Putting RCTs in their place: implications from an RCT of the Integrated Group Reading approach

Norwich, B. and Koutsouris, G.
Graduate School of Education, University of Exeter

Abstract

This paper describes the context, processes and issues experienced over 5 years in which a RCT was carried out to evaluate a programme for children aged 7-8 who were struggling with their reading. Its specific aim is to illuminate questions about the design of complex teaching approaches and their evaluation using an RCT. This covers the early development by the originator and work to develop and design a RCT funded trial. The experimental, process evaluation and case studies findings are summarised. It is argued that if RCT is the only credible evaluation approach, that there is no strong evidence for IGR use. But, if RCT as the first-choice evaluation approach needs to be supplemented by process evaluation, then a positive process evaluation might save IGR for further development and evaluation trials. However, it is suggested that conceptualising IGR as a complex teaching intervention also raises questions about RCT as the method of first choice. It is argued that a Designed-Based Research approach to scaling up IGR, an example of a Design & Research approach, might have been tried. The reasons why this was not done are explored with implications for the place of RCTs in improving teaching and learning.

Introduction:

Questions of what works well or best in enabling children and young people to learn is a continuing question in education for teachers, parents and policy-makers. Debates about how to establish such knowledge and the nature of this knowledge has divided educational researchers along epistemological and value lines, on one hand, and reflects different methodological positions, on the other. There is currently an alignment between 'what works' language - what is called 'evidence-based practice'- and the use of randomised controlled trials (RCTs). Most of the RCT literature in education research journals is either theoretical pieces, RCT studies or synthesis reports. The former tend to analyse the questions in abstract terms, while the latter tend to be specific and technical. There have been no papers which describe the experience and process of designing and conducting an RCT in the light of the theoretical and methodological issues. The purpose of this paper is to describe the context and process as well as the issues experienced over a period of 5 years in which a RCT came to be designed and carried out in the area of early reading teaching for children aged 7-8 who were delayed and struggling with their reading. Its specific aim is to use this example of an RCT to illuminate questions about the design of complex teaching approaches and evaluation approaches.

The reading approach which was evaluated using a RCT type of design came to be called the Integrated Group Reading (IGR) programme. It was developed by Jan Stebbing who had been working on the programme materials and trying out the approach as a teacher since 2009. This paper describes these early stages briefly and how Jan came to work with the first author of the paper which then led to the trial which was funded by the Nuffield Foundation. This includes an account of how IGR fits with current positions about the teaching of early reading, the response to teaching / instruction (RTI model) and inclusive teaching. The experimental findings and the process evaluation are summarised and the case studies used to further understand what was found. The process is discussed in relation to the academic-practitioner partnership, evaluating complex interventions in actual classrooms and methodological and philosophical issues. The paper concludes with some consideration of future research and development prospects.

What works and evidence-based practice

There has been a growing international movement since the 1990s to establish what has been called evidence-based practice in education. This was about educational research as not providing adequate quality evidence for policy and practice (Hargreaves, 1996), which led to much debate since then about teaching as a research based profession (Hammersley, 2007). Evidence-based practice has been closely linked to a focus on identifying 'what works' in teaching which has been translated as mostly meaning randomised control trials (RCTs).

More recently with Government support this approach has been presented as building evidence into education in the UK (Goldacre, 2013). It has been suggested that by collecting 'better evidence about what works' (page 1) and setting up a culture where this evidence is used routinely, children and young people will have better progress outcomes and teachers will be more independent professionally. Goldacre (2013) presents this as promising a 'revolution' to empower teachers. Torgeson (2009) argues that stakeholders need to know' what would have happened to children if they had not been exposed to an intervention' (page 313) and that a robust way to do this is using RCT designs. For Hutchinson and Styles (2010) an RCT is to be considered as the 'first choice' - what has also been called the 'gold-standard' - to find out what interventions work. The main reason given is that RCTs eliminate selection bias and so can support causal conclusions. This is presented by contrast with the 'misleading results from non-experimental work which has inadequately controlled for selection bias' (page 7). RCTs are said to promise 'quick and digestible conclusions of programme effectiveness that avoids lengthy caveats' (page 7). However, Katsipataki and Higgins (2016) have argued that meta-analyses of RCTs provide evidence of 'what has worked', suggesting what is likely to work in the future. But, this distinction between the generalised 'what works' and what has worked is a reminder that what has worked might not always work.

Hammersley has questioned whether the rise of RCTs has revived the paradigm wars (Hammersley, 2008), but whatever the scale of conflict, what matters is that the critiques of the 'what works' approach have persisted since its manifestation two decades ago (Thomas and Pring, 2004). More recent criticism has questioned what is seen as a narrow view of science expressed by advocates of evidence-based practice. For example, Torrance (2013) and Furedi, (2013) question whether RCTs can prove what works in a simple way while advocating for teachers to have more time to reflect on their practices. Some aspects of the debate reflect a clash of purist views for or against RCTs as the 'gold standard', a revival of the 'paradigm wars' and an acceptance of the incommensurability position about philosophical underpinnings of education research.

This idea that paradigms are mutually exclusive has been questioned by Toulmin (1972) who presented an evolutionary view of conceptual change that focused on what is common to all argumentation. So, it is interesting how the differences over RCTs are currently presented in less purist terms. Thomas (2016), for example, now urges a diversity of enquiry methods, but does not advocate abandoning experimental research. His position is that RCT designs might be relevant to less complex interventions, while teaching in schools, like psychotherapy, involves human relationships in which there is a 'systematic unpredictability', a term he derives from MacIntyre (1985). This is sometimes called the problem of reflexivity (Flanagan,1981), in which the objects of a scientific inquiry are subjects with agency. Classroom teaching and learning are done by teachers and pupils who can also take up a perspective on an experimental intervention and act according to their own goals and meanings. Acting in specific contexts they can confound classroom interventions and so undermine experimental controls. So, generalisations from RCTs are risky. Cartwright (2009), for instance, warns about the 'vanity of rigour in RCTs' and that they are not the only 'game in town'. She reminds us of the distinction between knowledge of the efficacy of a teaching method – can it have effects in test conditions? – and the effectiveness of a method – can it work for other learners under other conditions? What can be shown to work in some conditions might not work in different less controlled conditions. So, an RCT might show the valid claims that a method has certain learning effects in specific conditions (high internal validity), but this knowledge might not be valid when the method is used in different conditions (low external validity). RCT generalisations might not only be risky, but they might also 'decay' over time, as Cronbach (1975) pointed out over 40 years ago.

A contemporary perspective on evidence-based practice using RCTs is given by some RCT practitioners. Siddiqui, Gorard and See (2016) recognise that educational interventions are often complex and that outcomes could be due to factors not focussed on in the impact evaluation. So, they argue for process evaluations to be part of the evaluation design to understand the context of the outcome results from the RCT design. Connolly, Keenan and Urbanska (2018), from a similar perspective, have addressed some of criticisms of RCTs through a large scale systematic review of international RCTs from 1980-2016. Based on over a 1000 RCTs, most of which have been done in the last decade and the USA, though also Europe and UK, they show the diversity of interventions and settings in which RCTs have been done. These authors also show a sizeable number of RCTs (38%) did not ignore context by using process evaluations.

Humphrey (2018), someone who is a RCT practitioner. He identifies a range of issues with RCTs when used as the only research approach, the first of which is that it could suppress other research approaches and undermine practitioners' efforts to improve their practice. A second issue is the pragmatic one that there is simply not enough RCT based evidence to inform practice questions. This can be illustrated in the area of literacy, for instance in deciding between synthetic and analytic phonics approaches (Torgeson, Brooks, Gascoine and Higgins, 2018). In addition, there are issues in translating RCT findings into practice. A database search done for this paper to find studies about the effects of implementing evidence-based practice in literacy teaching shows very little implementation evidence, though several sources about how to implement evidence-based practice knowledge exist (Sharples, Albers and Fraser, 2018). There is also evidence in the social emotional intervention area that the specific intervention effect sizes of efficacy trials become reduced when the intervention is used under ordinary (effectiveness) conditions. There could be similar reductions in the literacy teaching area. Implementation variability is an issue in some areas of RCT efficacy and effectiveness studies and will be discussed later in this paper.

Though there were RCTs in education research before the establishment of the Education Endowment Foundation (EEF) in 2011, the EEF has contributed much to the growth in RCTs. It required that the impact of interventions be evaluated with RCTs where possible, and that quasi-experimental design (QED) be used when randomisation was not feasible (Education Endowment Foundation, 2018). There has been a move more recently for the EEF to emphasise implementation and process evaluations when in the past they varied in size and purpose (Humphrey et al., n.d.). So, an implementation and process evaluation is ’commissioned alongside every impact evaluation to understand how a project is implemented on the ground and the elements of successful delivery’ (page 4). This marks the recognition of the value of what is called ’mixed methods’ and acknowledges that the context and implementation process of an intervention is important and by implication recognises the limitations of straight RCTs (Wyse and Torgeson, 2017).

Though there have been moves that reflect some common ground between advocates and critics of RCTs as the ‘gold standard’, there may still be some key differences. Some may still favour a hierarchy in which RCTs are the first choice, while accepting QEDs only if RCTs are not feasible (Hutchinson and Styles, 2010), while others might prefer a matrix in which RCTs are one amongst many available designs (Hammersley, 2013; Thomas, 2016). For those with more plural and pragmatic inclinations, the selection of designs is more about the project purposes and less about adopting RCTs to secure general causal inference as the main priority. This difference might also be seen to be about the place of research in the development of teaching approaches. A distinction can be made between a traditional research and development (R&D) model in which knowledge is established (a summative type of knowledge) and then applied to practice and a development and research (D&R) model (Bentley and Gillinson, 2007) in which an innovative teaching approach is developed and then evaluated formatively. This D&R model, which includes forms of action research and design-based research approaches, is flexible, teacher oriented and more grounded in the needs of teachers and pupils.

In asking the question of how research can contribute to the improvement of teaching, Lewis, Perry and Murata (2006) identified two broad routes to improving practice, the Local Route and the General Route. Building on this distinction Norwich (2014) identified these routes as allied to the D&R and R&D models respectively. In this formulation, the Local Route involves developing practice in local contexts where the emphasis is on flexibility and local ownership. The D&R model is about innovation, continuous evaluation and adaptation. Methodologically this route is associated with Action research and Design-based research approaches. By contrast, the General Route is about establishing general causal claims and then disseminating these through the R&D model. The emphasis is on random allocation and fidelity of implementation, starting with efficacy trials and moving on to effectiveness trials as the basis for dissemination. Lewis et al. (2006) suggest that there are strengths and limitations to both models, which can also be seen to lie along a continuum, rather than forming a clear dichotomy, and to be interactive, where a D&R model could be basis for R&D and vice versa.

Despite the EEF approach now involving ‘mixed methods’ with RCT supplemented by implementation and process evaluations, its overall approach to improving teaching and learning has no place for a local route or D&R model (EEF, 2016). Its cyclical approach is based on generating evidence and then using evidence (the R&D model). It represents this as a sequence of making grants to test the impact of high potential projects, publishing independent rigorous evaluation, scaling up promising approaches and programmes and supporting teachers to use high quality evidence. Though this last phase is seen to lead back to grant-making for high potential projects, no part of the cycle recognises a place for local route development and evaluation research that involves teachers. It was a local route approach which was the start of the IGR approach which is the focus of this paper.

At the start:

Jan Stebbing originated what is now known as the IGR programme (originally Small Group Integrated Reading). The approach was rooted in her primary school teaching experience, her learning support and consultancy work, and her longstanding experience that, given the right kind of informed professional teaching approach, all children can learn to read. She had some Higher education research experience and a Master’s degree that involved literacy research. IGR also arose from her writing children’s early reading materials for well-known publishers and for BBC Education. In 2009 she began developing in southwest English schools a set of core materials for Year 2 and 3 children who were reading-delayed. Her emphasis was on the way current professional knowledge bases can be cross-referenced and integrated for the benefit of teachers working with children needing to progress in a systematic and lively way.

She approached the first author as someone who worked in a university close to where she was based and who had a general interest in special educational needs and inclusive education. His particular interests were in the relationship between general and specialist teaching and how teachers could develop inclusive ways of organising their classroom teaching. Jan was interested in the scaling up and systematic evaluation of what she had developed in the form of principles, procedures and materials for the small group teaching of children delayed in their reading. The first author was interested in this group-based approach with teacher-ready materials that could support a more inclusive way of supporting children with early difficulties in reading. He had recent experience of a school-based RCT and process evaluation study which interested Jan. These complementary interests were the basis of the partnership that developed.

The first author asked Jan Stebbing to write a brief paper about the principles that formed the basis of the teaching approach and the evaluation of her own teaching of Year 2 and 3 children with delayed reading in several primary classes (Stebbing, 2013). Drawing on Government end of Key Stage 1 performance data, this paper argued that the overall improvement in the percentage of children reaching expected standards in reading since 2006 could be attributed to the introduction of systematic phonics teaching but that, equally, between 7% and 18% of children across local authority areas were not responding to current literacy teaching (DFE, 2013). This provided the rationale for something additional to the currently available systematic phonics programmes, involving a multi-perspective intervention, to enable improved reading progress for delayed readers. This involved integrating systematic phonics teaching, systematic high frequency/exception word learning, and the reading of appropriately-levelled short reading books and story-specific word games with the learning taking place in the context of existing classroom literacy practice.

IGR principles and practices

In its initial conception the distinctiveness of IGR as a systematic approach was framed in terms of four principles:

1 Using a classification of English orthography in terms of phonic and high frequency word progression by ease of acquisition; cross-referencing this progression with early narrative texts and reading ages,
2 Distinction between expressive and receptive language,
3 Combining (1) and (2) to make learning to read congruent,
4 The importance of the small group dynamic.

IGR was based on a sequence for the teaching of grapheme–phoneme correspondences and out-of-context sight word acquisition of the first 100 (and next 200) High-Frequency words from the Letters and Sounds programme (Primary National Strategy, 2007). It was the cross-referencing and integration of this progression into narrative texts up to a reading age of 9 years, which was a central part of the IGR approach. This was the basis for the IGR reading books, some of which were rewrites of existing stories, other books were original. The second principle applied the process of language learning - where receptive language is in advance of expressive language - to early reading. In IGR children are introduced to orthography beyond their current phonic and high frequency word knowledge through story-specific lotto games. These games also enable teachers to notice individual children's phonological-to-visual matching and mismatching. The fourth principle deploys a small group dynamic in which children learn to read through a combination of choral reading, reading themselves, and learning from others' reading. Children are taught in IGR to approach the reading of text (story or rhyme) as they would the joint singing of a song, to listen and follow closely while others read alone. It was assumed that the experience of seeing another child stumble or hesitate prompts immediate cognition and problem-solving behaviour in the listeners.

Initial IGR evaluation and a further IGR trial

The first evaluation of an IGR approach was conducted by Jan Stebbing, teaching identified groups of delayed readers in Years 3 and 2 during typical class teacher-taught literacy lessons in two South-west English schools (2010-2013). IGR was used once a week for 30 minutes over a period of three terms. The Neale Analysis of Reading Ability was used to assess accuracy and comprehension, but could only be used with Year 3 children (ages 7 to 8 years old) because of their starting levels. For Year 3, starting reading accuracy was between <6.01 and 7.03 years. Results showed that accuracy and comprehension mean gains were 10.2 and 11.8 standard scores respectively. Mean ratio gains (ratio of reading age gains in months divided by the period of time in months) were 2.8 for accuracy and 3.6 for comprehension (sometimes judged as useful and substantial respectively; compared with a ratio gain of 1 which is standard expected progress).

Once it had been decided to seek funding to try out IGR nationally, it was important to introduce other teachers to the use of IGR as part of their group-based classroom organisation with teaching assistant support. This meant developing a training and support programme for teachers to understand not only the principles and procedures of the small group teaching, but also how IGR could be taught during usual literacy lessons using a group-based organisation. This is when the term ‘integrated group reading’ became established, as ‘integrated’ captured both the integration of different reading approaches for group teaching and the integration of those struggling to learn to read into a class organisation model where the teacher and teaching assistant rotated round each group on a weekly schedule. This was a key point in the formation of the IGR approach. An experienced primary PGCE tutor who specialised in literacy then became involved to assist in developing the class management aspects. Together with Jan Stebbing and the local literacy adviser they prepared 7 Year 2 and 3 teachers in 4 schools in one authority in the principles of IGR and the use of IGR materials in their classrooms. The trial lasted for one term in 2015 and provided the basis for the development of a training and support programme for the introduction of IGR on a larger scale.

The 30-minute IGR teacher-led lesson followed the sequence shown in Table 1; one way in which the group-based management of literacy lessons was organised is shown in Table 2. At this time, it was decided to use IGR more intensively than once a week, increasing it to twice a week for a reduced time period. A challenge for class organisation was to have enough sessions for the teacher to work with each group at least once a week where there were four other groups of pupils in addition to the IGR group having two sessions a week. In Table 2 there is an additional 30-minute session on one of the days, but there were also other ways of arranging this extra session.

Table 1: IGR learning model and cycle:

Teacher		
Activity	Process	Linguistic level
-Drawings from previous book prompt story recall -play story-specific Go-Fish game:	Consolidation and recall from previous book	Sentence and phrase
New book: storytelling	Narrative familiarisation	Story itself
Play story-specific Lotto game	Advance organiser (new words) Phonological-visual mapping	Receptive vocabulary
Reading new story between us	Collaborative reading and problem-solving	Words in story context
Swap phonics game	Words in more detail (analytic phonics)	Non-story words out of context
Teaching assistant		
Chose a picture from their books to draw in detail	Supports comprehension	
Write a brief accompanying sentence as explanation		
Re-read their books individually	Problem solving	
Play story-specific Word Pelmanism	Consolidation and recall at the level of the word	

Table 2: Classroom organisation for IGR

	Monday	Tuesday	Wednesday	Thursday	Friday (1)	Friday (2)
Group 1	TA					Teacher
Group 2		Teacher			TA	
Group 3			TA		Teacher	
Group 4				Teacher		TA
IGR group	Teacher	TA	Teacher	TA		

Conceptualising IGR as an inclusive targeted intervention

It made sense to think about the design of the IGR approach in terms of the wave model (Rose, 2009) given its international recognition for linking general to specialist teaching and how it has informed policy in the UK since the 2000s. The three-part wave (sometimes called tier) model distinguishes between wave 1 (universal or *Quality First* teaching), wave 2 or targeted teaching and wave 3 specialist teaching. However, in its current Response to Intervention (RTI) use, the relationship between the waves is unclear. Current practice is to provide 'Quality First' teaching that is meant to be differentiated, but might not be differentiated enough for pupils struggling to learn. So tailored teaching for those not progressing at the expected rate with targeted or specialist teaching is often offered as additional withdrawal sessions with people other than the class teacher (e.g. TAs). This could have two risks: i. creating 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 learning time lost – for instance, it has been found that children who had immediate access to additional support rather than waiting to fail, had improved reading outcomes at the end of Year 1 (Al Otaiba et al, 2014).

A systematic review of 64 international experimental evaluations between 1970-2017 focussed on how school-based reading interventions for struggling readers aged 5-8 years were delivered in a wave / tier 2 or targeted form by the RTI model (Stentiford, Koutsouris and Norwich, 2018). The review showed that the wave 2/3 interventions targeting pupils who did not respond to whole class teaching (wave / tier 1) were almost all delivered in pull-out sessions by people other than the classroom teacher. Assuming that the teaching in these evaluations reflect current practices, it is clear that pupils who are identified for wave 2/3 support might have less access to their teacher's time and expertise. It is in this context that IGR introduces tailored targeted teaching (wave 2) in a 'Quality First' (wave 1) teaching setting. It is in this sense that IGR has the potential for being an inclusive teaching approach (see Norwich and Koutsouris (2019) for discussion about the issues of inclusive teaching and IGR).

Designing the IGR programme

The design of the IGR programme for the funded evaluation was based on the pilot work and involved a team that included in addition to Jan Stebbing and the PGCE tutor four local authority literacy advisers who had expressed interest in the approach Preparing teachers was to be done through a national one-day workshop for all teachers followed by local training of teachers and teaching assistants in the four areas. Literacy advisers had a key training and support role for IGR teachers. The IGR materials, which consisted of 52 titles which were to be the hub for each lesson, consisted of learning packs containing book copies and related games. These had to be designed and printed in sufficient numbers.

By the stage that the proposal was submitted to the Nuffield Foundation in outline the rationale for using IGR had been further clarified. The point that practitioners found it hard to know how best to address the practical problem of integrating relevant research-informed teaching approaches for the benefit of reading-delayed children, was subjected to a detailed review of the literature that then served to help underpin and support the case for evaluating IGR. For instance, a US source was found which also talked about integrating research-based findings into a coherent system for reading teaching in a multi-tiered literacy intervention for children at risk of reading difficulties (Fien et al., 2014). Other elements of IGR that had a research base were, namely i. how early progress in reading depends on children's oral language skills (Muter et al., 2004), ii. how word games can accelerate reading skills for children who struggle to learn to read (Charlton et al., 2005), iii. detailed responses to reading in situations, as is built into the IGR approach associated with Reading Recovery (What Works Clearinghouse, 2013) and iv. a 'chiming in' system to support children as they read, with links to a 'paired reading' approach (Topping et al., 2011).

However, and in addition, IGR had the dual aspect of both integrating different professional knowledge elements into a carefully sequenced set of group activities led by a teacher and supported by teacher assistant follow-up on one hand, and organising IGR teaching as part of a group-based class organisation system on the other. This opened up design options. It was presumed that IGR group teaching would be led by a teacher and not an assistant, but it could have been done by a teacher other than the class teacher, e.g. a SEN coordinator, another class teacher or a literacy specialist, such as Jan Stebbing when she first evaluated IGR. The other question was about where and when it took place. IGR group teaching could take place as part of a group class organisation, but it could also be done outside the classroom during lesson time or class lesson times, e.g. during assembly. These options depend on who teaches IGR and a third consideration, whether IGR is an additional programme to what is provided for all (wave 1,'quality first' teaching') or a replacement of wave 1 teaching, even if a time-limited one. These three sets of options: i. class teacher taught – taught by another teacher, ii. in – out of classroom / during – outside literacy lesson time and iii. as additional – replacement of wave 1 programme are inter-linked.

In the funded IGR evaluation, IGR was to be taught by the class teacher (not other teachers), inside the classroom and during lesson time (not outside the classroom and outside lesson times) and to be a replacement targeted programme (not additional to wave 1 teaching). The importance of this design position became evident when the results of the evaluation were known, a point to be pursued later in this paper.

Designing the evaluation for submission

The perspectives about what works and evidence-based practice were considered in designing this national IGR evaluation. Both principled and pragmatic factors influenced the final design. A purist RCT design with no process evaluation was not favoured for the reasons discussed above. A combined RCT and process evaluation was preferable reflecting the mixed methodological consensus that has developed. But, it is interesting in retrospect that a purist rejection of a RCT design in favour of a local route or D&R approach, such as a Design-Based Research (DBR) approach was not considered seriously. This was not because the first author was unaware of such designs, as he had used DBR in other projects. It is probably because a more summative generalising style of evaluation was seen to be needed. This came from the partnership between the originator of IGR who wanted to establish the credentials of IGR and the first author who was interested in being involved in another RCT style evaluation. Both shared the view that RCT evaluation conclusions were the currency for raising the IGR profile in a policy and practice climate.

So, a mixed methodological design was designed for submission to the Nuffield Foundation. combining two separate strands: i. a cluster randomised control trial (CRT) with ii. a process evaluation of its implementation and of teachers' perspectives. The switch in reference from RCT to CRT reflects the randomisation of schools as clusters involving a slightly different design from those usually written about. The CRT was designed to assess the impact of IGR teaching for Year 2 and 3 children who most struggle with reading on reading accuracy, comprehension and attitudes compared to similar children experiencing the usual approaches to teaching reading. In order to check whether IGR focused on a sub-group might affect the progress of others who were not receiving IGR teaching and learning in other groups, the reading progress of the other children in IGR teaching and control classes was also assessed.

The process evaluation was informed by realist evaluation principles (Pawson and Tilley, 1997), through its focus on context and processes (mechanisms), though this was not emphasised in the submission proposal. It aimed to examine the school and class contexts in which the IGR programme was used, how IGR was implemented as part of the whole class organisation model, whether teachers using the IGR methods increased their self-efficacy to teach pupils struggling to learn to read, the perspectives and experiences of teachers and children about using the IGR programme and how reading was taught in the control classrooms.

The project was designed in two phases. Phase 1 randomised the schools into intervention and control groups, ran IGR in intervention classes for two school terms while teaching was as usual in control classes, with pre, post and final one year follow up assessments for all classes. Control schools were offered IGR in phase 2 in the second year, with a simple pre-post assessment. Process evaluation was conducted in both phases.

Method detail and issues

For word length reasons only key points and issues will be raised in this section for the CRT, not the process evaluation (full details are available in Norwich, Koutsouris and Bessudnov, (2018). Recruiting schools was a major challenge despite the support of the four local advisors who had recruited at least 40 schools to agree in principle to participate at submission stage. Some of these schools backed out at the point of signing an agreement to undertake what was involved for various practical reasons, e.g. staff changes. At one point 50 schools were actively interested, but some also backed out at the signing stage and even some head teachers who signed the agreement backed out later on just before the start. In the end the project was 8 schools short of target number of 40. With 32 schools there were 64 classes, so with between 110-120 pupils per group, medium size effects could be identified with a 0.8 power.

One of the design issues was about the control pupils and what teaching they would receive. This is about the nature of the experimental control in a RCT or CRT and what IGR is being compared to as well as the cost of running a trial with multiple comparisons. IGR (as a targeted intervention) could be compared with teaching as usual (which could range from no targeted teaching to considerable targeted teaching). While IGR as an intervention could be controlled to some extent, teaching as usual was more difficult to control ethically. If schools were already providing targeted support then it would be questionable ethically to ask them to withhold this support. Another option was to arrange another comparative intervention, perhaps a targeted phonics programme which could be controlled to some extent and would be delivered alongside IGR and teaching as usual. However, this involved additional costs not available in this project. This issue was resolved by comparing IGR with teaching as usual and monitoring the additional teaching provided in control classes.

One proposal reviewer called for a more independent type of evaluation with the programme arm of the project detached from the evaluation arm, as practised by the EEF. The risk was that with the IGR originator leading the programme training and support team, this might influence the evaluation team. So, the evaluation team's separation from the programme team was monitored by an outside evaluation advisor who reported to the funders. Assessments of children delayed in reading for IGR and control groups were done by assistants who were blind to their group status, randomisation was done after baseline assessments and done by the statistician who was detached from the programme arm. This worked well as confirmed by outside evaluation advisor, but there were some differences between IGR originator and evaluation team about interpreting the findings in the final report and summary, but not over the findings themselves.

There were expected issues over identifying the 4 pupils delayed in reading for the IGR or control group. Teachers who had taught the pupils completed a multi-dimensional reading rating form, based on an adapted US research informed scheme, to give a holistic overview of those struggling with reading. This was instead of just using test reading scores, which teachers were also advised to take into account in their ratings. However, there was sometimes an issue about confining selection to just 4 children and what to do about others considered to need additional support. It was suggested that IGR materials could be used with these others in their groups even if they were not receiving the intensive IGR intervention. For some teachers the suitability of IGR for a few children and reading level variations in the IGR group was a problem that emerged once they started using IGR which had to be dealt with.

There were also issues over selecting a reading accuracy and comprehension test suited to those with minimal reading that would at the same time be relevant to some children who could not score at the floor of the test while still being a useful measure of change as their reading improved. The Neale Analysis of Reading Ability was suitable in some ways but had out of date norms. The York Assessment of Reading Comprehension (YARC) was more recent but had complex administration procedures requiring a child to read two passages to receive a score. It had an associated word reading test which was used for all and so acted as a back-up for those not scoring on the YARC. A group sentence and word accuracy and comprehension reading test (Hodder Group Reading Test) was used for the whole class including those identified for IGR and control group children. But, the problem of the suitability of reading measurements was evident throughout the project. In addition, to measure attitude and perceived reading competence the Chapman Reading Self Concept Scale (RSCS) and to measure attitudes to school, the 'How I feel about my school' scale (HIFAMS) were used. The research assistants administered these directly, reading out the statements aloud. Reports from them indicated that some children did not engage with these scales despite their design as suitable for this age range. Doubts about the validity of these tests were also raised by there being no change in mean scores for them from before to after the intervention for the IGR group by contrast with evidence of positive attitude change in the process evaluation.

Key conclusions

Experimental evaluation

The 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. Participating children in schools using IGR in phase 1 and phase 2 made the same degree of progress in reading accuracy and comprehension (no statistically significant differences), compared to similarly struggling children in control schools, which the process evaluation showed were mainly using phonics approaches. The mean reading progress in intervention and control groups was equivalent to 11 months over the 7 months in phase 1, (often seen as a 'modest impact'; Brooks, 2016); and mean progress in phase 2 intervention groups of 14 months over the 7 months, (often seen as 'useful impact'). There were also no statistically significant changes for reading and school attitude in either the treatment or control group.

There were no statistically significant differences between boys/girls and Year 2/3 pupils in their responses in the IGR and control classes. Some analyses showed that pupils having English as an Additional Language (EAL) and being identified for Pupil Premium made significantly greater gains with IGR, but these findings were not replicated to the same level of significance across phases. However, the initial hypothesis that IGR in the classroom would not have any negative effect on the classroom pupils not having the intervention was confirmed. In Phase 1 there was no statistically significant difference in gains between treatment and control classes for non-IGR children.

Process evaluation

Overall, participants in both phases were enthusiastic about the intervention, the project materials, and accompanying support. Teacher-reported outcomes for IGR pupils included increased confidence, motivation and interest in reading, and improved reading, oral language and social skills. However, some teachers were concerned that these gains had not yet transferred outside of the IGR group setting. Most pupils were not worried about being seen in a low attainment group and did not see IGR as an intervention, but as an exciting classroom activity. Other class pupils were often very interested in the IGR resources, especially the games.

IGR was used with varied fidelity, and many teachers had limited understanding of the theory underpinning the programme, which could partly indicate a training limitation. In phase 1, the programme support team had to produce a table with acceptable and unacceptable implementation variations for teachers to take into account. Some departures from the suggested methodology were seen to be justifiable (such as, slowing down the pace of the programme in response to pupils' needs), whereas others were less acceptable (for instance, delivering all programme sessions in withdrawal sessions).

Significantly in the context of this trial, control schools did not just continue with typical teaching; teachers recognised that control pupils had significant additional needs, so they also had a great deal of additional, mainly phonics-based teaching input, making what was being compared to the experimental evaluation varied and complex.

Case studies

The project was based on the assumption that high fidelity IGR teaching would lead to relatively high reading gains. This was tested through a series of case studies of IGR teaching where there was a match between IGR fidelity and reading gains (high-high and low-low) and a mismatch (high-low and low-high). Data from two different teachers showed that when high reading gains followed high IGR teaching fidelity, several supportive factors were identified, e.g. teacher and pupil enthusiasm, school leader and adviser involvement, teacher understanding the theory and rationale of IGR and the IGR model fitting the pre-existing reading organisational arrangements. When low gains were followed by low fidelity in the cases of two other teachers, the above factors were not identified.

By contrast, in teacher cases where low or no reading gains followed medium to high IGR fidelity, there was evidence of barriers to reading progress, such as, a mechanical teaching approach that did not engage pupils, having a TA who could not manage the other groups during IGR teaching and unsatisfactory teacher job-sharing arrangements. In the case of another teacher where quite high reading gains followed low IGR fidelity, there was evidence that the low fidelity score was due to a change in teaching which did not affect the otherwise high-quality IGR teaching These case analyses are taken to suggest that IGR is not a simple intervention that can be applied or not irrespective of its teaching context. Its introduction as a programme was involved in a complex of interactions, resulting in what has been called a complex intervention (Moore et al., 2015). A complex intervention is one with interacting parts, difficulties in implementing working at several organisational levels, which is an apt description of the IGR approach

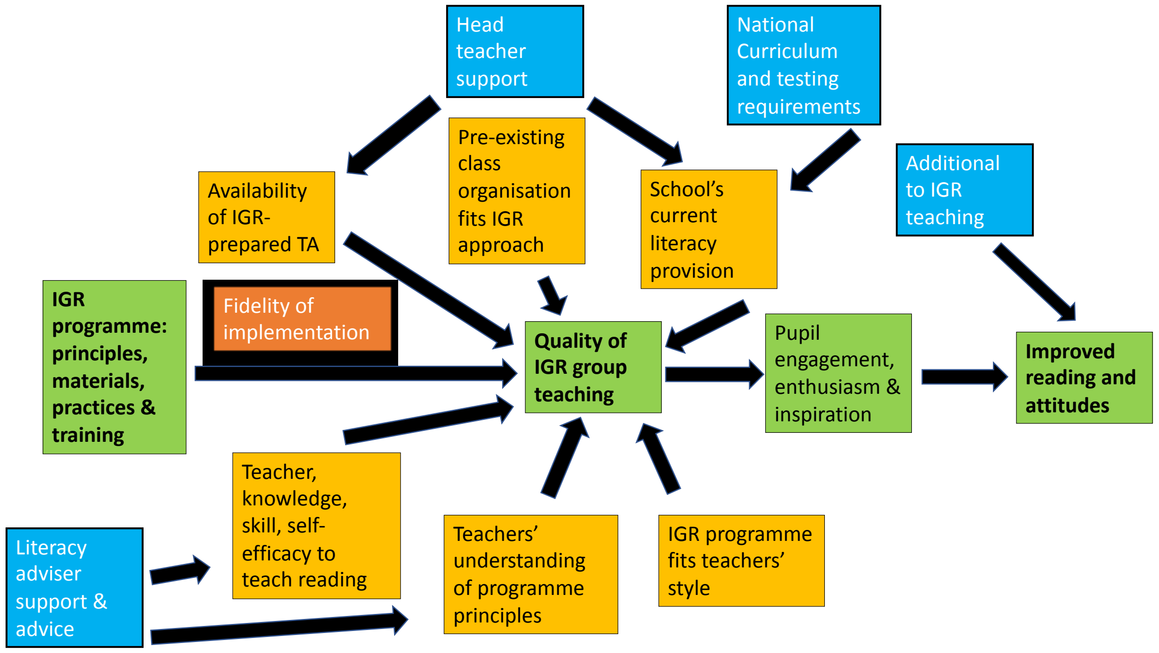
Discussion and conclusions

This paper has described the context and processes as well as the issues experienced in designing and evaluating IGR using a RCT. Its specific aim is to use this RCT example to illuminate questions about the design of complex teaching approaches and evaluation approaches.

Part of the complexity of IGR as a targeted teaching approach was due to its dual aspect as both a group based teaching programme that i. integrated a mix of principles and procedures and ii. as a classroom group organisation approach that enabled intensive targeted teaching while allocating time to all class groups. The IGR group approach could have been detached from the group organisation, or delivered once-weekly in the normal way within class group reading organisation, which an informal follow up of IGR project participants shows has been done in some schools. In retrospect it is clear that the design of the IGR evaluation could have compared two versions of IGR with teaching as usual. IGR (1), the one used in this trial, is a replacement for usual teaching of children struggling to learn, is taught by class teacher in the class and during lesson time using a group organisation model. IGR (2) is additional to usual class teaching, taught by a teacher outside the classroom and outside lesson times. The teacher could be another teacher (e.g. SENCo) or the class teacher (e.g. in assembly time).

It is evident from the process evaluation and case studies that IGR (1) as used in this project involved a range of local contextual factors and mediating processes that could affect programme implementation and outcomes. Koutsouris and Norwich (2018) have used these analyses to develop a process model of the main aspects relevant to the IGR programme implementation that seem to affect IGR group reading outcomes (see Figure 1).

Figure 1. Process model of contextual factors associated with IGR programme outcomes


Head teacher support
National Curriculum and testing requirements
Pre-existing class organisation fits IGR approach
Availability of IGR-prepared TA
School's current literacy provision
Additional to IGR teaching
IGR programme: principles, materials, practices & training
Fidelity of implementation
Quality of IGR group teaching
Pupil engagement, enthusiasm & inspiration
Improved reading and attitudes
Teacher, knowledge, skill, self-efficacy to teach reading
Literacy adviser support & advice
Teachers' understanding of programme principles
IGR programme fits teachers' style

Interpreting the process evaluation in terms of this model indicates there is scope for further programme development, especially in relation to professional learning (about the programme principles) and training (for some teachers to be more comfortable implementing unusual current practices, such as collaborative reading). With reference to the national curriculum requirements, IGR could include synthetic phonics games so that teachers would not feel they have to teach phonics outside of the programme.

The finding of a null effect size for IGR trial could be interpreted, if RCT is the first and only credible evaluation approach, as meaning that there is not strong evidence for its use. Future refinement and evaluation might not be justified given modest-useful gains for typical phonics and IGR approaches; the effort better spent on other approaches. But if RCT is the evaluation approach of first choice but needs to be supplemented by process evaluation, then the positive process evaluation might save IGR for further development and evaluation trials. However, securing funding from funders who prioritise RCTs over process evaluations might be challenging.

This is where the idea of IGR as a complex teaching intervention is useful as it calls into question whether RCT is the evaluation of first choice or just one of several intervention oriented evaluation designs, as discussed above. Perhaps in the case of this IGR trial it would have been wiser to continue with the Local Route model and continue to use D&R approaches as Jan Stebbing had done informally when initially working alone. The scaling up of IGR (1) could have been done using a specific version of a Designed Based Research approaches (Brown, 1992). This would have involved designing the intervention, in this case the complex IGR intervention, to be used by a range of teachers in a range of contexts. Systematic evaluation data could have been collected about the processes (informed by process model, such as in Figure 1) and outcomes, children's reading and attitude changes. There would be no control group, though baseline reading measures could be taken as an internal comparison, and after one cycle of IGR the data analysed to understand whether and how IGR works. This analysis would then inform another cycle of IGR using the same evaluation data approach.

If as suggested above, the Local Route / D&R approach and the General Route / R&D approach are on a continuum, then D&R findings could be generalised to some extent to produce contextualised generalisations. These generalisations might also be examined through RCT type designs. As explained above, the reason in the end why a design experiment approach was not used for IGR was because of the current dominance of the RCT approach and risk of not persuading an organisation to fund another type of evaluation. This is also where realist thinking could be helpful in teaching evaluations, to go beyond the 'catchy' language of 'what works' without grounding this in context and processes. 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). There is no reason why policy makers could not come to favour both D&R and R&D approaches to improving teaching and learning once researchers come to see RCT as one and not always the first choice amongst experimental approaches.

References

Al Otaiba, S., C.M. Connor, J.S.Folsom, J. Wanzek, L. Greulich, C. Schatschneider 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.

Bentley, T. and S. Gillinson 2007. *A D&R system for education*. London: Innovation Unit.

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

Brown, A. 1992. “Design experiments; theoretical and methodological challenges in creating complex interventions in classroom settings”. *Journal of learning Sciences*, 2, 2, 141-178.

Charlton, B., R.L. Williams, and T.F. McLaughlin 2005. “Educational Games: A Technique to Accelerate the Acquisition of Reading Skills of Children with Learning Disabilities”. *International Journal of Special Education*, *20*(2), 66–72.

Connolly, P., Keenan, C. and Urbanska, K. 2018. The trials of evidenced-based practice in education: a systematic review of RCTs in educational research 1980-2026. *Educational Research*, 60, 3, 276-291.

Cronbach, L.J. 1975. “Beyond the two disciplines of scientific psychology”. *American Psychologist*, 30(2) 116-127.

Department for Education 2013. *Phonics screening check and national curriculum assessments at key stage 1 in England*. https://www.gov.uk/government/publications/phonics-screening-check-and-national-curriculum-assessments-at-key-stage-1-in-england-2013

Education Endowment Foundation (EEF) 2015. *Making Best Use of Teaching Assistants: Guidance Report*. London: EEF.

Education Endowment Foundation (EEF) 2016 *EEF at 5*. https://educationendowmentfoundation.org.uk

Education Endowment Foundation (EEF) 2018 *Information for grantees about EEF’s approach to evaluation*. https://educationendowmentfoundation.org.uk/projects-and-evaluation/evaluating-projects/evaluation-information-for-grantees/

Fien, H., J.L.M. K. Smith, S.K. Smolkowski, S.K. Baker, N.J. Nelson and E. Chaparro 2014. “An examination of the efficacy of a multi-tiered intervention on early reading outcomes for first grade students at risk for reading difficulties”. *Journal of Learning Disabilities*, 46, 6, 602-621..

Flanagan, O.J. 1981. “Psychology, progress, and the problem of reflexivity: a study in the epistemological foundations of psychology”. *Journal of the History of the Behavioral Sciences*, 17,375-386.

Furedi, F. 2013. *Teaching is not some kind of clinical cure comment.* http://www.frankfuredi.com/newsite/article/teaching_is_not_some_kind_of_clinical_cure

Goldacre, B. 2013. *Building evidence into education*. https://media.education.gov.uk/assets/files/pdf/b/ben%20goldacre%20paper.pdf

Hammersley, M. 2008 “Paradigm war revived? On the resistance to RCTs and systematic reviews”. *International Journal of Research and Method in Education*, 31(1) pp. 3–10.

Hammersley, M. 2013. *The myth of research-based policy and practice*. London: Sage.

Hargreaves, D. 1996. *Teaching as a research-based profession: possibilities and prospects*. TTA Annual Lecture. London: TTA

Humphrey, N. 2018. “Are the kids alright? Examining the intersection between education and mental health”. *The Psychology of Education Review*, 42,1, 4-12

Humphrey, N., A. Lendrum, E. Ashworth, K. Frearson, R. Buck, and K. Kerr (n.d.) “Implementation and process evaluation (IPE) for interventions in education settings: An introductory handbook”. London: EEF

Hutchinson, D. and B. Styles 2010. *A guide to running randomised controlled trails for educational researchers*. Sough: NFER.

Katsipataki, M> and Higgins, S. 2016. What woks or what’s worked? Evidence from education in the U.K. *Procedia – Social and behavioural Sciences*, 217, 903-909.

Koutsouris, G., B. Norwich, 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* doi:10.1080/0305764X.2018.1438365

Koutsouris, G. and B. Norwich 2018. “What exactly do RCT findings tell us in education research?”. British Educational Research Journal https://doi.org/10.1002/berj.3464

Lewis, C., R. Perry and A. Murata 2006. “How should research contribute to instructional improvement? The case of lesson study”. *Educational Researcher* , 35 (3), pp. 3–14.

MacIntyre, A. 1985. *After virtue: a study in moral theory*. London: Duckworth. .

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

Muter, V., C. Hulme, M.J. Snowling and J. Stevenson 2004. “Phonemes, rimes and language skills as foundations of early reading development: Evidence from a longitudinal study”. *Developmental Psychology*, 40, 663-681.

Norwich, B. 2014. “Context, interests and methodologies, in Research in special needs and inclusive education: the interface with policy and practice”. *Journal of Research in Special Educational Needs*, 14,3,193-196

Norwich, B., G. Koutsouris 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 Programme*. Project Report for Nuffield Foundation. Access at http://www.integratedgroupreading.co.uk/evaluation-project/

Norwich, B. and G. Koutsouris 2019. "An inclusive model of targeting literacy teaching for 7-8 year old children who are struggling to learn to read: the integrated group reading (IGR) approach". In Boyle, C. S. Mavropoulou, J. Anderson and A. Paige (eds.) I*nclusive education: global issues and controversies*

Pawson, R. and N. Tilley 1997. *Realistic Evaluation*. London: Sage.

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

Primary National Strategy 2007. *Letters and Sounds: Principles and Practice of High Quality Phonics*, Department for Education and Skills, UK. https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/190599 /Letters_and_Sounds_-_DFES-00281-2007.pdf

Rose, J. 2009. *Identifying and Teaching Children and Young People with Dyslexia and Literacy Difficulties*. DCSF-00659-2009.London; DCSF.

Sharples, J., B. Albers and S. Fraser 2018. *Putting evidence to work: a school's guide to implementation: Guidance Report*. London EEF. 9.8.18 at: https://educationendowmentfoundation.org.uk/public/files/Publications/Campaigns/Implementation/EEF-Implementation-Guidance-Report.pdf

Siddiqui, N., Gorard, S. and See, B.H. 2016. The importance of process evaluation for randomised control trials in education. *Educational Research*, 60,3, 357-370.

Stebbing, J. 2013. *Putting it all together: Integrated Group Reading for raising the attainment of children needing systematic, integrated reading experience in Years 2 and 3*. (Unpublished paper: contact first author for copy).

Stentiford, L., G. Koutsouris and B. Norwich 2018. "A systematic literature review of the organisational arrangements of primary school-based reading interventions for struggling readers". Journal of Research in Reading (revised version under review)

Thomas, G. and R. Pring 2004. *Evidence-based practice in education* (eds.) Maidenhead: Open University Press.

Thomas, G. 2016. "After the gold rush: questioning the 'gold standard' and reappraising the status of experiment and RCT in education". *Harvard Review of Education*, 86, 3, 390-411.

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

Torgeson, C. J. 2009. “Randomised controlled trials in education research: a case study of an individually randomised pragmatic trial”. *Education 3-13*, 37,4, 313-321.

Torrance, H. 2013. “Building evidence in education: why not look at the evidence?” *Research Intelligence*, 121, 26–8.

Torgeson,C., G. Brooks, L. Gascoine and S. Higgins 2018. “Phonics: reading policy and the evidence of effectiveness froma systematic tertiary review”. *Research Papers in Education*, DOI: 10.10.1080/0267/1522.2017.1420816.

Toulmin, S.E. 1972. *Human Understanding, Volume I: The Collective Use and Evolution of Concepts*. Princeton: Princeton University Press.

What Works Clearinghouse 2013. *Updated intervention report: Reading Recovery*. U.S. Department of Education, Institute of Education Sciences.

Wyse, D. and C. Torgeson 2017. “Experimental trials and ‘what works?’ in education: The case of grammar for writing”. *British Educational Research Journal* 43, 6, 1019–1047.