicity information | 11.86%(7) | 21.86%(68) |
| No | 59.32%(35) | 55.31(172) | White British | 88.14%(52) | 75.56%(235) |
| Yes | 33.90%(20) | 35.04%(109) | Others | 0% | 2.57%(7) |

Summary of results

The Intention To Treat analyses showed that there was a positive effect for the primary outcome (NGRT) with an effect size of 0.42 (CI 0.01, 0.82) across all schools. This was statistically and educationally significant, amounting to a gain of around five months' additional progress although, given the caveats with regard to the MDES discussed in the earlier section about the sample size, this is a promising but not a secure finding.

The analysis of the results from the one-to-one tests of single word reading (SWRT) and non-word reading (PhAB) across all six schools found an effect size of 0.38 for SWRT (CI -0.14, 0.90) and 0.23 (CI -0.03, 0.49) for PhAB.

With respect to the primary outcome, there were reported problems with the implementation of the intervention and the administration of the assessments in the Phase Two schools, which should be considered when interpreting the scores. Also, there were reported difficulties with the computer version of the pre-test in the Phase One schools, which should be borne in mind.

Indications from sensitivity analyses were that when Phases One and Two were analysed separately, the effect sizes were of a similar magnitude to the analysis of the full sample but, with smaller sample sizes, did not reach statistical significance. When the potential influence of missing data was explored, the patterns suggested that it was unrealistic to make an assumption that this occurred at random.

Cost

Materials

The *Butterfly Book*, which is used as the course book by each child in Butterfly classes 1, 2, and 3, costs £9.50 and is published by Civitas from where it may be bought directly. It is also available from retail book websites. The *Butterfly Grammar Book* costs £9.50, has the same publisher, and is available from the same outlets. During the intervention, it was used only occasionally, when grammatical weaknesses were identified.

The *Advanced Butterfly Reader* and *Junior Butterfly Reader* cost the project £10.00 each. They are printed by Printmeit but have not yet been published. It is possible to pre-order printed versions of them through Real Action. The *Advanced Butterfly Reader* was used in this trial because it was most suited to the age group being studied.

Each pupil was provided with an exercise book and a pencil by the Real Action project.

The project therefore cost nothing for the schools to be involved in, apart from staff liaison time with the project, providing lists, and setting up the computer tests. Schools varied greatly in how much teacher time they were prepared to give the project in a supporting role, helping with discipline problems if necessary, and ensuring pupils' attendance at intervention lessons and test administration sessions.

Training

The Butterfly practitioners who taught the intervention were postgraduate students recruited from London University colleges, as described earlier, as well as local people with experience of working with pupils, some of whom were qualified teachers. Initial training was delivered by the Education Director of Real Action and the Project Manager for Butterfly Phonics. The practitioners were then required to observe classes at the Butterfly Saturday Reading School which is the project's model school and works as their training centre. Once complete, they received specialist training by Irina Tyk, the author of the *Butterfly Book*. When they started teaching the intervention, the practitioners were observed regularly by the Education Director and the Project Manager and given feedback during monthly teacher meetings. These teacher meetings also functioned as mini training sessions and were an opportunity for teachers to share experiences. Butterfly practitioners were paid £20 per hour. All Butterfly classes are delivered to pupils by two adults: either one teacher (at £20.00 per hour) with one teaching assistant (at £15.00 per hour) or two teachers (each at £20.00 per hour). Therefore the cost of the Butterfly teaching staff for a one-hour class was £35 or £40.

The specialist training from Irina Tyk cost £600 for half a day.

Assuming eight pupils in a class, the following costs have been estimated (fewer or more pupils in a class will alter these figures). The cost per pupil is calculated as £21 for course books and stationery, and £87.50 for the teaching of 20 hours of the intervention (20 hours is the average number of hours that a pupil in Phases One and Two received the intervention). The £87.50 is based on the £35 an hour cost for the project of a Butterfly practitioner and a Butterfly teaching assistant. The overall estimated cost per child in these circumstances is therefore £108.50, based on information provided by the project.

If an experienced teacher employed by a school were to teach alongside a teaching assistant, the teaching cost would be higher by about £10 an hour per lesson, increasing this estimated teacher cost to approximately £112.50 for a 20-hour course of instruction, and the overall cost rises to £133.50 per child.

The Butterfly practitioners undergo extensive training, which includes periods of observing experienced practitioners as well as specialist training. It is possible that experienced teachers may not require such intensive training but they would need some training in the Butterfly method. It is not known what form such training would take once it was rolled out nationally, but training courses and instructional videos are likely to be involved. An amount to cover these contingencies would have to be factored in to the costs for a school.

In conclusion, if pupils were to be taught in groups of eight by fully trained Butterfly practitioners, the estimated cost is £108.50 per pupil; if taught by experienced school teachers with teaching assistants, the estimated cost is £133.50, not including specialist training in the Butterfly method.

Process evaluation

The task for the delivery team was to transfer an intervention which began life as an extracurricular Saturday morning activity to one that could be effectively delivered in disadvantaged inner city secondary schools to pupils with, sometimes, challenging behaviour and for whom Butterfly would not be perceived in the same way as a voluntary weekend pursuit. The demands of delivering an intervention were also accompanied by the difficulties of conducting testing in a controlled environment, which is required by the rigorous standards of academic research.

To reiterate, the process evaluation consisted of observations of four lessons in two of the Phase One schools and interviews with six Butterfly practitioners all conducted by the independent evaluator. Additionally, Butterfly practitioners were invited to complete a survey about how they thought the pupils had responded to the intervention and about the organisation and content of the intervention.

Implementation: findings from observations

The focus of the stage 1 lesson was on encouraging decoding, on getting the pupils to read aloud in a meaningful way, and to blend phonemes together. Some instruction on cursive writing came up as a topic which needed to be addressed. The teacher reacted to the needs of the pupils and adjusted to areas covered according to the pupils' progress. There was some discussion of grammar and learning of grammatical rules. As with all the lessons observed, pupils were asked by name to answer specific questions to check their understanding of the main learning points of the lesson. Pupils also spontaneously offered ideas and examples to the teacher.

In the stages 2 and 3 lessons observed, there was an increasing shift away from time spent on explicit teaching of decoding towards the more holistic skills required to appreciate the story or passage that was the subject of the lesson. Families of words sharing similar sounds remained an important aspect of the tuition, and the course book continued to determine the structure of the lesson. Repetition of key points was in evidence in these lessons and reading aloud by the pupils was a central activity.

Pupils were encouraged to express key concepts learned in their own words, to give examples from their lives to illustrate the new vocabulary they had learned, thereby demonstrating their understanding of these new words and consolidating them in memory. In stages 3 and 4, the new vocabulary being learned was very abstract. Discussions of the words 'cynical' and 'gullible' in these classes were thought by the evaluator to be at a high level: the pupils clearly understood what these concepts meant, as they illustrated them with incidents when their acquaintances had shown these qualities, and some lighthearted disputes broke out between them as to the fairness of attaching these labels to their friends. It was the opinion of the evaluator that the teacher of the stage 4 class showed great skill in guiding the class discussion, involving everyone and reinforcing and elaborating on the new vocabulary encountered.

The seating of the pupils in rows and the teacher picking out individuals to ask specific questions did not appear to perturb them, and it seemed to aid the teacher in bringing the pupils' focus back after any incident, such as a clever comment by a pupil, which threatened to sidetrack proceedings. There was humour shown appropriately and respectfully by the teachers and pupils observed.

The evaluator noticed that some of the rooms provided were not appropriate. Two of the rooms were particularly inappropriate for the intervention. One room was a computer room which was cramped and full of equipment that was not apposite to the lesson and next to a very noisy corridor. Another was a science laboratory where the door to the technician's room was constantly in use. A fire drill was conducted during this lesson yet no warning had been given to the Butterfly practitioners. Apparently, it was usual at one of the Phase One schools to be ejected from the allocated classroom

and to have to locate to alternative accommodation. These matters illustrate the difficulties faced by projects attempting to deliver an intervention as visitors to a school and their dependence on the co-operation of the school staff to be able to deliver those lessons.

An example of the support that was given to the intervention by school staff was observed by the evaluator when the project asked a form teacher for help with a disruptive child. The teacher spoke to the pupil, who returned to the class and was co-operative for the remainder of the lesson. Discipline issues did, however, trouble the Butterfly lessons in one of the Phase One schools and the Real Action project called in a behaviour management consultant to train the Butterfly practitioners on how to deal with bad behaviour in the classroom. Thereafter, strategies were employed such as lining up outside the classroom and waiting until the pupils were calm before allowing them to enter the room. This permitted the intervention to continue with fewer disruptions to the lessons than before. If pupils were still not responding to the class situation, the project taught them on a one-to-one basis, as a last resort. The project manager told the evaluator that no pupil was expelled from any Butterfly classes.

An aspect of the lesson delivery which may have helped maintain an orderly atmosphere was that a teaching assistant from Real Action was always in the room to support the Butterfly practitioner. The role of the teaching assistant was to: make sure that the correct books were at the lesson; to give names of missing pupils to the members of the team whose role it was to collect up the pupils who were not in the classroom; keep the register; to make sure that all the pupils were looking at the right page of the book; and to assist pupils to keep up with the lesson, so that no pupil would panic or become frustrated at losing their place.

Implementation: findings from survey and interviews

The survey was distributed by Real Action but returned directly to the evaluator. It was explained that the survey was anonymous. The interviews were unstructured and informal, to complement the mainly structured nature of the survey. Most of this information came from Phase One of the study.

The results from the survey are presented in Table 14.

Table 14: Process evaluation survey responses

| Questionnaire Item | Score: mean*, based on six respondents | Response category tended towards |
|--|--|----------------------------------|
| 1. The progress built into the intervention was just right for the children. | 3.8 | Agree |
| 2. All the children were placed in the correct groups for their abilities. | 4.3 | Agree |
| 3. There was evidence that the children were doing more reading on their own outside of lessons. | 2.2 | Disagree |
| 4. I think that most of the children grew in confidence as a result of taking part. | 4.3 | Agree |
| 5. I would have liked more training in how to deliver the intervention. | 3.2 | Neutral |
| 6. Feedback from children about the intervention was positive. | 3 | Neutral |
| 7. Teachers in the school supported us in encouraging good behaviour from the children. | 3.2 | Neutral |
| 8. The method of picking a particular child to respond to a question never seemed to perturb the children. | 4.3 | Agree |
| 9. My training had prepared me well to deal with any bad behaviour from the pupils. | 2.7 | Neutral |
| 10. The lessons would have been improved had there been more interaction between the pupils. | 2.7 | Neutral |
| 11. Children only contributed when made to do so. | 2.2 | Disagree |
| 12. Children generally completed the homework they were given. | 2 | Disagree |
| 13. The lessons stretched the children. | 3.8 | Agree |
| 14. There was a lot of lateness amongst the pupils. | 4.3 | Agree |
| 15. The classrooms provided made it difficult to effectively deliver the intervention. | 3.5 | Agree |
| 16. Lessons were too difficult for the children. | 2 | Disagree |
| 17. There were some children who did not seem to benefit from the intervention. | 2.7 | Neutral |
| 18. The lessons were too long to maintain the children's sustained attention. | 2 | Disagree |
| 19. The reading materials could have been better suited to the children's abilities. | 2.3 | Disagree |
| 20. Reading materials could have been more appropriate to the interests of the children. | 3.7 | Agree |
| 21. Some children should have been excluded from lessons as they distracted others. | 4 | Agree |
| 22. I enjoyed my time teaching on the project. | 4.3 | Agree |
| 23. Children's spelling seemed to improve. | 4 | Agree |
| 24. I noticed the children's reading improve. | 4.5 | Agree |
| 25. The didactic approach worked very well. | 4.5 | Agree |
| 26. Teaching outside of school hours was a problem due to pupil absenteeism. | 5** | Strongly agree |

*Based on a five-point Likert scale where 5 was strongly agree, 4 agree, 3 neither agree nor disagree, 2 disagree, and 1 strongly disagree.

**From one response from a School Six teacher.

It should be noted that a sample of six for the survey was small and so the mean scores need to be interpreted with caution. They should be interpreted as one element of information alongside the interviews and observations.

The teachers surveyed in these ways considered their training to be good, and that it prepared them well for teaching the intervention. They approved of the formal arrangement of pupils seated in rows, and thought that the direct way of asking pupils questions in class was not threatening to the pupils and helped to engage them and to make them feel included. The rooms provided by the schools were criticised by the Butterfly teachers and the sometimes noisy conditions were also the main complaints. Some voiced concerns that they found some of the pupils' unruly behaviour in class difficult to deal with. A response from a Phase Two teacher was that the lessons should be held during school hours, because in School Six (where the children from the school bus were very late or too late for lessons), there was a lot of disruption caused by lateness and that it would not occur if the lessons had been during school hours. This lateness meant that these children missed much of the lesson and would be behind the rest of the class. In addition to these examples, Table 15 summarises the barriers and suggests ways of avoiding these in future.

Table 15: Barriers to implementation and evaluation

| Occurrence | Phase in which this problem was experienced | Consequence of problem | Conditions or measures that can be taken for success |
|--|---|--|--|
| 1. Withdrawal of the largest of the original participant schools prior to pre-testing | Phase One | Necessitated a Phase Two to reach the suggested power for the study | Commitment of school to the research timetable. Continued obligation should be observed by the school after a change of key liaison teacher or headteacher |
| 2. Lack of school staff support in post-tests | Phase Two | The presence of teachers, senior teachers and head teachers at the <i>NGRT</i> group pre-tests at these two schools contrasted with the absence of teachers when it came to invigilating the post-tests. Behavioural difficulties were reported in two schools in the post-tests | School support with tests |
| 3. Schools arranging tests for the last lesson of the last day of term | Phase Two | At Schools Five, the <i>NGRT</i> group post-test was timetabled for the last afternoon of term when pupils were reported to be disengaged. Similar disengagement was reported for the pupils in School Six where the post-test was initially scheduled for the last lesson of the last full day of term. | Schools accepting a duty to timetable tests responsibly |
| 4. Poor behaviour in tests | Phase Two | In the group <i>NGRT</i> post-tests at both Schools Five and Six, the children were reported as displaying boisterous, loud, uncontrolled behaviour | Pupil behaviour in tests should not be an issue if assessments are timetabled sensibly and enough teachers are in attendance |

| | | | |
|---|-------------------------|--|---|
| | | during the tests | |
| 5. Poor behaviour in class | Phase One | This was reported to be a major problem in the largest of the Phase One schools | The backing of the school staff can encourage good behaviour in intervention lessons. Advice was taken from a consultant on countering poor classroom behaviour. If attempts to maintain an appropriate environment failed, one-to one teaching of problem children was adopted in this study in preference to total exclusion from lessons |
| 6. Late attendance at lessons | Phase One and Phase Two | Children were frequently late for their intervention lessons | The project team always sent along back-up staff whose job it was to find the children who were not present at the start of the lesson and bring them along to the intervention |
| 7. Absence from lessons | Phase One and Phase Two | At the largest school in Phase One, there was a higher degree of absenteeism than at the other schools in the phase School Six in Phase Two was the only school where the intervention could not be delivered within school hours. Absenteeism here was higher than elsewhere. This absenteeism was systematic in that it overwhelmingly affected the children who depended on the school bus. The bus was perennially late, so that those children travelling on it were unable to attend the intervention. Requests to timetable the lessons during the school day were declined by the new headteacher | Co-operation of teachers in encouraging pupils to attend Where the intervention has to be conducted outside of school hours, more contact is needed between the project and pupils' parents in order to promote regular attendance. If children are physically not available to attend, because they are on the school bus for example, then it requires the agreement of the school to alter the time of the intervention |
| 8. Teachers withholding pupils from intervention lessons | Phase One and Phase Two | On occasion, some teachers refused to allow children to miss their lessons to attend the intervention | Commitment to the intervention needs to be shared by all the teachers in the school |

| | | | |
|---|--------------------------|---|--|
| 9. Poor allocation of rooms for the intervention | Phase One and Phase Two. | Noisy conditions so that the children could not hear the lesson properly | Proper allocation of suitable teaching rooms, avoiding noisy or specialised rooms, such as science laboratories |
| | | Ejection of intervention classes from rooms and searching around for alternative classrooms with the resulting shortened lesson duration and interruption to children's concentration | Respect for the visiting teachers of the intervention by the teachers in the school and recognition that the intervention staff need suitable facilities |

The Real Action team faced many problems to the successful delivery of lessons but they put into place a system of self-critical monitoring, including observations of lessons by the project's management staff and continuing training and regular meetings of the Butterfly practitioners, in an effort to adapt to the difficulties encountered and share experience and knowledge so as to act pre-emptively where possible. They attempted to deal with poor behaviour in the classroom by engaging the services of a consultant. The claim by the team that no pupil was excluded from lessons suggests that they learnt and successfully responded to this issue. Without classroom discipline, the application of Butterfly to the school environment would not have been possible. It involves a formal arrangement of the teacher facing rows of pupils and targeted questioning of pupils, which some might find unfamiliar. However, the evaluator's lesson observations, and the survey responses from the Butterfly practitioners, suggest that the pupils were comfortable in this environment. The extent to which it contributed to maintaining the pupils' concentration cannot be established from this study but the author of the Butterfly method, Irina Tyk, contends that it does help.

Other barriers to the delivery of the intervention could be seen to be centred around the different levels of support offered by school staff to aid the smooth running of the project, for example: the provision of suitable rooms; sending or withholding pupils from the intervention lessons; encouraging pupils to behave well; and timetabling the intervention at realistic times so that the children could attend. An example of problems with the latter occurred in School Six where 22% of the intervention group did not attend at all and 50% were regularly late or very late as a result of lessons being timetabled before or after school when the pupils had to be on the school bus.

Fidelity

There was a variation in the total number of hours over which the intervention was delivered in the schools. This was because of the different lesson durations at schools and fitting in the intervention around existing school commitments. However, all the schools could be said to have received sufficient hours from the team to constitute a full training course. The intended number of hours was 40: 2 hours a week for 20 weeks. The intervention in Phase One schools ranged between 10 and 12 weeks' duration. One of the Phase Two schools (School Six) received 14 weeks of intervention and the other Phase Two school received only four weeks. Schools received a mean of 12, 17, 23, or 31 hours of intervention in Phase One, and either 16 (School Five) or 12 hours (School Six) in Phase Two. This figure of 12 hours was for the children who attended at least one lesson. Ten out of the 45 in the intervention group failed to attend a single lesson. When all 45 pupils are included, the mean attendance drops to 9 hours.

The lessons that pupils missed in order to attend the intervention are shown in Table 3 above. The lessons included: learning for life, maths, science, English, PE, languages, and other subjects. In School Six, no lesson was missed as the intervention was held outside school hours.

There were concerns about absences at the Phase Two school lessons, particularly at School Six where absences were systematic, overwhelmingly affecting the pupils who relied on the school bus. The lessons in this school were conducted outside school hours, whereas all the rest took place in school time.

The fidelity with which the Butterfly teachers delivered the scheme within lessons was high, according to the judgement of the evaluator who observed lessons.

Outcomes

The perceived outcomes of the intervention, by the Butterfly practitioners who were surveyed, matched the quantitative findings of this trial: that pupils' comprehension and word reading improved. There were other more qualitative factors which were thought to have been improved in the pupils, such as a perceived improvement in their confidence to express themselves in class.

Six Butterfly practitioners returned their email questionnaire to the evaluator. They agreed on the following three negative outcomes: there was a lot of lateness by the pupils; pupils did not do their homework; and they saw no evidence of the pupils doing much reading outside of lessons. The positive outcomes that the teachers agreed with were: the programme was well suited to the pupils; that they had been placed in the right groups for tuition; pupils had grown in confidence over the period of the intervention; that picking individual pupils to answer questions in class did not perturb the pupils; the teachers had enjoyed teaching the course; they had noticed the pupils' reading improve; the didactic teaching method worked well; and the four teachers from Phase One who responded said that the teachers in the schools had been supportive in encouraging the good behaviour of their pupils.

Qualitative responses from these five practitioners were that there needed to be more co-ordination with the schools, especially over disciplinary matters, and that class sizes should be kept to less than eight. The one respondent from the Phase Two schools said that at School Six lateness had disrupted lessons and that pupils arriving late fell behind the rest of the class, so that the intervention should be conducted within school hours.

There were challenges that arose as a result of poor classroom behaviour. Some pupils were very disaffected by school and so it is not known if this indiscipline was due to the intervention or was displayed in the school in general. The team did remark that some pupils, whom they had been told by staff were difficult, were well behaved in the intervention sessions. The reported timing and lack of preparation of pupils for the NGRT post-test could possibly have had negative consequences. For example, one of the schools timetabled the post-tests for the last afternoon of term when the pupils might not have been at their most motivated. However, these circumstances applied similarly to the control group as well as the intervention group.

Formative findings

How could the intervention be improved? This process evaluation was intended to be a 'light touch' one and so did not interrogate the various aspects of the intervention to establish which areas could be improved. However, one possibility is for experienced, qualified teachers to deliver the intervention. Slavin et al. (2011) found that reading interventions were optimally effective when delivered by professional teachers. The lesson observations found that all the lessons delivered by the Butterfly practitioners were of a high standard, which implies that their training was effective, but the kinds of skills demanded in discussing text, as is essential to Butterfly as a catch-up programme,

are likely to be more finely honed by, say, an experienced teacher of English literature than by a newcomer to teaching. On the other hand, if school teachers were to carry out the intervention, they would probably benefit from receiving some training about the specific aspects of Butterfly Phonics.

It is recommended that in order to get the most out of Butterfly Phonics, an efficient way of training teachers in its method should be devised, if it were to be rolled out nationally.

Control group activity

No evaluation of control group activity was included in this project. It is possible that some aspects of the intervention leaked into their school lessons, but it is less likely in this rigidly structured programme with its formal classroom routines than perhaps in other interventions, and the school teachers were not present during the intervention lessons run by the team. Additionally, the intervention group was withdrawn from a mixture of lessons (for example, maths, science, P.E.) and so the different subject teachers for the control group were unlikely to have focused on direct phonics instruction with the control group in those lessons. We did not evaluate the control group as part of this evaluation but we consider it likely that they followed the usual subject lessons and did not receive additional literacy input.

No reports reached the evaluator of the pupils in the control group displaying any negative behaviours associated with their not yet having received the intervention. There were many pupils in these schools who were on report or spent some of their time under close supervision because of their poor behaviour, but their distribution between the treatment and control groups is unknown.

Conclusion

Limitations

The generalizability of the results is limited for reasons common to many studies, such as any peculiarities in the sample (in this instance the participating schools were chosen by opportunity sampling and limited to a particular small area in London), as well as limitations due to unexpected situations which occur within a particular study that may not occur in another piece of research. In this study, a number of factors reduce confidence in the findings.

The project delivery team was experienced in delivering the teaching of the intervention in Saturday morning schools but had no experience of randomised controlled trials. The 'light touch' process evaluation model meant that it was not the evaluator's role to be present at testing and so the delivery team was very dependent on the test administrators who were recruited for the study to ensure the kind of test conditions for the project tests that the school teachers themselves would demand in examinations. Support to ensure appropriate conditions was reported to be provided by the schools in Phase One but not in Phase Two.

The administration of the pre- and post-tests requires an appropriate environment. Lack of distraction would be even more crucial for poor readers. It was unexpected that School Five, which participated in Phase Two, organised the post-tests for the last period of the last day of term when the pupils' motivation and concentration were likely to be low. The EEF judged a re-test session in this school at the start of the following term to be appropriate, with the evaluator as an observer. The evaluator considered that this re-test was carried out in appropriate conditions. As the pupils had not received Butterfly lessons during the holiday, two revision lessons were suggested by the evaluator as an attempt to compensate for the hiatus between the ending of the intervention and the repeat post-test. These took place and the pupils sat the same paper as before in the disrupted post-test. The re-test results were used as the post-test data in the analyses. This arrangement represents a departure from the protocol and differs from the other schools, and should be borne in mind when interpreting the generalisability and quality of the findings. Difficulties with the administration of the post-test in School Six were encountered, as detailed in Appendix 2. It would be an improvement on the present situation if the independent invigilators were trained and employed by the EEF or the evaluator.

Another limitation is the reported difficulties associated with the pupils' completion of the computer version of the *NGRT* test, which was used as the pre-test in Phase One. A change was made to paper tests for the post-tests, but although the test proprietor suggested that the scores from the two formats could be considered to be equivalent, the possibility exists that they are not.

Inconsistencies of implementation arose whereby the intervention was delivered within the school day for five of the six schools and outside of school hours for School Six.

The Butterfly intervention was delivered at a different time to regular English lessons and was conducted in small groups. The pupils in the control group did not receive additional time or small-group work and therefore it is not possible to say whether the impact on the post-test scores was an effect of the Butterfly intervention alone or a combination of the intervention and additional time spent on literacy and small-group work.

Interpretation

The brief theory of change diagram in Figure 2w summarises the project.

Figure 2: Theory of change diagram



The intention-to-treat analysis of the primary outcome measure suggested a significant and positive effect for Butterfly Phonics; however, the level of attrition in addition to the reported implementation and test administration issues should be borne in mind. Additionally, when clustering was taken into account, the positive impact was promising but inconclusive. The secondary outcome measures suggested a beneficial impact, however, these results were not statistically significant.

The results of the primary and secondary outcome measures suggest that the various elements of Butterfly are contributing to the impact and that it is not simply a straightforward phonics scheme. There are essential vocabulary and reading comprehension skills which are taught in Butterfly, in addition to the phonological skills, grapheme–phoneme mapping, and phoneme blending skills, which would qualify it as a phonics course: it could be called a phonics ‘plus more’ reading intervention.

The mixed skills promoted by Butterfly seem to be well suited to the needs of the sample of Year 7 pupils in this study. It would be expected from a developmental viewpoint that phonological skills are most important at less skilled stages of reading. The later concentration on improved comprehension skills may help with access to the Year 7 curriculum. The full range of reading abilities of those eligible to receive the intervention has been catered for by the programme. No exclusions were made for participants with very low reading ages or those with special needs. Hence, there were many pupils with a variety of special needs who progressed with the help of the intervention. It has features that make it particularly interesting from the viewpoint of dyslexia. Children with dyslexia have been found to have a problem with the first sound in a word (King, Wood & Faulkner, 2008). Butterfly tuition encourages the combination of the first consonant in a word with the vowel that follows it. This new sound unit will be longer than the brief initial phoneme (for example, the sound ‘ca’ as opposed to the phoneme /c/) and it poses the question of whether a longer sound at the start of a word might make it easier to hear and/or to store in memory. Throughout the course, Butterfly continues to promote the use of larger units of sound and spelling in words than those of the smallest units of words, that of phonemes and graphemes. The Advanced Butterfly Reader concentrates on families of polysyllabic

words and their commonalities. It would be useful to discover if the Butterfly approach helps children with the phonological difficulties associated with dyslexia (for example, Snowling, 2000) by boosting their sound memories for words. It could, instead, help children identify common spelling patterns which would assist their sight reading. The likelihood is that Butterfly improves both. Writing and spelling are also integral parts of the intervention but were not assessed in the trial. Impact on these skills could be investigated in the future.

The comprehension skills that Butterfly promotes are varied. There is stress on understanding and acquiring new vocabulary from the text passages, as well as appreciating gist, grammatical constructions and inferences, and other essential elements of reading comprehension. Its teaching, with the use of extracts from classic texts from English literature, is in tune with the current National Curriculum for English. Teaching the interpretation of these passages is carried out in conjunction with a focus on key words and the sound/spelling families to which they belong. This is a novel mixture of teaching methods, but one to which the pupils appear to respond, judging from the results of this trial.

When an intervention is the subject of a randomised controlled trial, it is tested in its entirety. It is not possible to extract an aspect from it and choose it as the element to which its success or failure can be attributed. So it is with Butterfly. In addition to the novel methods described above, there are other factors such as the formal style of teaching, involving children seated in rows facing the teacher. There are two factors, over and above the intended intervention, that could have contributed to the difference in post-test scores between the intervention and control groups. The intervention was delivered outside English lessons in small groups and this additional input could have had a positive effect over and above the Butterfly Phonics programme itself. Additionally, for the treatment group, Hawthorne effects cannot be ruled out, whereby a change in circumstances may improve a pupil's test performance. Similarly, a type of placebo effect whereby the act of being chosen for an intervention group rather than the treatment per se may motivate pupils to improve in tests.

As for Butterfly's wider application, absenteeism at the one school where the intervention had to take place outside of school hours suggests that incorporation into the school day would be beneficial. A high standard of training for the staff responsible for its delivery would also be important.

Future research and publications

Given the reported difficulties with this study, it would be useful to conduct a further randomised controlled trial with schools that have been selected well in advance of the commencement of the trial and whose commitments and responsibilities to supporting the testing as well as the intervention were firmly established. In this way, the problems encountered in this trial would hopefully be avoided. In addition, if the pupils in the intervention group are to receive the intervention in addition to their usual allocation of English lessons, it is recommended that the control group receives an equivalent amount of small-group time to be spent on literacy.

It is clear that the pupils in the intervention group received more literacy instruction than the pupils in the control group; up to 2 hours extra per week for up to 20 weeks. They were taught in small groups, which meant that the teachers could monitor their understanding and respond more readily to the pace of learning of the group. The observations of lessons suggested that pupils responded well to the format of instruction and the content of the programme when delivered in this way. Further research could tease out the level of impact of each of these factors, for example, by comparing the following intervention arms with a control group:

- An additional 40 hours of literacy lessons for the whole class rather than withdrawing the struggling readers and teaching them in a smaller class;
- An additional 40 hours of literacy lessons to a small class of struggling readers following the methods usually used in each school;

- 40 hours of using the Butterfly books with the small group of struggling readers at the same time as the rest of the class followed their usual literacy lessons. In other words, the intervention group receive the Butterfly intervention instead of their usual literacy lessons.

Future research could also be carried out in different contexts to investigate its success with different age groups. A study in primary schools would gauge whether Butterfly could be an effective reading catch-up programme at an earlier stage in children's school careers. Equally, Butterfly could be tested as a programme for beginning readers. It is already used in this way at the independent school where its author is headteacher. With its reduced demand on memory load, when blending sounds by using larger word subunits as early as possible (for example, 'spr' instead of /s/ /p/ /r/), it might work well for children with moderate learning difficulties; this could be investigated. Adults with reading difficulties could also be studied, as Butterfly appears from this trial to promote a quick progression. One of the participating schools conducted the trial over only four weeks; such a short intensive course might be appealing to adults with literacy difficulties whose patience with themselves may be short after years of frustration with their attempts at learning to read.

References

- Burge, B., Styles, B., Brzyska, B., Cooper, L., Shamsan, Y., Saltini, F. and Twist, L. (2010) *New Group Reading Test (NGRT)*. Third Edition. London: GL Assessment.
- Coltheart, M. (1978) 'Lexical access in simple reading tasks'. In G.Underwood (Ed.), *Strategies of information processing*. London: Academic Press.
- Coltheart, M., Rastle, K., Perry, C., Langdon, R., & Ziegler, J. (2001) 'DRC: A dual route cascaded, model of visual word recognition and reading aloud'. *Psychological Review*, 108(1): 204-256.
- Ehri L.C., Nunes S.R., Stahl S.A. & Willows D.M. (2001) 'Systematic phonics instruction helps students learn to read: evidence from the national reading panel's metaanalysis'. *Review of Educational Research* 71, 393–447.
- Foster, H. (2007) *Single Word Reading Test 6-16*. London: GL Assessment Limited.
- Frederickson, N., Frith, U. & Reason, R. (1997) *The Phonological Assessment Battery*. London: GL Assessment.
- Galuschka, K., Ise, E., Krick, K. and Schulte-Körne, G. (2014) *Effectiveness of Treatment Approaches for Children and Adolescents with Reading Disabilities: A Meta-Analysis of Randomized Controlled Trials*. PLoS ONE 9(2): e89900. doi:10.1371/journal.pone.0089900
- GL Assessment Limited (2012). *NGRT Digital*. Retrieved 12th June 2014 from <http://www.gl-assessment.co.uk/products/new-group-reading-test-digital>
- Hedges, L.V. (2007) 'Effect sizes in cluster-randomized designs'. *Journal of Educational and Behavioral Statistics*, 32 (4): 341-370.
- Hutchinson, D. and Styles, B. (2010) *A guide to running randomised controlled trials for educational researchers*. Slough: NFER [Online] Available from: <http://www.nfer.ac.uk/nfer/publications/RCT01/RCT01.pdf>
- King, B., Wood, C. and Faulkner, D. (2008). 'Sensitivity to visual and auditory stimuli in children with developmental dyslexia'. *Dyslexia*, 14(2) pp. 116–141.
- National Reading Panel (2000) *Teaching children to read: An evidence based assessment of the scientific research literature on reading and its implications for reading instruction* (NIH Publication No. 00–4769). Bethesda, MD: National Institute of Child Health and Human Development, Washington, DC: US Government Printing Office.
- Slavin, R.E., Lake, C., Davis, S., and Madden, N. (2011) 'Effective programs for struggling readers: best-evidence synthesis'. *Educational Research Review*, 6: 1-26.
- Snowling, M.J. (2000) *Dyslexia*. Second edition. Malden, MA: Blackwell.
- Suggate, S.P. (2010) 'Why what we teach depends on when: Grade and reading intervention modality moderate effect size'. *Developmental Psychology*, 46: 1556–1579.
- Tyk, I. (2007) *The Butterfly Book: A Reading and Writing Course*. London: Civitas.
- Wedderburn, A. (2011) *Learning to Read, Reading to Learn*. Retrieved 20th June 2014 from <http://www.realaction.org.uk>

Appendix 1: Parental Consent Letter



Dear Parent,

READING BETTER...

Our classes aim to help your child to read better.

The classes:

- will start after the February half term OR
- in September 2013. We will tell you at the family open day when your child will have their classes.

In the classes your child

- will be taught reading using our Butterfly Books.
- will be given 2 hours per week for 20 weeks. Classes will be taught by our staff who have received specialist training and will be CRB checked.

The Butterfly Project:

We want your child to read better. What we will do:

- give reading tests to your child before they start the classes to see their level in comprehension and single word reading.
- give them the same tests again after they have had all their classes to see their progress. The results will give us information for research being done by The Education Endowment Foundation.

Children who don't get classes now will be given their classes in September 2013. This is to inform the research to see the difference in reading levels between children given the classes in February and those not given them. This shows the effectiveness of the way we are teaching children how to read. Selection for classes is done randomly by an independent organisation.

Confidentiality and Feedback:

- All information about your child will be confidential.
- The Education Endowment Foundation may keep information for research purposes.
- You can take your child off the programme at any time.
- You can withdraw the data about your child at any time.
- We will share the test results with the school and you may know your child's results at the end of the programme.
- After it has finished, Real Action will supply information about the success of the project to parents.

TEA AND CAKES...

Come and have some tea and cakes with us.

We look forward to seeing you!

ASK US MORE ABOUT IT.....You can:

- call us anytime on 0208 960 2065 or
- e-mail us at **admin@realaction.org.uk** or
- come and see us anytime at The Learning Store in Mozart Street W10 4LA

If you **DON'T** want your child to come to classes:

tell us or give the form below to the school.



REAL ACTION Reading Lessons

Name of Child _____

Class _____

I do **NOT** wish my child to take part in these lessons.

Signature of parent /guardian _____ Date _____

Please return to _____ by _____

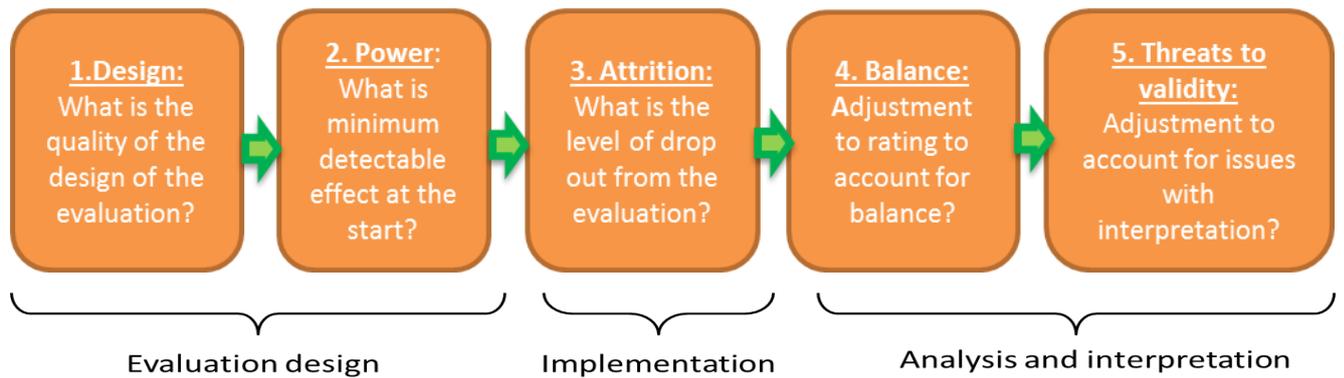
Appendix 2: Details of Reported Disruption to Test Conditions at Phase Two ‘School Six’

School Six informed the project team by email at 3:40pm on the afternoon of Wednesday 9th April that the post-test had been arranged for 2:30pm the following day, Thursday 10th April, which was the last period of the penultimate day of term. Written and verbal accounts of the test administration by those present relate that the pupils arrived in unruly groups without an accompanying teacher, that they hammered on the door and burst through into the hall, threw the test booklets over the floor and that a stampede ensued, which caused the project team members and independent invigilator to fear for their personal safety. The arrival of a teacher subdued the situation sufficiently for the pupils who were present to sit down and begin the test. One-third of the 90 pupils were absent. Members of the project team reported that a short while into the test, the teacher departed to attend a meeting and that the indiscipline of the pupils flared once more. A member of the team looked for the teacher and when they returned, some degree of order was restored. However, most of the pupils had given up on the test and there was an air of unrest, with persistent talking by the pupils. It was reported that when the teacher approached one group, they would fall silence, only for another group to start a conversation at the other end of the hall, so that the distraction continued to the end of the test. It was reported that some pupils did appear to want to do well and that at least one girl asked the adults in the room if they could stop the other pupils from talking. The project team reported that some pupils adopted an aggressive demeanour towards them during the test, including foul and abusive language, and that the experience left them shaken. The project manager, who was present as an observer, wrote a note at the end of the test, reporting the pupils’ behaviour, and gave it to the teacher who was present to pass on to the head of Year 7: it is not known if that teacher received it.

The delayed start of the test and the disturbances where no proper testing was possible meant that the testing session (according to those who were present) lasted only about half an hour, when it should have been twice as long.

The EEF carefully considered the evidence about the post-testing at School Six and came to the conclusion that the children were unlikely to respond in any more of a positive way to a re-test, and concerns were voiced about the safety of the invigilators and assisting team at any such re-test, given the reported apparent lack of support from the school staff. The EEF therefore decided not to re-test at School Six. The Real Action team has told the evaluator that they have reluctantly ceased contact with the school in the light of the abusive behaviour they were subjected to by the children, and cannot offer them any further teaching. The evaluator understands that, since their work began in 1999, this is the first time that they have not felt able to continue to work with a group of children.

Appendix 3: Padlock rating



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

The final security rating for this trial is 0 . This means that the conclusions have very low security.

The trial was designed as an efficacy trial and could achieve a maximum of 5 . This was a well conducted trial, with low levels of attrition. However, due to the variability between schools, the trial was considerably underpowered. Therefore, despite balance at baseline and only limited threats to validity from issues with the testing, the overall padlock rating is 0 .

Appendix 4: Cost Rating

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

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

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