An evaluation of Fresh Start as a catch-up intervention: A trial conducted by teachers

Stephen Gorard, Nadia Siddiqui and Beng Huat See
School of Education, Durham University
s.a.c.gorard@durham.ac.uk

Abstract

This paper describes a randomised controlled trial conducted with 10 secondary schools in England to evaluate the impact and feasibility of Fresh Start as an intervention to help new entrants with low prior literacy. Fresh Start is a synthetic phonics programme for small groups of pupils, here implemented three times per week over 22 weeks. The intervention was led by the schools themselves and assessed in co-operation with the independent evaluators. A group of 433 Year 7 pupils (first year of secondary school) were identified by schools as having literacy attainment below 'secure' KS2 Level 4, and individually randomised to a treatment group or a waiting-list control. The pupils were assessed via GL's New Group Reading Test. Missing data at pre- and post-test amounted to 3% of the total. The overall 'effect' size in terms of gain scores from pre- to post-test was +0.24, and this was repeated in a sub-group analysis involving only FSM-eligible pupils. However, there was some imbalance between the two groups at the outset, and this must lead to a slight caution about the findings, and to some doubts about whether one or more schools unwittingly subverted the randomisation. Other than this, the aggregated trial shows that schools can conduct evaluations of their own interventions with firm guidance from experts, and under favourable conditions such as individual randomisation and lack of extended involvement by developers with a conflict of interests.

Introduction: literacy and the transition to secondary schools

Pupils struggling to achieve their 'expected' reading level at primary school would generally find it difficult to access the full secondary curriculum, since literacy is such a fundamental gateway for further study. Around 25 per cent of primary school pupils in England do not achieve the expected level 4 at Key Stage 2 (KS2) in English, and over 750 mainstream primary schools do not achieve the minimal expected level in terms of pupils reaching level 4 (Department for Education 2013). Of course, these expectations may have been set too high (or too low) but this is still evidence of a potential problem for secondary schools and their lowest attaining pupils.

The Department for Education (DfE) and others have implemented a range of policies to try and reach national targets for reading at KS2, including the National Literacy Strategy and more recently giving extra funding to schools in proportion to their level of disadvantaged (mostly free-school meal eligible) pupils. This 'pupil premium' money must be used to help raise the attainment of these pupils who are more likely to be struggling than their more advantaged peers. This help might involve hiring extra assistants, or purchasing resources or interventions to support children at risk of not reaching an appropriate level of literacy. Such interventions and approaches need to be both effective, and practical in the sense of being able to be integrated within the wider school timetable and curriculum. This paper contributes

to these steps by presenting the results of an evaluation of a phonics intervention called Fresh Start, when used for children on arrival at their secondary with assessments of KS2 level 4c or below in English.

Can schools conduct their own robust evaluations?

In addition to consideration of the effectiveness of Fresh Start as a catch-up intervention, this paper also addresses the question of whether schools and teachers are able to conduct robust evaluations, such as randomised controlled trials, of their own interventions. If they can then the general quality of available evidence in education could improve. The cost of robust evaluations could be reduced, making them more feasible across a range of situations. And, perhaps most promisingly, a series of large, ongoing, almost automatic trials could be conducted nationally, similar to those espoused by Goldacre (2012) for medical GP treatments.

In 2013, three secondary school clusters independently applied to the Educational Endowment Foundation (EEF) to work on Fresh Start, and try it out in their own schools. Individually, the applications were judged by EEF to be not of a sufficient scale to enable the intervention to be evaluated robustly. However, the scale was sufficient for an aggregated efficacy trial if the schools co-operated, and so the separate applications for funding were treated as one.

The authors of this paper were then selected by EEF to act as independent evaluators for the trial. However, their role was to be rather unusual - advising the schools on how to conduct research, including how to create fair comparator groups, observing the schools as they conducted both the intervention and their own evaluations, aggregating the results to give an overall efficacy finding, and assessing the extent to which schools are indeed capable of conducting their own trials.

There was no direct developer (of Fresh Start) involvement in any of the proposals. However the developers were used by schools to provide training. Testing was done via an external company (GL), and the evaluators were provided with independent access to the results.

A summary of evidence on the effectiveness of the general phonics approach

Schools under pressure to show improvement in their literacy goals have various approaches available, but it is not always clear to them which approaches to teaching literacy will be the most effective, with national policies liable to change over time (House of Commons 2005). From 1998, England had a National Literacy Strategy for primary schools, based at least partly on an ‘analytic approach’ (Select Committee on Education and Skills 2005). Children first learnt the alphabet, and words were then introduced to illustrate the sounds associated with each letter. Subsequently, children used the whole word as the context to work out the sound of each letter (Johnston and Watson 2005). The Rose (2006) report was based on a large-scale review of this approach. It suggested that there was no evidence that the strategy had been effective, and so proposed changes. This led to a greater use of phonics (defined here as the teaching of letter-sound correspondences in an organised, regular, explicit and

sequenced manner), and teaching reading through phonics. However, evidence of the effectiveness of phonics approaches is not entirely secure.

Two 'experiments' seem have had a considerable influence on the move to phonics after the Rose report. One involved 304 first year primary school children, allocated to three groups to receive different literacy interventions (Johnston and Watson 2004). But the groups were not randomly allocated – indeed they were not even matched and the most disadvantaged pupils received the synthetic phonics intervention (p.12). This makes the finding insecure as the greater progress made by the phonics group could be because, starting from a lower base, more improvement was possible in a short time. The other study in the same report allocated the groups better, via matching, but involved only 92 first year pupils and these were divided into three groups – synthetic phonics, analytic phonics, and analytic phonics with phonological awareness training. This is a small study with only about 30 pupils per group, and it was ended early for ethical reasons. Other commentators have suggested the implementation of the three conditions may have led to bias (Wyse and Goswami 2008).

A subsequent systematic review of phonics trials concluded that systematic phonics teaching was more effective than not using phonics or using phonics non-systematically (Torgerson et al. 2006). But importantly, it excluded the first Johnston and Watson 'experiment' (above) on the basis of their lack of a valid control (but included the second because the authors personally communicated that the cases had been randomised even though this contradicts what they said in the original paper). Systematic phonics involves teaching letter and sound links in a clear sequence. The overall effect size for systematic phonics compared to other approaches was estimated as +0.27, and the results largely confirmed a previous review by Ehri et al. (2001). However, a combination of relatively few trials, and poor evidence or poorly-reported methods in some existing trials meant that the result cannot be seen as definitive, especially in relation to exactly how phonics should be taught.

There have been four RCTs of the phonics-based 'Sound Partners', a similar approach to Fresh Start, intended for below average readers in K-3. Overall, these suggest a benefit for beginning readers in terms of letter recognition, fluency and comprehension. But there is no discernible benefit for reading 'achievement' (What Works Clearing House 2010).

Johnston et al. (2012) conducted two follow-up analyses using some of their cases from their assessment of the impact of synthetic phonics teaching in Scotland (see above) by comparing them with cases from England, unmatched on prior attainment. The cases being compared have therefore been neither randomly selected nor randomly allocated. They can have no standard error by definition. Despite this, the authors analysed their results using techniques such as analysis of variance that are based on standard errors (and they similarly cited p-values erroneously in their 20o4 study). This is a common mistake in the field (Gorard 2015). Such approaches to analysis take no account of sample quality or attrition (Lipsey et al. 2012), being predicated on complete random samples of a kind never encountered in real-life research (Berk and Freedman 2001). They do not make sense anyway (Carver 1978), are routinely misinterpreted (Watts 1991), and can lead to serious mistakes for policy and practice (Falk and Greenbaum 1995).

Two generic meta-analyses of reading interventions for struggling readers, like the previous reviews described above, reported that phonics was the most promising approach. Galuschka et al. (2014) found 22 randomised controlled trials of phonics interventions. McArthur et al.

(2012) found 12 studies using a variety of evaluation designs. Their conclusion was that teaching phonics was more effective than other methods for reading accuracy, but not for spelling or reading fluency.

More recently, two studies in England have come to opposing conclusions. An evaluation of 'Rapid Phonics'- a popular synthetic phonics programme used as a catch-up literacy intervention for pupils moving to secondary school – found no benefit. In fact the pupils in the treatment group did worse than those in the control (King and Kasim 2015). However, 'Butterfly Phonics' was found to be effective for pupils who were not achieving expected reading levels in the transition stage from primary to secondary school (Merrell and Kasim 2015). The evidence is thus mixed. The generic phonics approach may be effective for some measures and in some contexts but not others.

A summary of evidence on the effectiveness of Fresh Start

The specific Fresh Start (FS) intervention used for the new evaluation described in this paper is a 'systematic synthetic approach' (sometimes known as rml2). The individual letters are sounded out within words, and these sounds then blended to form the pronunciation of the word, and so to 'read' it. The 44 basic sounds, used as building blocks, are taught first rather than the letters. When writing, the combination of sounds is said aloud and then converted to letters and written on the page (Brooks 2003). The intervention is described more fully later.

Fresh Start is produced by Read Write Inc., whose literacy programmes are cited by OFSTED (2010) as used by the 'best' performing schools. However, as with phonics more generally, the prior evidence related to FS is far from strong, again because studies have too often been small, non-randomised, with high dropout or poorly reported. There has been no clear report of its cost-effectiveness, which would be important information for schools planning to use their pupil premium funding on such an intervention.

A study by Brooks et al. (2003) intending to evaluate FS for use with low attaining pupils at Key Stage 3 (KS3) only managed to retain 30% of its initial 500 pupils, making any claims for the success of the intervention weak (Gorard 2014). One local authority in England adopted FS in all of its secondary schools for pupils not meeting or likely to meet expected levels of literacy (Lanes et al. 2005). The impact was never evaluated properly. The 'evaluation' report shows that the approach was popular and considered effective by teaching staff, but the only evidence of impact came from before-and after-figures in one school with no true comparator. A later summary of reading interventions for KS3 included these studies of FS, reporting effect sizes of +0.25 to +0.34 for reading comprehension (Brooks 2007). There was no benefit for spelling, perhaps precisely because of the phonic nature of FS. All of the samples were small, with one study having only 29 cases, and there was high dropout, with studies not clearly reporting the comparator groups, the allocation of cases, and whether the groups were equivalent at the outset.

Overall, therefore, the direct evidence for Fresh Start is limited, and mostly from small-scale studies not randomising pupils to treatments. What follows is a description of the largest, perhaps the only, randomised controlled trial of FS so far in the UK.

Evaluation methods

Three heads of school clusters in different regions of England independently proposed conducting FS as an intervention. The funders (EEF) felt that each cluster was too small for a feasible efficacy trial. So they suggested that each cluster run their own intervention but that they should be constrained to use the same evaluation design, and the results should aggregated by an independent evaluator. The independent evaluator would also train the school research leads, advise on design, oversee implementation and conduct a process evaluation from start to finish. This is therefore a school-led trial, and in addition to the substantive results it provided evidence on whether schools and teachers can conduct robust research with advice.

Study design

This new evaluation of FS was conducted as a randomised controlled trial, in the school year 2013/14. A group of eligible low-attaining pupils was selected in each school and each pupil was individually randomised to receive FS in the first term or to wait until later in the school year. The eligible pupils who had to wait for FS carried on with their English lessons as normal, and took part in any pre-existing interventions. This evaluation design, in which the intervention is not rationed but provided for all eligible pupils, has a few disadvantages (most notably there is no permanent control group that can be followed up in later years). But it is ethical, and tends to reduce any possible demoralisation caused by an individual knowing which group they have been allocated to (Gorard 2013). And because each school had both treatment and control pupils in the first phase this reduced the chance of schools wanting to drop out of the study.

At the start, the independent evaluators held a one-day workshop for the heads and research leads in each of the 10 schools. This covered the craft of conducting a randomised controlled trial. A key issue was how to randomise the eligible pupils into the two groups, making the allocation fair and without bias. A second workshop was conducted by the evaluators with the cluster and school leads before the post-test phase. This explained the conduct of the test process (the need for 'blinding' or at least observation to prevent bias), and how to calculate and interpret the results for each cluster. The schools reported that the workshops were very useful. Attendance was high. And the evaluators also found them informative about the kinds of challenges teachers faced when conducting research projects in their schools.

The sample and allocation to groups

The study was based on 10 schools in three clusters from Harlow, Holderness and Telford. All were urban, mixed, reasonably diverse secondary schools. A target group of 433 eligible pupils from the fresh intake to Year 7 was identified by the schools, based on them having KS2 scores at or below level 4c in English (considered to represent less than secure literacy for transfer to secondary school). On the basis of the pre-test when in year 7, all but 29 of the 423 pupils were reading at National Curriculum level 4c or below, and 237 were reading at level 3c or below.

The background data and identity of these pupils was sent to the independent evaluators. The clusters then conducted the individual randomisation based on techniques learnt from the workshop, such as using a well-shuffled pack of cards or a computer pseudo-random number

generator. The two aggregated groups were reasonably well-balanced in terms of pupil background characteristics at the outset (Table 1).

Table 1 - Percentage of participants in each group by sex, FSM-eligibility, SEN, EAL and ethnicity

	Treatment	Control
Male	54	50
Female	46	50
FSM-eligible	34	38
Special educational need	25	22
English as an additional language	4	4
Non-White UK	15	8

N=433

By the end of the intervention, a total of 14 pupils provided either no pre- or no post-test (3% overall attrition). Of these, eight were in the treatment group and six in the control. Reasons for absence included that they had left the country, were long-term ill, suspended, or no longer attended the school and did not provide details of their new school. The average score of the tests (either pre- or post) taken by pupils dropping out in each group was 251, suggesting that the dropout was balanced. There were therefore 419 pupils in the final analyses, of which 215 were in the treatment and 204 in the control group.

Outcome measures

The pre-test agreed by the schools, funders, and independent evaluators was GL Assessment's New Group Reading Test A (NGRTA). The post-test was New Group Reading Test B (NGRTB). The age-appropriate level was selected, suitable for the pupils aged 10 to 13 (Year 5 to Year 8) (http://www.gl-assessment.co.uk/products/new-group-reading-test/test-detail). Both versions of the test include 20 sentence completions, 4 short passages for context comprehension, and reading comprehension. The areas of assessment are as follows:

- Vocabulary
- Grammatical knowledge
- Inference skills
- Ability to recognise
- Authorial intent
- Deduction skills

This evaluation used the digital version of the test, which has no time limit but is estimated to take no more than 45 minutes to complete. The difficulty level of the test adapts depending on the first few answers by each pupil making it adjust the difficulty accordingly (which also reduces the chances of cheating among pupils).

The headline findings are based on the 'overall reading score' provided by the software. This is used because our prior work has shown that there is floor effect created by the minimum achievable score when using the 'standardised age scores' (Gorard et al. 2015). However, for comparison the results are also presented in terms of standardised age scores.

Impact analysis

The difference in scores between the pre- and post-test was used to create a gain score. The average difference in gain scores between the two groups was used to create an 'effect' size (Hedges' g). All of the pupils initially identified as eligible and randomised to groups were used in this analysis, as far as possible, regardless of the time actually spent on the intervention (for example, some pupils moved schools before the post-test, and were followed up wherever possible). Further analyses used the same 'effect' size calculation for sub-groups such as FSM-eligible pupils, boys and pupils reported by their schools as having special educational needs.

The 'number needed to disturb the finding' of the effect size (NNTD) can be calculated by creating a counterfactual score consisting of the mean gain score for the treatment group, and adding its standard deviation. The number of such scores that can be added to the control group before the effect size disappears is the 'number needed to disturb the finding' (Gorard and Gorard 2015). The treatment counterfactual is used here because the control group is slightly smaller (and so is easier to 'disturb'). This NNTD is a measure of the robustness of the finding both in terms of chance and in terms of any bias caused by attrition.

Process evaluation

The role of the independent evaluators was to monitor the school leads and assess whether the protocol of the intervention was being followed. The results also relate to how well the trial was being conducted by the schools themselves. The purpose of the process evaluation was to assess fidelity to treatment and to collect the views of participants to get a sense of whether there was any resentment or resistance to the programme. It also helps to identify potential barriers to implementation and the key features necessary for successful implementation.

The training, implementation and monitoring of the intervention as well as data collection were all managed by the cluster leads with the assistance of staff members. As part of the process evaluation the independent evaluators attended the initial training sessions as participant observers to evaluate the quality of training. They also observed the delivery of the intervention and the conduct of the post-test. Observations of classroom delivery were carried out once at the beginning of the intervention and once towards the end of the trial to assess progress and changes in pupils' and teachers' behaviour. Altogether 10 observation visits were made. In addition another three visits were made to observe the post-testing.

Face-to-face interviews with staff, pupils and project leaders were arranged during these visits. These interviews were intentionally kept informal and as open as possible to allow respondents to speak freely without the constraints of a structured protocol. Observations of staff training and delivery of the programme were as simple and integrated and non-intrusive as possible.

The intervention

Fresh Start is a reading and writing programme. The aim of FS is to provide additional support for children who have missed earlier opportunities of learning to fill the learning

gaps. Therefore, FS is tailored to get the pupils to catch up with their peers so that they can participate in mainstream literacy activities without falling further behind.

FS modules are arranged to be completed in 33 weeks. The modules are graded according to reading age. It is recommended that one session be conducted each week with each session lasting an hour. However, in this trial FS was conducted three times a week for one hour each over 22 weeks.

The complete FS resource pack includes module sets, assessment charts, magnetic sound cards, speed sound cards, sound charts and poster, lesson plans, pronunciation DVD for teacher, teacher training books and handbooks to support the delivery of FS.

The key features of FS include:

Initial screening
The programme begins with an initial assessment of pupil's phonics and word recognition. Pupils are assessed individually by teachers. The assessment involves pupils reading aloud or sounding out the letters and words to the teacher.

After the initial screening pupils are put into four groups according to the initial scores. This is to ensure homogeneity within the groups, which is believed to encourage progress. Depending of the individual pupils' progress, teachers may also provide additional 20-minute regular one-to-one sessions.

Phonic lessons
The phonic lessons involve the systematic teaching of 44 sounds in English. Using a sound chart and Speed Sound Cards the teacher introduces the sounds and graphemes. Pupils practise blending the sounds through Sound Talk (sounding-out) by repeating the sounds after the teacher. This process is assisted using a number of learning aids such as picture cards, picture books, Fred puppet and talking fingers. Nonsense words are also used for pupils to practise independent blending of sounds. Pupils practise writing, although the letters are not mentioned by their names.

Assessment and modules
In the third week, pupils are withdrawn from the regular literacy class and put into groups for the phonics lessons. Pupils begin with an entry-test to determine the starting lesson level.

There are 33 modules altogether and pupils may start with different modules depending on their entry level. The modules are graded in six sets and each set consists of a pack of five booklets with different titles. Pupils are assessed after completion of each set to see if they are ready for the next module.

Each booklet in a module contains nine reading activities based on a text which could be fiction or non-fiction, but with appropriate to the interest of Year 7 (aged 11-12). Words in the text are colour-coded. Green words are decodable, while red words are non-decodable. There are also Challenge Words. These are new or unfamiliar words. Each story is read three times in one FS session to achieve fluency and decoding. Each session begins with teachers explaining the new vocabulary and reading aloud. This is to generate interest and discussion.

Pupils then read it as partners. Pupils read the story a third time with expression and a storyteller voice.

Each module contains writing activities including spelling checks, mime punctuation, building sentences for composition, editing work and composition writing.

Classroom management
Inherent in the FS programme is a set of classroom management routines. These are used in each FS session and form the way teachers communicate with pupils. This is how it works:

1) When the teacher speaks, they point towards themselves and say, 'my turn'.
2) When pupil repeats what the teacher sounds out, the teacher points to the pupils and says, 'your turn'.
3) If the teacher wants the pupils to practise with their partners, they say, 'Talk to your partner'.
4) If the teacher wants to stop an activity, they raise their hand or count down from 5 to 1.

In addition to receiving the resource pack, teachers also took part in a two-day training workshop provided by Read Write Inc. The latter, along with the schools, decided on all aspects of the intervention, independently of the evaluators.

The impact results

The main outcome measure was the post-test score on NGRT. However, because the two groups had differing average scores at the outset, gain scores from NGRTA to NGRTB were used to assess the impact (Table 2). The control group had higher pre- and post-test scores than the treatment group (the pre-test 'effect' size was -0.36, and the post-test 'effect' size was -0.19). Using the gain scores, the intervention showed a small positive impact on reading comprehension (+0.24).

Table 2 - Gain scores for reading – overall reading scores

	N	NGRTA pre-test	Standard deviation	NGRTB post-test	Standard deviation	Gain score	Standard deviation	'Effect' size
Intervention	215	251.8	65.4	279.5	59.9	27.5	47.7	+0.24
Control	204	274.2	58.2	290.6	53.3	16.7	42.1	-
Total	419	262.7	62.9	284.9	57.0	22.2	45.3	-

The same analysis was conducted using the standardised age scores, with the same finding in terms of effect size (Table 3).

Table 3 - Gain scores for reading – standardised age scores

	N	NGRTA pre-test	Standard deviation	NGRTB post-test	Standard deviation	Gain score	Standard deviation	'Effect' size
Intervention	215	82.3	13.1	86.0	13.3	3.64	8.15	+0.24
Control	204	86.8	12.2	88.2	12.3	1.48	9.72	-
Total	419	84.5	12.9	87.1	12.9	2.59	9.00	-

(Note: very similar results were obtained using the standardised age scores in place of the raw overall reading scores for all of the analyses that follow)

To help assess if the results are due to chance or attrition, we take the mean post-test score for the treatment group, and add one standard deviation to create a score that would be counterfactual to the control group. It is estimated that over 35 such scores would need to be added to the existing control group to eliminate the effect size reported in Tables 2 and 3. This is considerably more than the level of dropout, suggesting that the results are reasonably robust and not due to bias caused by missing data.

Because of the initial imbalance between groups, we have to treat the results as more tentative than if the randomisation had led to more equal average scores. As the treatment group has a lower pre-test score than the control group, the difference in outcomes could be due to regression to the mean, since the pupils with the lowest scores have the most opportunity for rapid improvement.

To help determine if this was the case we repeated the analysis with only those pupils whose pre-test score was below 262.7, the pre-test mean in Table 3. Although, as noted in the discussion of allocation to groups, more of the low-scoring pupils are in the intervention group, the reading scores of the two low-scoring groups are well balanced at the outset (Table 4). The results showed that the low-scoring pupils in the treatment group actually had a slightly lower average gain score than the low-scoring pupils in the control group, even though the 'effect' was small (-0.04). This suggests that overall the result was very unlikely to be due to regression towards the mean.

Table 4 - Gain scores for reading, low-scoring pupils

	N	NGRTA pre-test	Standard deviation	Gain score	Standard deviation	'Effect' size
Intervention	116	204.4	47.9	40.4	52.3	-0.04
Control	67	204.1	42.3	42.9	48.3	-
Total	183	204.3	45.8	41.0	50.7	-

To assess the relative impact of the intervention on FSM pupils and boys, further analyses were carried out. The intervention is equally effective with FSM-eligible pupils (Table 5), but slightly less effective for boys (Table 6). These sub-group results do not have the force of a trial, but they do suggest that the intervention is equally effective for poorer children, and therefore could be used to reduce the pupil premium attainment gap.

Table 5 - Gain scores for reading, FSM-eligible pupils

	N	NGRTA pre-test	Standard deviation	NGRTB post-tesst	Standard deviation	Gain score	Standard deviation	'Effect' size
Intervention	52	243.1	54.2	271.6	59.8	28.5	49.2	+0.24
Control	52	264.7	58.4	283.2	52.5	19.5	38.5	-
Total	104	253.9	57.1	277.5	56.2	24.0	44.2	-

Table 6 – Gain scores for reading, boys

	N	NGRTA	Standard	NGRTB	Standard	Gain	Standard	'Effect'

		pre-test	deviation	post-test	deviation	score	deviation	size
Intervention	140	243.1	69.2	269.6	64.3	26.0	49.0	+0.17
Control	121	265.7	64.6	282.3	56.5	17.0	41.3	-
Total	261	253.6	67.9	275.5	61.0	21.8	45.8	-

Several prior studies of the impact of Fresh Start have dealt primarily with pupils labelled as having special educational needs (SEN). It is therefore interesting to see how promising the intervention was for such pupils in the present study (Table 7). The SEN-labelled pupils in the treatment did make greater gains than the control, but not by as much the overall group of all pupils.

Table 7 – Gain scores for reading, SEN pupils only

	N	NGRTA pre-test	Standard deviation	NGRTB post-test	Standard deviation	Gain score	Standard deviation	'Effect' size
Intervention	117	231.3	61.0	259.9	59.3	28.3	40.0	+0.14
Control	81	246.7	66.5	268.8	55.2	22.6	41.3	-
Total	198	237.6	63.6	263.6	57.7	26.0	40.5	-

In summary, although they cannot be considered definitive due to the initial imbalance between groups, these are promising results overall.

Formative evaluation outcomes

FS is a teacher-led highly structured intervention with prescriptive classroom management and teacher-pupil communication techniques. To implement the intervention successfully requires training in the use of these communication routines and the strategies suggested in the teacher's handbook.

In this study, teachers attended a two-day training workshop given by experienced and professional trainers. In all, 65 school teachers and teaching assistants attended the FS training, the majority of whom had no previous experience of using FS. The classroom management strategies are inspired by reception years and primary school teaching. FS teachers are expected use body language, praises and dramatisation to get pupils' attention. As such, the secondary school teachers found these strategies and management styles more difficult to adopt.

One school leader reported that parents expressed initial concern that the intervention was too low level for secondary school pupils. Consequently a meeting was organised to explain to parents the purpose of the intervention. There was also some initial resistance from pupils who felt that the activities were too patronising. Private discussions with a few individual pupils eventually convinced them of the purpose of the activities. There was also resistance from some of the more experienced teachers. In fact a head of English in one school refused to take part, and found a substitute.

Teachers reported difficulties with pupil grouping because of different test scores giving different results. Teachers had the KS2 scores in English, pre-test NGRT-A results, and the screening test scores that came with the FS package. The grouping could be quite different

depending on which test scores teachers used. In this study, pupils were put into ability groups according to the FS screening test.

Delivery of the intervention

Classroom observations suggested that the FS teaching strategy was quite effective in keeping pupils engaged. In each session the trained teacher was supported by one trained TA. The pupils received a lot of support and individual attention which they would not otherwise have had. The attendance records showed that pupils were attending the sessions very regularly. Many pupils reported that they preferred coming for these sessions rather than attending regular classes.

Generally pupils said they liked the structured phonics lessons. However, some pupils felt that some of the things they were learning were very basic, or something that had already been taught in their primary school. Some felt that the reading tasks were too low level. However, the teachers often did not agree with these pupils' views

Teachers mostly believed that the highly structured and prescriptive nature of the programme was crucial to effective implementation, but others felt that it was not necessary for every session. Some teachers also felt that it was confusing to pupils not to use the traditional names of the letters.

Teachers thought that although FS helped improve pupils' reading and decoding abilities it had no impact on their writing skills. As writing was not assessed in this study we could not confirm teachers' apprehension.

One of the other aspects commonly reported by teachers was that FS provided positive results for pupils who have learning difficulties (although the results of the impact evaluation did not support this). Teachers reported positive changes in pupils' confidence in reading and class participation. A number of teachers reported that pupils who were usually quiet or disruptive in class were more focused and confident in the small group FS sessions.

Control group activity

Control group pupils continued their usual classroom activities. Although FS materials were not made available to the pupils in the control group or in regular classes, teachers reported that it was hard not to use some of the FS strategies in the regular classes. There is therefore some potential of diffusion. This is the problem with randomisation within schools. However, all teachers interviewed said they were aware of the research trial and had made a conscious effort not to bring FS concepts deliberately into their teaching. If this problem did occur it would anyway only tend to dampen or reduce the estimated effect size of the intervention.

Discussion

Limitations of the study

One of the limitations of the study is the lack of generalisability of the findings since the schools were not representative of the larger population in the area. This study also cannot

establish whether the impact could be sustained once the pupils are integrated into the mainstream classes. There was no opportunity to follow the pupils, although the schools themselves may continue to monitor their progress. This, however, was not the purpose of the study.

There was also a possibility of a Hawthorne effect since the trial was initiated by the schools themselves who were very keen on the programme, and there was no placebo control.

The schools themselves conducted the randomisation and there is some evidence that this might have been subverted to some extent as one cluster asked to re-randomise their pupils after discussion with the developers during training. It is possible that they did do at the request of the developers. The initial imbalance between the groups may also be an indication that some schools were somehow putting forward their weaker pupils to receive the intervention first.

Since the NGRT test was a test of reading comprehension, the impact of the programme on spelling and writing could not be ascertained.

The promise of Fresh Start as an approach

The findings suggest that FS as a programme for supporting pupils at risk of failing has some promise. It was well-received by teachers and pupils.

The trial itself was well-conducted with no school dropout and a relatively low pupil attrition rate of 3%. The strength of the evidence was medium with an effect size of +0.24, judged by EEF to be equivalent to three months' of progress in reading age over one year. However, these findings have to be treated with slight caution given the initial imbalance between groups.

The cost of running the programme for 22 weeks was largely the cost of the module booklets which is estimated at £95.90 per child, plus the cost of training the teachers received and the TA time.

Schools conducting their own evaluations

Before the trial started, the evaluators and funders provided a workshop concerned with the conduct of an RCT, how to randomise cases, and issues such as 'blinding' and avoiding bias. The workshop meeting also agreed on the timing of pre-tests, start of intervention, duration, frequency of intervention, ages of pupils, and post-tests. Most important was agreement on a date for randomisation. Schools had to compromise from their original plans to some extent to allow greater co-ordination and so make aggregation of the individual school results feasible. This workshop was seen as crucial by all parties. In any future RCTs by schools such a meeting would be needed. Schools need to understand that getting the highest quality result is more important than what the result is. Once this is understood, the 'craft' of an RCT becomes easier to explain.

A second meeting was held before the post-testing phase. Here the evaluators learnt how to conduct the test and how to analyse and interpret the results. Again this meeting was useful to

all parties, and cleared up a few misunderstandings. A few schools then felt able to calculate their own effect sizes, and as far as it is possible to tell, they did so correctly.

Schools are reasonably good at implementing new packages, and all appeared to follow the programmes. However, it must be recalled that these schools were self-selected and chose these specific programmes. The situation might be worse if programmes were imposed on less willing schools.

School leaders are seemingly able to take responsibility for the implementation of the intervention in their schools, and an evaluation at the same time. Their involvement, added to the fact that randomisation was at an individual level giving all schools a treatment group, meant that attrition was low. The experience of the evaluators is that the closer a trial is to the schools, with the fewest parties involved, and the lower the level of randomisation, the lower the attrition is. It is likely that randomisation of schools would fail because the control schools would be more prone to dropping out.

Getting permission to innovate was easier than it would have been for an external agency. Schools were also generally good at monitoring attendance and progress. During school visits we observed that the teachers always had in-depth data on pupils' performance. Teachers were using this to make decisions such as what level of intervention should be introduced to pupils and when to proceed to the next level.

The schools allowed the evaluators access to limited background data on each relevant pupil. If, in future, schools run their own trials it is possible to envisage a process whereby schools handle this step and no personal data on named individuals is passed on.

No schools and few pupils dropped out after being deemed eligible and, once the training was completed, there was no developer pushing the advantages of their product. In terms of managing the intervention school leaders were free to make decisions regarding venues for the sessions, purchasing materials, choice of equipment, timings and class adjustment without any developer's direct involvement.

In addition to the conduct of the trial, the process and training involved builds the capacity of practitioners in reading and critiquing research claims. If conducting such research was seen as a part of schools' functions then the overall cost of research could go down. It may even be possible to create some kind of nationwide ongoing trial with all willing schools contributing to an on-line database.

It was observed that most staff involved increasingly became advocates for their programmes during the trial, and schools had already made arrangements to continue with and expand its use for future years. They did not always retain the mental equipoise needed to conduct a fair test.

It is not certain that schools can be trusted to conduct the randomisation entirely by themselves, perhaps because they allow practicalities and concern for some individuals to over-ride the demands of the evaluation. This may be especially so when there is any conflict of interest, such as the involvement of the developer (as happened here in the training). The waiting-list design was partly intended to reduce any subversion. Although conducted by the school leads, evaluators were monitoring and strongly advising on the process throughout the

period of intervention. Evaluators wanted to ensure independence of allocation by using methods such as receiving the names of target pupils and matching the names after group allocation. However, it was not completely within the control of evaluators to convince teachers that the randomisation should be a blind allocation of pupils in two groups. Part of the evaluators' role was to judge whether, given expert advice, schools were able to conduct RCTs. In this respect, and this respect alone, the answer has to be that schools cannot necessarily be trusted to randomise, and this step at least must be independent.

School leaders did not always appreciate the importance of some aspects of the evaluation. For example, when pressed they were happy to support the evaluators who were trying to locate and test missing pupils. But they did not do this on their own initiative, and had no real concept of the dangers from attrition (despite discussion in the training days). Although the school leaders were given guidance, materials and a template to develop school reports, in addition to the two training days on evaluation, none of the schools submitted an individual school report on the results. The design of an RCT makes analysis simple – the headline result is just the standardised average difference between the two groups at post-test. But this is seemingly beyond the capacity of some school leaders.

Acknowledgements

This work was funded by the Educational Endowment Foundation. A report for practitioners appears at https://educationendowmentfoundation.org.uk/uploads/pdf/Fresh_Start_(Final).pdf

References

Berk, R. and Freedman, D. (2001) *Statistical assumptions as empirical commitments*, http://www.stat.berkeley.edu/~census/berk2.pdf, accessed 030714

Brooks, G. (2003) *Sound Sense: the phonics element of the National Literacy Strategy. A report to the Department for Education and Skills*, DfES website, 20/8/03: http://www.standards.dfes.gov.uk/pdf/literacy/gbrooks_phonics.pdf

Brooks, G. (2007) *What works for pupils with literacy difficulties? The effectiveness of intervention schemes*, London: DCSF Publications

Brooks, G., Harman, J. and Harman, M. (2003) *Catching Up at Key Stage 3: an evaluation of the Ruth Miskin [RML2] pilot project 2002/2003*, A report to the Department for Education and Skills, Sheffield: University of Sheffield

Carver, R. (1978) The case against statistical significance testing, *Harvard Educational Review*, 48, 378-399

Department for Education (2013) *Statistical first release: National Curriculum Assessments at Key Stage 2 in England, 2013* (revised), DfE: National Statistics https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/264987/SFR51_2013_KS2_Text.pdf

Ehri, L., Nunes, S., Stahl, S. and Willows, D. (2001) Systematic phonics instruction helps students learn to read: Evidence from the National Reading Panel's meta-analysis, *Review of Educational Research*, 71, 3, 393-447

Falk, R. and Greenbaum, C. (1995) Significance tests die hard: the amazing persistence of a probabilistic misconception, *Theory and Psychology*, 5, 75-98

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*, 26, 9, doi: 10.1371/journal.pone.0089900

Goldacre, B. (2012) *Bad Pharma*, London: HarperCollins

Gorard, S. (2013) *Research Design: Robust approaches for the social sciences*, London: SAGE

Gorard, S. (2014) A proposal for judging the trustworthiness of research findings, *Radical Statistics*, 110, 47-60, http://www.radstats.org.uk/no110/Gorard110.pdf

Gorard, S. (2015) Rethinking "quantitative" methods and the development of new researchers, *Review of Education*, 3, 1, 72-96, doi: 10.1002/rev3.3041

Gorard, S. and Gorard, J. (2015) What to do instead of significance testing? Calculating the 'number of counterfactual cases needed to disturb a finding', *International Journal of Social Research Methodology*, http://dx.doi.org/10.1080/13645579.2015.1091235

Gorard, S., Siddiqui, N. and See, BH (2015) An evaluation of the 'Switch-on reading' literacy catch-up programme, *British Educational Research Journal*, 41, 4, 596-612, DOI: 10.1002/berj.3157

House of Commons Education and Skills Committee (2005) *Teaching Children to Read: Eighth report of session 2004-05:* http://www.publications.parliament.uk/pa/cm200405/cmselect/cmeduski/121/121.pdf

Johnston, R. , McGeown, S., and Watson, J. (2012) Long-term effects of synthetic versus analytic phonics teaching on the reading and spelling ability of 10 year old boys and girls, *Reading and Writing*, 25, 6, 1365-1384

Johnston, R. and Watson, J. (2004) Accelerating the development of reading, spelling and phonemic awareness skills in initial readers, *Reading and Writing: An Interdisciplinary Journal*, 17, 4, 327-57

King, B., and Kasim, A. (2015) *Evaluation of Rapid Phonics*, Education Endowment Foundation: London, https://educationendowmentfoundation.org.uk/uploads/pdf/Rapid_Phonics_(Final).pdf

Lanes, D., Perkins, D., Whatmuff, T., Tarokh, H. and Vincent, R. (2005) *A survey of Leicester City Schools using the RML1 and RML2 literacy programme*, Leicester: Leicester City LEA (mimeograph)

Lipsey, M., Puzio, K., Yun, C., Hebert, M., Steinka-Fry, K., Cole, M., Roberts, M., Anthony, K. and Busick, M. (2012) *Translating the statistical representation of the effects of education interventions into more readily interpretable forms*, Washington DC: Institute of Education Sciences

McArthur, G., Eve, P., Jones, K., Banales, E., Kohnen, S., Anandakumar, T., Larsen, L., Marinus, E., Wang, HC and Castles, A. (2012) Phonics training for English-speaking poor readers, *Cochrane Database of Systematic Reviews*, 12, doi: 10.1002/14651858.CD009115.pub2

Merrell, C., and Kasim, A. (2015) *Evaluation of Butterfly Phonics*, Education Endowment Foundation: London, http://www.realaction.org.uk/wp-content/uploads/2014/07/Butterfly_Phonics_Final-EEF-report.pdf

OFSTED (2010) *Reading by six: How the best schools do it*, London: OFSTED

Rose J. (2006) *Independent review of the Teaching of Early Reading*, London: DFES http://www.literacytrust.org.uk/assets/0000/1175/Rose_Review.pdf

Select Committee on Education and Skills Eighth Report (2005) *Teaching Methods*, http://www.publications.parliament.uk/pa/cm200405/cmselect/cmeduski/121/12106.htm

Torgerson, C., Brooks, G. and Hall, J. (2006) *A Systematic Review of the Research Literature on the Use of Phonics in the Teaching of Reading and Spelling*, London: DfES, Research Report 711, https://czone.eastsussex.gov.uk/sites/gtp/library/core/english/Documents/phonics/A%20Systematic%20Review%20of%20the%20Research%20Literature%20on%20the%20Use%20of%20Phonics%20in%20the%20Teaching%20of%20Reading%20and%20Spelling.pdf

Watts, D. (1991) Why is introductory statistics difficult to learn?, *The American Statistician*, 45, 4, 290-291

What Works Clearing House (WWCH) (2010) *Sound partners. US Department of Education: Institute of Education Sciences*, http://ies.ed.gov/ncee/wwc/pdf/intervention_reports/wwc_soundpartners_092110.pdf

Wyse, D. and Goswami, U. (2008) Synthetic phonics and the teaching of reading, *British Educational Research Journal*, 34. 6, 691-710