Interpreting RCT, process evaluation and case study evidence in evaluating the Integrated Group Reading (IGR) programme: a teacher-led, classroom-based intervention for Year 2 and 3 pupils struggling to read

Koutsouris, G., Norwich, B.,

Graduate School of Education, University of Exeter

and A. Bessudnov

School of Sociology, Philosophy and Anthropology, University of Exeter

Abstract

Almost 20% of English pupils still experience difficulties in reading despite a predominantly phonics approach that works well for most children, but not for all; so other approaches need to be explored. The IGR programme involves an inclusive approach to targeted teaching led by class teachers using a group-based class organisation and the integration of diverse research-based approaches (language and phonics-based). IGR has been evaluated in thirty-four English schools in five varied local authority areas using a cluster randomised design and a process evaluation. IGR was found to support enjoyment of reading with as much reading gains as the more phonics-oriented programmes used in control classes. Following its use, there were gains in teachers' self-efficacy in teaching reading, and no negative effects on the class pupils' reading. This study shows what a more inclusive approach to targeted reading intervention can achieve with a well-resourced programme. Questions can be about the interpretation of RCT findings when it comes to classroom-based educational interventions, and about teacher choice in opting for alternate teaching approaches.

Keywords: *randomised controlled trial*, *phonics*, *targeted interventions*, *response to intervention, reading programme*

Introduction

Using RCTs in educational research

Randomised control trials (RCTs) have often been presented as the 'gold standard' of educational evaluation (Goldacre 2013), providing the most robust evidence available to educational practitioners and policy makers, thus cultivating a 'what works' culture (as represented e.g. by the *Education Endowment Foundation* toolkit). Although evaluation research has a role in educational research, a side effect of a 'what works' culture is that experimental research often does not discuss explicitly that RCT findings are particular to specific contexts (for example area, school and children's characteristics) and mediational factors (such as teachers' and pupils' self-efficacy or motivation). These factors are interacting with each other in such complex ways that programmes that were found to be successful in some contexts might prove less so in different ones. This issue can become even more apparent and it is indeed more openly discussed (Vaughn et al. 2016), when there is no difference between the treatment and control group (null findings), and the researchers have to explore the possible reasons behind their inconclusive findings. Concerns about the relevance of RCTs in evaluating social 'real-life' interventions have often been raised by authors (indicatively Goodman et al. 2018; Hammersley 2015; Thomas 2016), and are a matter broadly acknowledged by researchers involved in large-scale RCTs (e.g. Humphrey et al. 2016 – guidance on process evaluation for EEF trials). This paper echoes some of the arguments about the complexity associated with

educational evaluation based on the findings of the trial of the Integrated Group Reading (IGR) programme, an inclusive targeted early reading intervention.

The context

Almost 20% of children in English Primary Schools on entering Key Stage 2 (KS2) do so as delayed or non-starting readers (DfE 2017), and analysis of the DfE phonics test in 2016 indicates that around 10% did not reach the nationally set threshold level at the end of Year 2 (DfE 2016). This persistent challenge can be attributed to several factors including the opaque nature of English orthography (Wyse and Goswami 2008). Though impressive attempts are made in Key Stage 1 (KS1) to ensure that all children acquire and can deploy the phonic knowledge that they will need as a basis for both encoding for writing and decoding for reading, national statistics only show modest gains between 2006-13 (DfE 2013), though the picture is improved after 2013 (DfE 2016). The significance of this is that a small minority is unable to make sufficient progress in reading to be able to benefit fully from an increasingly lively and diverse KS2 curriculum in the context of classroom 'Quality First' teaching.

How teaching is geared towards the needs of pupils who are struggling to learn to read has increasingly been approached from the perspective of the Response to Intervention (RTI) model (Fien et al. 2011; Griffiths and Stuart 2013) to distinguish between what is offered to all (tier 1) and what to some (tier 2) or to a few (tier 3) (we will use the term '*tier*' in this paper and not the alternative term '*wave*'). Current practice is to provide 'Quality First' (tier 1) teaching that is meant to be differentiated but might not be differentiated enough for pupils struggling to learn. Tailored teaching for those not progressing at the expected rate is often offered in higher tiers (2 or 3) as additional pull-out sessions with people other than the class teacher (such as teaching assistants). This has two potential implications: i) it can create a 'separation' effect (EEF 2015) by limiting the opportunities of these pupils for quality time with the class teacher and peer interactions; and ii) it can mean the loss of crucial learning time – for instance, it has been found that children who had immediate access to tier 2/3 additional support had improved reading outcomes at the end of Year 1 (Al Otaiba et al. 2014).

An additional matter relevant to tier 2/ 3 remedial programmes is the approach to teaching reading. Since the Rose (2006) Report English policy favours synthetic phonics, yet the small but persistent percentage of pupils who were taught through synthetic phonics and still experience difficulties seems to suggest that an additional teaching approach could be tried for these children.

The IGR programme

The Integrated Group Reading (IGR) programme has been designed in response to the above issues. The programme is a tier 2 intervention targeting Year 2 and 3 pupils who are delayed in reading and is taught by class teachers in small groups during the existing small group organisation of lessons (Guided Reading or other form of group reading organisation). It is part of a classroom-wide model, with all pupils being in groups receiving teacher attention over a period of a week, supported by a teaching assistant. It was expected that this arrangement would not disadvantage the other children in the class despite the investment of teacher time to few struggling readers, as it was built on the existing group reading classroom organisation.

In this sense, IGR is original in two ways: Firstly, it introduces a tier 2 targeted intervention into the 'Quality First' setting. A literature search done by the authors to explore the nature of additional support for struggling readers found no systematic evaluations of programmes using similar delivery arrangements (the closer examples were in Scandinavia, e.g. Brinchmann et al. 2016 report the trial of a teacher-delivered literacy programme – yet, administered in pull-out sessions). It was also found

that the people responsible for teaching tier 2 interventions in UK-based trials were in most cases teaching assistants in pull-out sessions (e.g. See et al. 2015; Clarke et al. 2010; Duff et al. 2008; Hatcher et al. 2006). This can be to some extent indicative of how additional support is organised in UK primary schools. Although there is a risk of over-generalising to situations where appropriate support is provided by teaching assistants, IGR is influenced by the principle that struggling readers need the teacher's attention and time the most. Secondly, as discussed, an additional approach needs to be tried for the 20% of children who were taught through synthetic phonics but still experience difficulties. IGR adopts a multi-perspective approach to the teaching of reading, integrating (analytic) phonics (Torgerson et al. 2018), story-telling for oral language development (Clarke et al. 2010), word games (Raffaele Mendez et al. 2016), and elements from Paired Reading (Topping et al. 2011) and Reading Recovery (Clay 1994) (for which there is extensive research evidence that cannot be explored here). So, the term 'Integrated' refers not only to the inclusive aspect of the class-based organisation that enables pupils' identified for tier 2 support access to teacher expertise alongside their peers, but also to the integration of the above professional and research-based approaches to literacy education.

IGR adopts a systematic story-led teaching approach, designed to replace group reading for struggling Year 2 and Year 3 children in the classroom until such time as they have become confident and engaged early readers and can access higher level reading and comprehension work. The programme used a range of 52 specially written reading books with simple illustrations and accompanying story-specific games, developed with the narrative requirements of later-learning readers in mind, and deliberately short so that one story could be completed in each lesson. At the time of the evaluation, the programme materials began at red/yellow readability level (reading age equivalency 5.07 yrs.) and progressed through to turquoise readability level (reading age equivalency 7.01 – 7.04 yrs.). Through the programme, children experience each narrative text at several interdependent but interlocked levels (story, sentence and word), with game-playing functioning as a key motivator and memory tool at each level. The cross-referencing of orthographic progression and readability levels (specially refined by the programme developer, Jan Stebbing) allows to define words for reading as expressive vocabulary (the known words) and receptive vocabulary (words beyond current Instructional level). This receptive first encounter with words-yet-to-be-mastered makes learning to read into a comfortable activity for children who need support to re-engage with reading. For details about the programme visit: http://www.integratedgroupreading.co.uk/

Table 1 presents the structural logic behind the programme's teaching routine.

Insert here Table 1

The IGR evaluation

IGR was trialled by the Graduate School of Education of the University of Exeter with Year 2 and 3 pupils in 34 English schools in five varied local authority areas across two years (2015-2017). The project was funded by Nuffield Foundation. The programme and evaluation arm teams were both based in the Graduate School of Education, University of Exeter, but operated separately. IGR was run for 28 weeks (i.e. 7 months) in phase 1 (November 2015 to May 2016) and phase 2 (October 2016 to April 2017).

The study explored the following questions:

1. What were the immediate effects of the IGR programme – in reading accuracy and comprehension, reading attitude and overall attitude to school – after its first and second year

of implementation (phase 1 and 2) with Year 2 and 3 children identified as most struggling in reading, compared to similar children experiencing usual teaching?
2. What were the immediate effects in reading for classroom children in the Year 2 and 3 classes that used the IGR programme with those most struggling in reading, compared to similar children in classes experiencing usual teaching?
3. What were the long-term IGR effects for children having the intervention and classroom pupils?
4. What was the context of the programme implementation and what processes were involved?
5. How was reading taught in the control schools?

Methods

Study's design

In Phase 1, the IGR phase 1 programme evaluation had a mixed methodological design, involving:

- A clustered randomised control trial (clusters at the school level) with the comparison group in control schools on a waiting list to use the intervention.
- A process evaluation of implementation and teachers' and pupils' IGR experiences. This involved in-depth school level case studies, and a 2-weekly log to monitor the fidelity of implementation.

The phase 2 evaluation involved:

- A quasi-experimental study, in which pupil outcomes in the control schools in phase 1 were compared with pupil outcomes in the same schools in phase 2 when IGR was used.
- A process evaluation of implementation as in phase 1.

So, this design allowed us to make two comparisons in this trial: the outcomes in the randomly assigned treatment and control schools in phase 1 and the outcomes in the control schools in phase 1 and the same schools that were treated in phase 2. Some children were control cases in phase 1 and treated cases in phase 2.

Participants

The project had the support and co-operation of Literacy Advisers in 4 local areas [in the South West (1), West Midlands (2) and Greater London (1)] who were actively involved in the recruitment process (and later in a supportive role).

In phase 1, 32 schools participated across the 4 local areas. Randomisation was applied at school level and took place before the time 1 assessments (September 2015). 33 schools were randomly assigned into the treatment and control groups, with one school deciding to pull out just after the procedure (details in figure 1). At the end of phase 1 (and during phase 2 – 2016/17), phase 1 intervention schools could decide to continue using the programme or not. In most cases, teachers reported that they were using the programme materials and aspects of the routine, but in a looser way. Phase 1 intervention pupils were assessed again in May 2017 for long term effects. In phase 2, the intervention was offered to the phase 1 control schools (16 schools) which had no access to the programme in phase 1. Of these schools, 3 decided not to use the programme due to organisational and staffing difficulties, and 2 new schools were recruited from another South West Local Authority. Altogether in phase 2, there were 15 intervention schools (13 which were control schools in phase 1, plus 2 newly recruited), and there was no control group (see figure 1). Classes that used IGR in these

schools were not selected randomly. This was not possible when a school had only one year 2 and 3 class; when more than one class per year, we left it to the school to decide which class used IGR.

Insert here Figure 1

Pupil identification for using the programme

Both the IGR and control focus groups were identified by teachers following a standard procedure, using a template and instructions adapted from Speece et al. (2011). Speece et al. (2011) had found that teacher rating is an accurate and efficient predictor of early reading difficulties, with the additional advantage that teachers have intimate knowledge of their pupils and that a teacher rating system is less time-consuming and more cost-effective compared to a standardised assessment. The teachers were asked to identify four pupils who would benefit from literacy support and were given the selection form and instructions. The selection was based on a teacher report scale that included reading attainment and attitude; the teachers determined the attainment levels by reference to class reading levels. In only a few cases, less than four pupils were identified.

Assessments

There were 4 assessment times: September 2015 (phase 1, time 1), July 2016 (phase 1, time 2), September 2016 (phase 2, time 3), and May 2017 (long-term effects, time 4 and phase 2, time 4).

Assessments included individual and whole class assessments, as below:

1. Individual assessments

Individual assessments (only for the pupils identified for the IGR and control groups) were conducted by specially recruited and trained research associates blind to the allocation of schools to intervention and comparison conditions.

York Assessment of Reading for Comprehension (YARC) (Snowling et al. 2009): The YARC test gives separate scores for reading accuracy, rate and comprehension, of which we were interested only in accuracy and comprehension. Yet, due to the high proportion of missing values, the test could not be used in the analysis. This is discussed in the section about the study's limitations.

Single Word Reading Test (SWRT): The SWRT was developed by Foster (2007) and is part of the YARC test (a tool to select an appropriate starting passage). All pupils could access the test.

'How I Feel about Reading' (HIFAR): The HIFAR covers reading attitude and competence items using a 5-point scale. HIFAR has good psychometric characteristics (Cronbach alpha 0.86) (Chapman and Tunmer 1995). Though the sentences in this scale were read to the pupils, many Year 2 pupils seemed to find the scale difficult to follow.

'How I Feel about My School' (HIFAMS): HIFAMS covers school experiences and well-being (7 items) and has satisfactory psychometric characteristics (Cronbach alpha 0.62 to 0.67) (Allen et al. 2017).

2. Whole class assessments

Hodder Group Reading Test (HGRT): A whole class reading assessment of reading comprehension at word, sentence and text levels was used to explore any effects of the delivery of IGR on other class pupils (as IGR was delivered by the teacher in whole class sessions, the teacher had to make an investment of time to a small group of struggling readers). The HGRT is reported to have very good reliability (Cronbach alpha 0.92 to 0.95) (Hodder Education 2000). The paper tests were sent to the

schools by the research team, accompanied by detailed step-by-step instructions (and an envelope with prepaid postage), and were delivered by the class teacher to all pupils (IGR and class pupils). The completed tests were then returned by post for scoring.

Power analysis

Before the data collection, we conducted power analysis that showed that with the suggested design for the pupils having the IGR intervention we would be able to identify medium size effects ($d > 0.4$) with a power of 0.8, but not smaller effects. For the analysis at the whole class level, the sample size was larger, and we could reliably estimate even smaller effects. We conducted power analysis with the pwr package in R, adjusting for the effects of clustering at the school level (Campbell and Walters 2014)[1].

Statistical analysis methods

We considered five outcome variables for the IGR group analysis: the SWRT score, HGRT score, attitude to school scale, reading self-competence and attitude to reading scales. For the class pupils, we only considered the HGRT. YARC scores are not reported due to high volumes of missing data.

In all statistical analyses, we applied the following procedure. First, for each pupil we calculated the difference between the values of the outcome variables before and after receiving IGR. Then we tested whether the pre- and post-treatment difference in the outcomes was statistically significant in the treatment and control groups by regressing it on the treatment status variable, controlling for gender, Year Group, special educational needs (SEN) status and English as additional language (EAL).

Given the clustered design of the study, we needed to correct standard errors for within-cluster correlation. This can be done either by applying cluster-robust standard errors to the linear regression (as in the survey R package) or by fitting multilevel models (using the lme4 R package). We applied both methods, and the results were similar. Here, we report the models with cluster-robust standard errors. For the IGR group analysis, the observations were clustered at the level of the IGR group. For the rest of the class analysis, the observations were clustered at the class level.

Intervention implementation

IGR was delivered 4 times a week for 30 minutes as part of the group reading session for all pupils (e.g. Guided Reading). The teacher taught the IGR group twice a week and introduced a new book at each session. The teaching assistant worked with the group in-between the teacher sessions for consolidation. Teacher and teaching assistants had discrete yet interconnected roles, with the teacher keeping the main role (see table 1 for details about the programme teaching routine). During teacher-led IGR, the rest of the classroom worked independently or with a teaching assistant on various reading-related activities (such as comprehension tasks, dictionary work, and computer literacy programmes). IGR was designed to be part of the usual group reading classroom schedule, while allowing teachers to organise their group reading rota in a more structured and efficient way for all pupils. Teachers and teaching assistants were encouraged to communicate daily regarding the pupils' reading progress, using a record form for communication.

[1] When doing the statistical analysis, we clustered the observations at the level of IGR group (usually 4 pupils), so class rather than school level. When the number of clusters was larger and the observations within clusters were positively correlated, the required sample size was smaller, so the calculations in this section represent a more conservative estimate. Note that clustering only affects standard errors, not the effect size, and turned out not to be important for our main results.

Fidelity of implementation

Implementation fidelity was monitored by an online fortnightly log (about 10 per phase) in which teachers were asked to summarise their classroom organisation, the number of teacher and teaching assistant sessions and teaching routine, and to comment on pupil attainment and attitude. The log was reviewed and revised several times during the two years to better capture departures from the suggested organisation/ implementation. In addition to the log, in each intervention school at least one teacher-led IGR session was observed by independent researcher, and the observation notes were compared to the programme team observations conducted for support purposes.

The observations and logs revealed that IGR was implemented with varied fidelity across different schools and teachers. Most common variations observed or reported included delivering IGR out of the classroom, confusing the teacher and teaching assistant roles, delivering fewer than the suggested four sessions, and not following the IGR lesson routine.

Programme training and support

Teachers received a full day's training, covering the programme methodology and aspects of classroom organisation. Programme team members visited all the participating schools at least once (and up to 3 times in Phase 2). This was also done in collaboration with Local Literacy Advisers and Education Improvement Officers who also visited schools in their areas at least once during a phase. Training was also organised locally for teaching assistants supporting IGR.

Implementation cost

IGR involved the one-off cost of materials and training for teachers and teaching assistants, calculated as £1,600 per participating class (for a group of four pupils receiving the programme). Subsequent years of implementation do not incur any materials/ training costs, as the same set of materials can be re-used and teachers can train other teachers and teaching assistants in using the programme. However, having a regular teaching assistant during group reading sessions is important to ensure a smooth implementation of the programme.

Process evaluation methods

14 schools (8 in phase 1 and 6 in phase 2) (mixed range of rural, sub/urban schools), each acting as different cases, were visited across the four local authority areas. In each school one (or more) teacher-led IGR session/s was observed and one (or more) teacher/s was interviewed.

Self-efficacy questionnaire

Teaching self-efficacy was also measured for treatment teachers (both phases) at the training day and again at end-of-the-year review meetings using a 28-item 9-point scale focusing on reading, informed by Leader-Janssen and Rankin-Erickson (2013) and Tschannen-Moran and Johnson (2011) (Cronbach's alpha for phase 1 training day: 0.91).

Ethical considerations

The project had ethical clearance from the University of Exeter. All participating schools signed a memo of understanding outlining the project's procedures and a consent form. Informed passive consent was sought from parents, for both pupils in the identified groups and class pupils, and letters were sent explaining the randomisation process and were distributed before the randomisation took place. Some schools requested an extra consent form to be produced for the collection of the demographic data for the participating class and pupils. Anonymity and

confidentiality has been applied to every aspect of the project, and school/ individual participants had the right to withdraw at any time. In order to not affect the evaluation, the scores from the study's assessments were sent to the schools after the end of phase 2 (January 2018) as aggregated results, and schools were able to access their own individual results only.

Findings

Participant characteristics

The baseline HGRT score was statistically significantly higher in the treatment group (90.4) compared to the control group (85.9) in phase 1 (see Table 2). The baseline SWRT scores were approximately the same in both groups.

For most of the demographic characteristics, there was no statistically significant difference between the treatment and control groups (table 2). Yet, in phase 1 there were more pupils identified for Special Educational Needs (SEN) School Support in control schools; this might suggest that in phase 1 some treatment schools perceived IGR as a substitute to SEN School Support provision.

Insert here Table 2. IGR: IGR and comparison groups

Immediate effects for pupils having the intervention

Tables 3 and 4 report the IGR effects from Phases 1 and 2 for five outcome variables: HGRT and SWRT scores, attitudes to school, reading self-confidence and reading attitudes. The IGR effects were calculated as the coefficients for the treatment status in the linear models where the differences between the pre-treatment and post-treatment measurements were the outcome variables. The models applied cluster-robust standard errors (at the IGR group level) and controlled for Year Group, gender, SEN and EAL status.

The p-value is the p-value for these coefficients from the same model. Cohen's d is the standardised effect size (IGR effect divided by the standard deviation of the outcome).

None of the IGR effects for any of the outcomes in either Phase 1 or 2 were statistically significant at the 90% or 95% level (tables 3 and 4). The effect sizes were mostly close to zero and never larger than 0.25. Yet, both in the treatment and control groups, pupils showed progress on the standardised reading test scores between the pre-treatment and post-treatment measurements.

Insert here Table 3. Phase 1 results: IGR pupils

Insert here Table 4. Phase 2 results: IGR pupils

There were no consistent statistically significant interactions between the IGR programme and gender, Year Group, and having English as an additional language (EAL); yet, there were some indications of positive interactions (i.e., the IGR having a stronger positive effect) for pupils having EAL and being identified for Pupil Premium, not replicated across the phases and measures.

Immediate effects for classroom pupils

We also tested whether IGR affected the reading progress of the pupils who were not directly involved in the IGR programme (the rest of the class). This was seen as important because IGR was delivered by the classroom teacher during whole-class literacy-related sessions (group reading), so the teacher was making an investment of time to a particular group of pupils.

As the sample size for these pupils was larger, the analysis has more statistical power and we were able to identify smaller effects (table 5 and 6).

Insert here Table 5. Phase 1 results: class pupils

Insert here Table 6. Phase 2 results: class pupils

In both phases, there was a small positive effect on the reading progress (as measured by the HGRT) of the class children. The standardised effect size was 0.13 in Phase 1 and 0.23 in Phase 2. In phase 2 the effect was statistically significant at the 95% significance level ($p = 0.03$).

Yet, when we fitted the interaction effect between IGR and gender in the models for the class pupils, the positive IGR effect was only present for girls. An examination of the data suggested that this was due to the unusually high baseline HGRT measures for girls in the control group. As a result of this, the girls in the control group did not show improvement between the pre- post-treatment tests, as measured on the HGRT standardized scale (see Table 7). We discuss this later in the paper.

Insert here Table 7. HGRT measures for girls and boys in the control and treatment classes

Long-term effects for pupils having the intervention and classroom pupils

Long term effects refer to a time period of about two years after the beginning of phase 1 implementation (September 2015) and 9-10 months after the programme evaluation was ended (May 2017 – time 4). We assumed that some teachers might continue using IGR to some extent or form, since they had received training and the programme materials stayed with the schools.

The only statistically significant long-term IGR effect for pupils receiving the intervention was for HGRT (table 8) – IGR effect was negative (approximately -0.3). Yet, at time 4 the mean HGRT scores in the control and treatment groups were very similar, and the negative IGR effect was most possibly due to the lower HGRT scores in the control group at time 1 possibly reflecting not entirely accurate HGRT measures at the baseline in the control schools. There were no significant (both in the statistical and substantive sense) long-term effects as measured on the SWRT scale. The long-term IGR effect on class pupils (table 8) was positive, but small and not statistically significant.

Insert here Table 8. Long-term phase 1 IGR effects

Reading ages are provided in tables 9 and 10, showing that both treatment and control groups made the same degree of modest progress, and that these children were still behind their classmates.

Insert here Table 9. Phase 1 results: mean reading ages (years: months)

Insert here Table 10. Phase 2 results: mean reading ages (years; months)

Context and processes

IGR proved to be a demanding programme as far as teacher skills were concerned, since it adopts a multi-perspective approach that can be seen as different from the current approach to early literacy that emphasises synthetic phonics and comprehension. Teachers had mixed views on this: some younger teachers who had been trained with a focus on phonics tended to alter the delivery of IGR slightly to be closer to a more phonics-driven instruction. In a similar way, the story-telling element of IGR for some teachers tended to be altered into a more inference-driven approach to text with teacher questions and pupil responses, in a teaching style closer to the Guided Reading approach

that aims to make pupils independent readers. On the other hand, many teachers appreciated the simplicity of IGR that combined a variety of light touch approaches to re-engage pupils in reading.

The IGR organisation was described as '*marginally more demanding*' by one teacher, and this seemed to reflect the overall attitude of all interviewed teachers. Most teachers could see the value of keeping all the pupils in the classroom during the intervention, but there were a few teachers who saw a tension between the inclusive aspect of IGR (keeping all pupils in the class) and the difficulty of maintaining concentration in a busy and lively class. Giving the main role to the teacher meant also that a teaching assistant had to be available to work with the rest of the class, with some schools reporting issues with teaching assistant availability.

The main issue schools and teachers had was with the number of teacher sessions when there were more than four reading groups in the class. With four reading groups already, this meant that teachers had to fit these four sessions into three days for the period of the intervention. Teachers came up with a variety of solutions to this issue with the most common being the delivery of one of the two teacher-led IGR sessions in the classroom but during a school assembly. Many schools made clear that all reading groups and pupils should have an equal entitlement to the teacher's time.

A relevant finding was the increase in teaching reading self-efficacy for treatment teachers between the training day and end-of-the-year review meetings that was statistically significant in both phases (table 11). However, this measure was not taken with teachers in control groups. We also did not find any consistent correlations between pupils' reading gains and teachers' self-efficacy scores, gender, age, experience and route to the teaching profession.

Insert here Table 11. Teacher self-efficacy

Teaching in the control classrooms

The control teaching data were collected from the phase 1 control schools only (there was no control group in phase 2), using two online surveys sent in autumn 2015 and again in autumn 2016. When control teachers were asked how much time they and their teaching assistants spent with the identified pupils, they reported giving considerable additional time to the identified pupils (figure 2). In addition, a number of literacy programmes was used in control schools, including *ReadWrite Inc, Toe by Toe* and *The Five-Minute Box for Literacy.*

Insert here Figure 2

Discussion

Key findings

Participating children in schools using IGR in both phase 1 or 2 made the same degree of progress in reading accuracy and comprehension, compared to similarly struggling children in control schools: taking into account both SWRT and HGRT, mean progress of 11 months in 7 months in phase 1, and mean progress of 12 months in 7 months in phase 2 – often seen as 'modest impact' (Brooks 2016). There were no statistically significant changes for reading and school attitude in either the treatment or control group. There were also no significant (both in the statistical and substantive sense) longer-term effects on the measures used. This suggests that our initial hypothesis that IGR would improve reading gains and attitudes for the IGR group compared to the control group was not supported by the findings. However, this hypothesis assumed that control pupils would not receive as intensive additional teaching as the study has shown that they did.

In phase 1 there was no statistically significant difference in gains between treatment and control classes for class children. This confirms our initial hypothesis that IGR in the classroom would not have any negative effect on the reading for classroom pupils. In phase 2, classroom children showed somewhat better progress on the Hodder standardised scale in the treatment classes compared to the control classes (d = 0.2). This effect was statistically significant, but this is possibly due to a significantly higher baseline HGRT mean score for girls in the control group that we cannot account for. This matter is further discussed in the section about the study's limitations.

IGR was used with varied fidelity, and many teachers found it demanding, but viable. In addition, control schools did not just continue with typical teaching; teachers recognised that control pupils had significant additional needs, so they also had considerable supplementary, but mainly phonics-based teaching input, making for a complex comparison.

With regards to treatment teachers, the change in their teaching reading self-efficacy between the training day and the end-of-the-year review meetings was found to be statistically significant across phases 1 and 2. Self-efficacy was not measured for control teachers; so, this finding only suggests that IGR use can be associated with teachers becoming more confident in their literacy teaching. This might possibly be because the IGR organisation model made it possible for teachers to work with their pupils who struggled the most without having to leave the classroom.

The IGR approach to reading

Schools in England largely use phonics approaches to teach early reading, with explicit phonics teaching shown to produce good results (e.g. Torgerson et al. 2018). Yet government statistics, discussed in the introduction, reveal that a consistently present small minority of pupils cannot overcome difficulties in reading through this teaching approach. Thus, other approaches might be tried with those pupils for whom phonics has not resulted in enough progress.

Lovett et al. (2017) discuss the importance of multi-perspective remedial programmes that have the potential to address a number of issues beyond phonological difficulties. The implication of this is that teachers have some scope to select the method they feel better suits their teaching style and pupils' needs, whether this method is mainly phonics-centric or not. Our findings suggest that some teachers felt comfortable in using more phonics-oriented approaches, and others were attracted to more multi-perspective approaches to reading (such as IGR). This is particularly evident in the way storytelling was used by the teachers in the study, with some finding it very challenging, and others experiencing it as a natural activity. Thus, the selection of an appropriate method for teaching reading need not be determined by specific policy prescription – as in the case of the Rose (2006) report for synthetic phonics – but rather be research-informed and involve teacher decision-making. This also means that IGR could be considered by teachers as an alternative to the current pattern of tier 2 interventions that involve more phonics-based programmes delivered by teaching assistants.

The IGR organisation

The IGR programme was delivered by the class teacher during whole class group reading sessions two times a week, followed by teaching assistant consolidation sessions (also in the classroom). So, the teacher (and to a lesser extent the teaching assistant) had to make a considerable investment of time to a particular group of pupils. Despite this arrangement, the change in reading outcomes (measured by HGRT) was approximately the same in the control and treatment groups across phases for class pupils. This is consistent with the IGR teaching of a sub-group not affecting how other pupils

progressed either positively or negatively. Similar findings were reported in a recent EEF report (Patel et al. 2017) but the intervention was delivered by teaching assistants in pull-out sessions.

There was also evidence that pupils, with a few exceptions, were not concerned about being visible in the IGR group, a low attainment group, and that being in the IGR group was often seen as a privilege because of the out of the ordinary activities. Assumptions that pupil grouping, as practised in IGR, as always leading to devaluation and stigma underpin the general rejection of ability grouping advocated in some inclusive pedagogy perspectives (Florian and Black-Hawkins 2011). However, the process evaluation indicates that the inclusive features of IGR teaching are compatible with the temporary reading ability grouping of some pupils.

The implication of this is that it is i) practically possible and ii) does not result in any harm to pupils for the teacher to offer tailored tier 2 support in the 'Quality First' teaching setting, during a whole-class teaching session. This can allow pupils identified for tier 2 support to access tailored teaching and spend quality time with their teacher/ peers.

What was compared in the trial: IGR fidelity and control teaching

As discussed, on the one hand IGR was implemented with varied fidelity, and on the other hand, control schools offered intensive additional support to control pupils.

Discussing the matter of fidelity to a real-world intervention, Moore et al. (2015) note that fidelity is best seen as a matter of degree rather than as a fixed quality. In the study, most teachers tried to stay faithful to the study's protocol, but some found this practically difficult. This is consistent with findings from other studies, e.g. Gorard et al. (2015) and See et al. (2015), where the fidelity of implementation was found to have varied. Also, it is indicative that a lot of time was spent on preparing for the practical, organisational aspects of the programme, but few teachers devoted time to reflect on the teaching approach of the programme. There are teacher case studies for phase 2 (Koutsouris and Norwich 2018) which show how fidelity relates to pupil outcomes and how IGR is one of many factors that influence reading outcomes. The case studies showed how the same programme can be implemented differently in local circumstances. The ways in which IGR was used reflected various factors including how teachers experienced the pressures of the national curriculum, their attitude to the IGR approach to reading, the school ethos and the resources and support available. The teacher cases also did point to particular combinations and interactions that may be associated with successful or less successful results. For instance, an important factor was whether the teachers felt that the programme fitted well with their teaching style. This reinforces Moore et al. (2015) with regards to how the local context is crucial for intervention implementation.

The analysis suggests that IGR is not a simple intervention that can be applied well or not irrespective of its teaching context. Its introduction as a programme was involved in a complex of interactions (involving pupils, teachers, schools or the broader context) that might have affected programme implementation and outcomes. The conclusion drawn is that in addition to the statistical effect sizes evaluators ought to pay attention to the context in which a programme is implemented and processes involved, especially when it comes to interventions evaluated in real classrooms.

In addition, we examined the teaching in the phase 1 control classes in some detail and found that typical teaching included intensive additional support of various forms (teacher and teaching assistant-led, in and out of the classroom) and often the use of other literacy programmes – in many cases with a focus on synthetic phonics (e.g. *Toe by Toe*). This is a matter which was found in other

studies as well (Vaughn et al. 2016). As we had decided not to intervene with the teaching decisions in the control schools, we were in fact comparing IGR to an intensive programme of (mainly phonics-oriented) support, driven by the national curriculum and the assessment requirements. We were also aware that in the longer term follow up of IGR in phase 1 some teachers will have continued to use IGR to some extent in the period after the intervention. We were unable to control for this. Though we could have analysed longer term effects for those who did not use IGR in this period, the statistical power was lacking due to reduced sample size.

Study's limitations

Pupil identification

The study's identification procedures adapted from Speece (2011) were adopted as a quick, simple, cost-effective and research-evidenced approach that did not require teachers to administer a standardised assessment. Yet, one of the consequences was that a few IGR groups included children with standardised reading scores already above 85 (1 standard deviation below test mean of 100) or pupils with very varied abilities. This was related to the fact that some classrooms had high mean reading levels and few struggling pupils (often less than 4), whereas others had low mean reading levels and more than 4 struggling pupils. The latter case did not pose a problem to the study as in such cases, the teachers selected 4 pupils to be individually monitored for the evaluation, and they were encouraged to use IGR with other non-individually monitored pupils as they saw fit. Many teachers followed this advice. However, in the former case of the high achieving classrooms, identification for the programme was in some cases seen as problematic for effectively teaching the group and choosing appropriate materials for all in the group. This matter was more evident in some local areas than in others revealing the differences in reading attainment across project areas. To avoid similar issues, future research studies could consider introducing a cap of around 85 in standardised score terms.

Issues with the YARC test and measuring reading progress

The YARC test was chosen as the main reading test of the study based on its recent norms and design. However, only a minority of pupils could complete the baseline assessments (e.g. 42% of the control pupils in September 2015). This was due to the test requiring the completion of two whole reading passages and accompanying questions (as opposed to one for similar tests). Using two passages can improve the accuracy of the test results, but it proved a difficult task for those struggling to read for one reason or another. The YARC was found to be a complex test to administer by the team of assessors. Administration issues might account for the low proportion of pupils who could complete the baseline measure; yet, we found that even with the assessors having been even more experienced with the test in phase 2, there was also a large minority of pupils who did not complete the baseline in phase 2.

We planned for this contingency by also using SWRT and HGRT as backup measures. However, we conclude that in future studies with children of this age and reading level, individual assessments need to use a simpler test that does not require two passages of text reading as does the YARC.

In addition, the HGRT was delivered by the class teacher and not by independent researchers, and although specific guidance and support was provided, there was space for delivery-related errors.

This might be the reason why there was a significantly higher baseline HGRT mean score only for girls in the control group that, as discussed, we cannot account for.

Measuring reading and school attitudes

As mentioned above, the administration of these scales (HIFAR and HIFAM) indicate some difficulties in understanding the read-out statements, and the findings showed no change in any of these measures in both phases. Yet, the process evaluation (Koutsouris et al. 2018; Norwich et al. 2018) gave consistent indications of very positive response by pupils using the IGR. This raises questions about how best to measure these affective characteristics with pupils of this age and level of reading in future studies.

Conclusion

The study's findings show that IGR, an approach to reading that incorporates multiple research-informed strategies (including analytic phonics), trialled with Year 2 and 3 pupils in primary schools across England, can bring the same results as the currently dominant synthetic phonics approach. This has important policy implications, as the assumptions of the Rose (2006) report, which endorsed synthetic phonics in England, have often been questioned (e.g. Wyse and Goswami 2008), and there is a persistent percentage of struggling readers not progressing despite the use of this kind of approach.

The IGR study suggests that approaches using phonics alongside other perspectives have the potential to bring similar results to a mainly phonics teaching approach, with the advantage that they can better support the enjoyment of reading as is shown by the IGR process evaluation, though not captured through our attitude measures. The IGR approach could be considered for pupils for whom phonics teaching has failed to bring positive results.

The significance of this study is that it opens up another approach to targeted early reading teaching for those struggling to learn to read. When research evidence is inconclusive (Torgerson et al., 2018), it is reasonable that teachers exercise some autonomy in deciding about a teaching approach to reading for struggling readers, which could be based on their judgement about what best suits their own teaching style and their pupils' needs. Clearly there is a need for further studies using IGR, building on the current one. But, the suggestion is that policy should recognise the complexity of teaching decisions and that research-informed teacher decision-making is compatible with some teacher autonomy in the field of targeted early reading interventions.

Also, the IGR organisation that enables teachers to offer tailored tier 2 provision in a 'Quality First' setting proved to be viable but challenging. This has consequences for the way additional provision is organised for pupils identified as being in need of tier 2 support. It shows how it is practically possible for the teacher to take responsibility for the learning of all pupils – even by offering extra time to some who most need it – without any negative effects for the rest of the class.

A final point is that with regards to programme evaluation, it might be too simple to seek to answer whether a programme works or not, while being silent about the circumstances. As Pawson and Tilley (2004) note, it is better to ask the question: 'for whom, in what circumstances, in what respects, and how?' (p. 2). This might not provide the certainty that some researchers might aspire to nor policy makers would prefer, but it does capture the complexity associated with programme

evaluation. It can also give an insight into the factors that make a programme more or less successful and give directions for revisions and further development.

Acknowledgement: The IGR project has been funded by the Nuffield Foundation. The views expressed are those of the authors and not necessarily those of the Foundation.

References

Allen, K., Marlow, R., Edwards, V., Parker, C., Rodgers, L., Ukoumunne, O. C., et al. and T. Ford 2017. "How I Feel About My School: The construction and validation of a measure of wellbeing at school for primary school children". *Clinical child psychology and psychiatry*: 1–17.

Al Otaiba, S., Connor, C. M., Folsom, J. S., Wanzek, J., Greulich, L., Schatschneider, C., and R. K. Wagner. 2014. "To Wait in Tier 1 or Intervene Immediately: A Randomized Experiment Examining First-Grade Response to Intervention in Reading". *Exceptional Children:* 81 (1), 11–27.

Brooks, G. 2016. *What works for pupils with literacy difficulties*. The Dyslexia-SpLD Trust.

Campbell, M. and Walters, S. 2014. *How to Design, Analyse and Report Cluster Randomised Trials in Medicine and Health Research*. Chichester: Wiley.

Chapman, J. W., and W. E. Tunmer. 1995. "Development of young children's reading self-concepts: An examination of emerging subcomponents and their relationship with reading achievement". *Journal of Educational Psychology* 87 (1): 154.

Clarke, P. J., Snowling, M. J., Truelove, E., and C. Hulme. 2010. "Ameliorating Children's Reading-Comprehension Difficulties: A Randomized Controlled Trial". *Psychological Science* 21 (8): 1106–1116.

Clay, M. M. 1994. "Reading Recovery: The Wider Implications of an Educational Innovation". *Literacy, Teaching and Learning* 1 (1): 121–141.

DfE. 2013. *The national curriculum in England: Key stages 1 and 2 framework document*. London: DfE.

DfE. 2016. *Phonics screening check and Key stage 1 assessments in England*. London: DfE.

DfE. 2017. *SFR49 2017 KS1 National tables*. London: DfE.

Duff, F. J., Fieldsend, E., Bowyer-Crane, C., Hulme, C., Smith, G., Gibbs, S., and M. J. Snowling. 2008. "Reading with vocabulary intervention: evaluation of an instruction for children with poor response to reading intervention". *Journal of Research in Reading* 31 (3): 319–336.

EEF. 2015. *Making Best Use of Teaching Assistants: Guidance Report*. London: EEF.

Fien, H., Smith, J. L. M., Baker, S. K., Chaparro, E., Baker, D. L., and J. A. Preciado. 2011. "Including English Learners in a Multitiered Approach to Early Reading Instruction and Intervention". *Assessment for Effective Intervention* 36 (3): 143–157.

Florian, L. and K. Black-Hawkins. 2011. "Exploring Inclusive Pedagogy". *British Educational Research Journal* 37 (5): 813–28.

Foster, H. 2007. *Single word reading test 6-16*. London: GL Assessment Limited.

Goldacre, B. 2013. *Building evidence into education*. London: Department for Education.

Goodman, L. A., Epstein, D. and M. Sullivan. 2018. "Beyond the RCT: Integrating rigor and relevance to evaluate the outcomes of domestic violence programs". *American Journal of Evaluation* 39 (1): 58–70.

Gorard, S., Siddiqui, N., and B. H. See. 2015. "An evaluation of the Switch-on Reading literacy catch-up programme". *British Educational Research Journal* 41 (4): 596–612.

Griffiths, Y., and M. Stuart. 2013. "Reviewing evidence-based practice for pupils with dyslexia and literacy difficulties". *Journal of Research in Reading* 36 (1): 96–116.

Hammersley, M. 2015. "Against 'gold standards' in research: On the problem of assessment criteria". Paper presented at Was heißt hier eigentlich 'Evidenz'?, Fru€hjahrstagung 2015 des AK Methoden in der Evaluation Gesellschaft fu€r Evaluation (DeGEval), Fakult€at fu€r Sozialwissenschaften, Hochschule fu€r Technik und Wirtschaft des Saarlandes, Saarbru€cken, Germany, May. Available online at: http://www.degeval.de/fileadmin/users/Arbeitskreise/AK_Methoden/Hammersley_Saarbrucken.pdf

Hatcher, P. J., Hulme, C., Miles, J. N. V., Carroll, J. M., Hatcher, J., Gibbs, S., Smith, G., Bowyer-Crane, C., and M. J. Snowling. 2005. "Efficacy of small group reading intervention for beginning readers with reading-delay: a randomised controlled trial: Efficacy of small group reading intervention". *Journal of Child Psychology and Psychiatry* 47 (8): 820–827.

Henbest, V. S., and K. Apel. 2017. "Effective Word Reading Instruction: What Does the Evidence Tell Us"? *Communication Disorders Quarterly*: 1–9.

Hodder Education. 2000. *Hodder Group Reading Tests (HGRT) II*. London: Hodder Education.

Humphrey, N., Lendrum, A., Ashworth, E., Frearson, K., Buck, R. and K. Kerr. 2016. *Implementation and Process Evaluation (IPE) for interventions in education settings: An introductory handbook*. London: Education Endowment Foundation.

Koutsouris G., and B. Norwich. 2018. "What exactly do RCT findings tell us in education research? ". *British Educational Research Journal*. Advance online publication.

Koutsouris, G., Norwich, B., and J. Stebbing. 2018. "The significance of a process evaluation in interpreting the validity of an RCT evaluation of a complex teaching intervention: the case of Integrated Group Reading (IGR) as a targeted intervention for delayed Year 2 and 3 pupils". *Cambridge Journal of Education*. Advance online publication.

Leader-Janssen, E. M., and J. L. Rankin-Erickson. 2013. "Preservice teachers' content knowledge and self-efficacy for teaching reading". Literacy Research and Instruction 52 (3): 204–229.

Lovett, M. W., Frijters, J. C., Wolf, M., Steinbach, K. A., Sevcik, R. A., and R. D. Morris. 2017. "Early intervention for children at risk for reading disabilities: The impact of grade at intervention and individual differences on intervention outcomes". Journal of Educational Psychology 109 (7): 889.

McKenna, M. C., Kear, D. J., and R. A. Ellsworth. 1995. "Children's attitudes toward reading: A national survey". *Reading research quarterly*: 934–956.

Moore, G. F., Audrey, S., Barker, M., Bond, L., Bonell, C., Hardeman, W., Moore, L., O'Cathain, A., Tinati, T., Wight, D. and J. Baird. 2015. "Process evaluation of complex interventions: Medical Research Council guidance". *BMJ*: 350:h1258.

Nation, K., and M. J. Snowling. 2004. "Beyond phonological skills: broader language skills contribute to the development of reading". *Journal of Research in Reading* 27 (4): 342–356.

Neale, M. D. 1997. *Neale Analysis of Reading Ability II: Second Revised British Edition*. Windsor, England: NFER-Nelson.

Norwich, B., Koutsouris, G. and A. Bessudnov. 2018. *An innovative classroom reading intervention for Year 2 and 3 pupils who are struggling to learn to read: Evaluating the Integrated Group Reading (IGR) programme* – Project Report. Accessed from: http://www.integratedgroupreading.co.uk/

Patel, R., Jabin, N., Bussard, L., Cartagena, J., Haywood, S., and M. Lumpkin. 2017. *Switch-on Effectiveness Trial: Evaluation report and executive summary*, *May 2017*. London: EEF.

Pawson, R. and N. Tilley. 2004. *Realist evaluation*. Available online at: http://www.communitymatters.com.au/RE_chapter.pdf (accessed 27 February 2018)

Raffaele Mendez, L. M., Pelzmann, C. A., and M. J. Frank. 2016. "Engaging Struggling Early Readers to Promote Reading Success: A Pilot Study of Reading by Design". *Reading and Writing Quarterly* 32 (3): 273–297.

Rose, J. 2006. *Independent review of the teaching of early reading: Interim report*. London: DFES.

See, B. H., Gorard, S., and N. Siddiqui. 2015. "Best practice in conducting RCTs: Lessons learnt from an independent evaluation of the Response-to-Intervention programme". *Studies in Educational Evaluation* 47: 83–92.

Snowling, M. J., Stothard, S. E., Clarke, P., Bowyer-Crane, C., Harrington, A., et al. and E. Truelove. 2009. *YARC: York Assessment of Reading for Comprehension*. London: GL Assessment.

Speece, D. L., Schatschneider, C., Silverman, R., Case, L. P., Cooper, D. H. and D. M. Jacobs. 2011. "Identification of reading problems in first grade within a response to instruction framework". *Elementary School Journal* 111: 585–607.

Thomas, G. 2016. "After the gold rush: Questioning the 'gold standard' and reappraising the status of experiment and randomized controlled trials in education". *Harvard Educational Review* 86 (3): 390–411.

Topping, K., Miller, D., Thurston, A., McGavock, K. and N. Conlin. 2011. "Peer tutoring in reading in Scotland: thinking big: Peer tutoring in reading". *Literacy* 45 (1): 3–9.

Torgerson, C., Brooks, G., Gascoine, L., and S. Higgins. 2018. "Phonics: reading policy and the evidence of effectiveness from a systematic 'tertiary' review". *Research Papers in Education.* Advance online publication. doi: 10.1080/02671522.2017.1420816

Tschannen-Moran, M., and D. Johnson. 2011. "Exploring literacy teachers' self-efficacy beliefs: Potential sources at play". *Teaching and Teacher Education* 27 (4): 751-761.

Vaughn, S., Solís, M., Miciak, J., Taylor, W. P., and J. M. Fletcher. 2016. "Effects from a Randomized Control Trial Comparing Researcher and School-Implemented Treatments with Fourth Graders with Significant Reading Difficulties". *Journal of Research on Educational Effectiveness* 9(sup1): 23–44.

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

Figure 1. Participants' flowchart

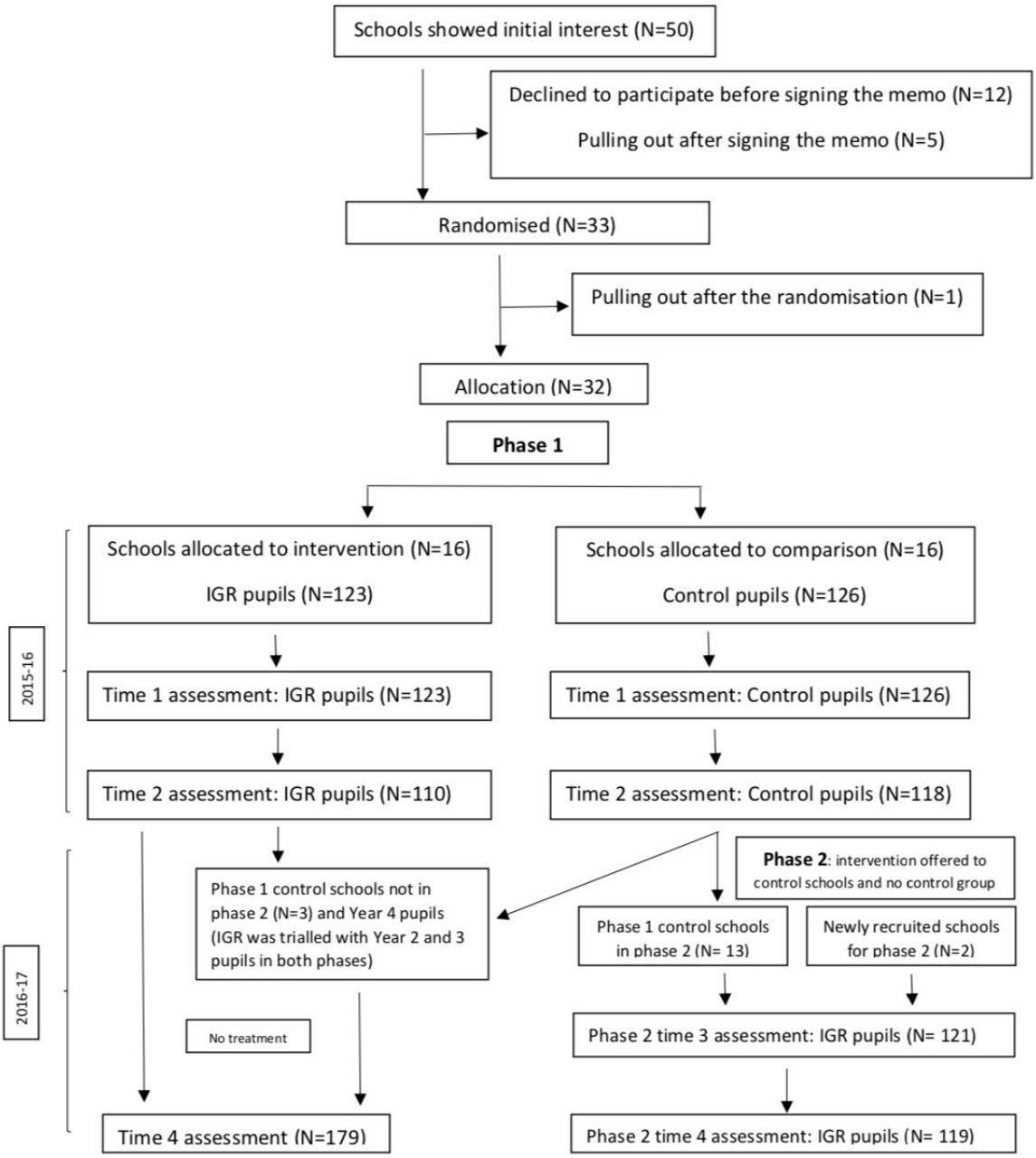

Figure 2. Control teaching: time spent with identified pupils compared to other class pupils

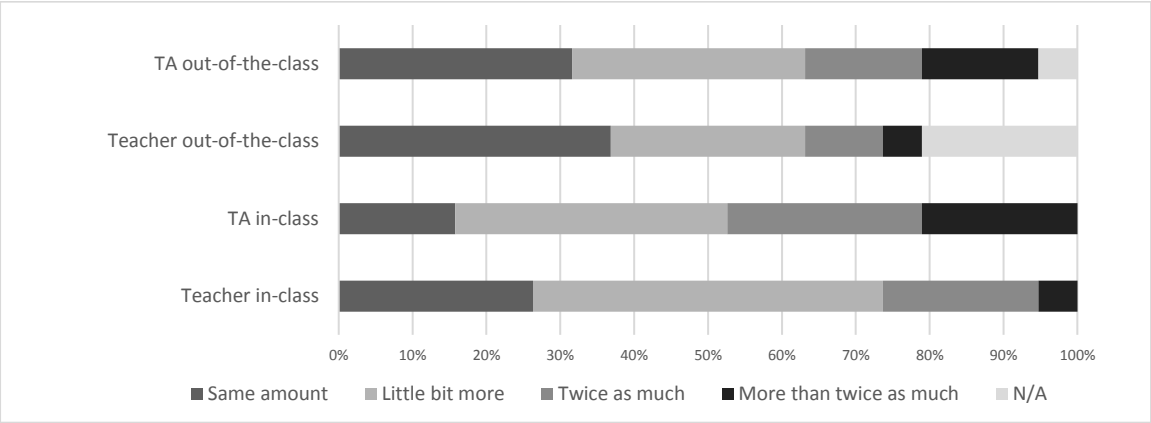

Table 1: The IGR teaching approach

Teaching strategies	The IGR routine	Linguistic Level
Teacher session (twice a week)		
Previous book		
Consolidation and recall	Drawings prompt story recall A game of GoFish	The sentence or phrase
New material		
Storytelling	Narrative familiarisation	The story itself
Phonological-visual mapping	A game of Lotto	Receptive vocabulary
Collaborative reading and problem-solving	Reading words in context	The new story between us
Words in more detail		
Analytic phonics	a SWAP phonics game	Non-story words out of context
TA session (in-between the teacher sessions)		
Consolidation	Word pelmanism	Story-specific words

Table 2. IGR and comparison groups

	Treatment phase 1 (N=131)	**Treatment phase 2** (N = 126)	**Control phase 1** (N=132)	Missing values treatment phase 1 (%)	Missing values treatment phase 2 (%)	Missing values control group (%)
Boys	84 (64%)	70 (56%)	89 (67%)	0 (0%)	0 (0%)	0 (0%)
Year 2	63 (48%)	63 (50%)	71 (54%)	0 (0%)	0 (0%)	0 (0%)
Ethnic background	25 (19%)	34 (27%)	23 (17%)	22 (17%)	7 (6%)	18 (14%)
EAL English as an additional language	19 (14%)	25 (20%)	18 (14%)	22 (17%)	7 (6%)	18 (14%)
SEN Education, Health and Care Plan (EHCP)	5 (4%)	6 (5%)	7 (5%)	22 (17%)	7 (6%)	18 (14%)
SEN (School Support) *	38 (29%)	49 (39%)	63 (48%)	22 (17%)	7 (6%)	18 (14%)
Pupil Premium	30 (23%)	37 (29%)	34 (26%)	22 (17%)	7 (6%)	18 (14%)
Child in Care	0 (0%)	1 (1%)	4 (3%)	22 (17%)	7 (6%)	18 (14%)
Mean phonics score (40 maximum)	24.0	28.4	27.4	36 (27%)	37 (29%)	61 (46%)
Mean HGRT standardised score *	90.4	90.0	85.9	4 (3%)	3 (2%)	4 (3%)
Mean SWRT standardised score	86.8	89.2	85.5	0 (0%)	7 (6%)	6 (5%)

Note: The variables where the difference between the treatment and control groups in Phase 1 is statistically significant at the 95% level are marked with an asterisk (*).

Table 3. Phase 1 results: IGR pupils

	Control Time 1	Control Time 2	Control n	Treatment Time 1	Treatment Time 2	Treatment n	Control T2 - T1	Treatment T2 - T1	IGR effect	Cohen's d	p value
HGRT standard score	86.1	90.3	117	90.5	92.2	112	4.18	1.77	-2.42	-0.23	0.20
SWRT standard score	85.6	89.6	118	86.3	89.5	118	4.03	3.10	0.13	0.01	0.93
HIFAMS: attitude to school	1.6	1.6	118	1.7	1.6	118	-0.07	-0.04	0.02	0.07	0.63
HIFAR: reading self-competence	3.4	3.5	117	3.4	3.5	116	0.11	0.13	0.05	0.06	0.68
HIFAR: reading attitude	3.9	4.0	117	3.9	4.2	116	0.14	0.22	0.10	0.12	0.42

Note: All averages were calculated with the balanced panel, i.e. pupils with valid observations at times 1 and 2. Cohen's d shows the effect size in standard deviations. The IGR effects were calculated by regressing the difference between the post intervention and baseline measures on the treatment status, controlling for year, gender, special educational needs (SEN) and English as additional language (EAL) status. Cluster-robust standard errors were applied. P-values come from the same models.

Table 4. Phase 2 results: IGR pupils

	Control Time 1	Control Time 2	Control n	Treatment Time 3	Treatment Time 4	Treatment n	Control T2 - T1	Treatment T4 - T3	IGR effect	Cohen's d	p value
HGRT standard score	86.1	90.3	117	90.2	96.2	119	4.18	6.01	2.49	0.24	0.15
SWRT standard score	85.6	89.6	118	89.2	92.7	118	4.03	3.46	-0.05	-0.01	0.96
HIFAMS: attitude to school	1.6	1.6	118	1.6	1.6	118	-0.07	-0.01	0.06	0.16	0.23
HIFAR: reading self-competence	3.4	3.5	117	3.3	3.5	118	0.11	0.16	0.09	0.11	0.42
HIFAR: reading attitude	3.9	4.0	117	3.9	4.0	118	0.14	0.17	0.05	0.06	0.69

Note: All averages were calculated with the balanced panel, i.e. pupils with valid observations at times 1 and 2 (3 and 4 for Phase 2). Cohen's d shows the effect size in standard deviations. The IGR effects were calculated by regressing the difference between the post intervention and baseline measures on the treatment status, controlling for year, gender, special educational needs (SEN) and English as additional language (EAL) status. Cluster-robust standard errors were applied. P-values come from the same models.

Table 5. Phase 1 results: class pupils

	Control Time 1	Control Time 2	Control n	Treatment Time 1	Treatment Time 2	Treatment n	Control T2 - T1	Treatment T2 - T1	IGR effect	Cohen's d	p value
HGRT standard score	106	108	573	105	108	586	1.7	3.3	1.7	0.13	0.24

Note: All averages were calculated with the balanced panel, i.e. pupils with valid observations at times 1 and 2. Cohen's d shows the effect size in standard deviations. The IGR effects were calculated by regressing the difference between the post intervention and baseline measures on the treatment status, controlling for year, gender, special educational needs (SEN) and English as additional language (EAL) status. Cluster-robust standard errors were applied. P-values come from the same models.

Table 6. Phase 2 results: class pupils

	Control Time 1	Control Time 2	Control n	Treatment Time 3	Treatment Time 4	Treatment n	Control T2 - T1	Treatment T4 – T3	IGR effect	Cohen's d	p value
HGRT standard score	106	108	573	105	109	598	1.7	4.5	2.8	0.23	0.03

Note: All averages were calculated with the balanced panel, i.e. pupils with valid observations at times 1 and 2 (3 and 4 for Phase 2). Cohen's d shows the effect size in standard deviations. The IGR effects were calculated by regressing the difference between the post intervention and baseline measures on the treatment status, controlling for year, gender, special educational needs (SEN) and English as additional language (EAL) status. Cluster-robust standard errors were applied. P-values come from the same models.

Table 7. HGRT measures for girls and boys in the control and treatment classes

Treatment status	Gender	HGRT baseline measure	HGRT post-treatment measure	HGRT difference (post treatment - baseline)	n
Control	boys	104	108	4.20	259
Control	girls	108	107	-0.56	302
Treatment phase 1	boys	105	108	3.28	292
Treatment phase 1	girls	105	108	3.22	292
Treatment phase 2	boys	104	108	4.39	313
Treatment phase 2	girls	106	110	4.64	284

Table 8. Long-term (phase 1) IGR effects

	Control Time 1	Control Time 4	Control n	Treatment Time 1	Treatment Time 4	Treatment n	Control T4 - T1	Treatment T4 - T1	IGR effect	Cohen's d	p value
HGRT standard score	85.6	90.2	106	90.2	90.8	110	4.62	0.55	-3.72	-0.32	0.05
SWRT standard score	85.4	89.1	108	86.2	89.1	107	3.70	2.93	0.39	0.04	0.80
HIFAMS: attitude to school	1.6	1.6	108	1.6	1.6	107	-0.05	-0.06	-0.02	-0.06	0.72
HIFAR: reading self-competence	3.4	3.4	107	3.4	3.4	105	-0.06	0.01	0.02	0.02	0.89
HIFAR: reading attitude	3.9	3.8	107	3.9	4.0	105	-0.07	0.10	0.18	0.18	0.24
HGRT standard score: non-IGR pupils	106	107	520	105	108	548	0.73	2.8	1.8	0.15	0.21

Note: All averages were calculated with the balanced panel, i.e. pupils with valid observations at times 1 and 4. Cohen's d shows the effect size in standard deviations. The IGR effects were calculated by regressing the difference between the post intervention and baseline measures on the treatment status, controlling for year, gender, special educational needs (SEN) and English as additional language (EAL) status. Cluster-robust standard errors were applied. P-values come from the same models.

Table 9. Phase 1 results: mean reading ages (years: months)

	Control Time 1	Control Time 2	Control n	Treatment Time 1	Treatment Time 2	Treatment n	Control T2 - T1	Ratio gains*	Treatment T2 - T1	Ratio gains	IGR effect**
HGRT	5:7	6:7	117	5:10	6:9	112	0:11	1.5	0:11	1.5	0:0
SWRT	6:1	7:0	118	6:1	7:0	118	0:11	1.5	0:11	1.5	0:0
HGRT: class pupils	7:3	8:5	573	7:4	8:8	586	1:2	2	1:4	2.2	0:1

Notes: *Ratio gains: reading gain in months divided by duration of programme in months (7); **Effects take into account rounding errors

Table 10. Phase 2 results: reading ages (years; months)

	Control Time 1	Control Time 2	Control n	Treatment Time 3	Treatment Time 4	Treatment n	Control T2 - T1	Ratio gains*	Treatment T4 - T3	Ratio gains	IGR effect**
HGRT	5:8	6:7	117	5:10	7:0	119	0:12	1.7	1:2	2	0:3
SWRT	6:1	7:0	118	6:3	7:1	117	0:11	1.5	0:10	1.4	-0:2
HGRT: class pupils	7:4	8:6	573	7:3	8:6	598	1:2	2	1:3	2.1	0:1

Notes: *Ratio gains: reading gain in months divided by duration of programme in months (7); **Effects take into account rounding errors

Table 11. Teacher self-efficacy

	Phase 1		**Phase 2**	
Reading teaching self-efficacy	T1	T2	T3	T4
Mean	7.4	8.2	7.2	8.0
Standard deviation	0.89	0.64	0.95	0.5
n	27	27	28	28
Cronbach's alpha	0.94	0.94	0.96	0.93
T test:	Mean difference = 0.7, t=5.5, df=26, p<0.001		Mean difference = 0.78, t=4.7, df=27, p<0.001	

Note: Only teachers who completed the questionnaires in both times 1 and 2 (or 3 and 4) were included in the analysis. Paired t-test was used.