What exactly do RCT findings tell us in education research?

Koutsouris, G. & Norwich, B.

Graduate School of Education, University of Exeter

Abstract

The paper addresses issues related to whether null RCT findings can by themselves be a secure indicator of programme failure. This is done by drawing on the findings of the evaluation of the Integrated Group Reading (IGR) programme using a number of teacher case studies. The case studies illustrate how the same intervention can be implemented differently in local circumstances, with different outcomes. The different ways in which IGR was implemented reflect how teachers experienced the pressures of the national curriculum, their attitudes to the IGR approach to reading, the school ethos and the resources and support available – and point to how IGR use might be enhanced to result in more significant reading gains. The paper argues that in addition to the statistical findings evaluators ought to pay attention to the context in which a programme is implemented, especially when it comes to complex interventions trialled in real classrooms. It is also concluded that it is preferable to avoid asking whether a programme works or not for all and under any circumstances. A focus on the different ways that programmes work under different circumstances and when implemented by different people is a more useful perspective. This might not provide the certainty that policy makers would likely opt for, but it captures more the complexity associated with teaching programme evaluation.

Keywords: Case study, reading intervention, randomised control trial, complex intervention

The importance of understanding context in programme evaluation

Educational interventions often include aspects of process evaluation (e.g. Humphrey et al., 2016), yet in some cases there is a silence about the importance of context. For instance, Clarke et al. (2010) report findings from a reading intervention where all treatment groups made statistically significant improvements compared to the control condition. The authors concluded that their findings have 'important educational implications for large numbers of children' (p. 1113) with similar difficulties. However, the paper focuses exclusively on discussing the statistical findings and makes no reference to the local contextual factors where the interventions were trialled, with only a passing reference to intervention fidelity. This appears to be common practice for some educational research where significant results are reported (e.g. Duff et al. 2012), and an exploration of local factors is more likely to be deemed necessary in the case of null findings (e.g. Vaughn et al., 2016). In addition, in the Government Department for Education endorsed argument for building evidence into education through randomised trials (Goldacre, 2013), where there are 14 references to 'what works' in a 15-page paper, there is no reference to context or process evaluation.

Pawson & Tilley (2004) point out how the questions asked in programme trials have gradually moved from 'what works' (often followed by a silence about the importance of context) to different ways that programmes work under different circumstances and when implemented by different people. This represents a shift from the long-assumed evaluation approach that focuses exclusively on the observed outcome which is attributed to a specific input, but overlooks the complex processes associated with context. As Stame (2004) argues, when programme designers introduce an input, they 'often gloss over what is expected to happen, the how and why' (p. 58); from this approach to

evaluation, the expectation held is an answer in the form of 'go/no go' that can easily be translated into policy decisions.

Most educational interventions can be seen as 'complex interventions' (a concept originating in health research) that Moore et al. (2015) define as interventions with:

> '…multiple interacting components, although additional dimensions of complexity include the difficulty of their implementation and the number of organisational levels they target' (p. 1).

When it comes to complex (or any real-world) interventions, effect sizes are not enough to shed light onto the contextual factors that might serve as barriers/ facilitators of implementation (for instance the presence or absence of supportive factors, the enthusiasm or indifference of the people involved etc.). Yet, without sufficient exploration of how and why (in addition to whether) a programme works or does not work in local circumstances, it is doubtful that an intervention can be successfully replicated in a different context (for good outcomes) or effectively revised (for null results).

A pertinent issue is the distinction between efficacy and effectiveness. Efficacy is about whether an intervention can produce good results under ideal conditions (when researchers are better able to reduce/ account for typically encountered threats such as selection bias and poor implementation), while effectiveness is about real-world applicability (when the conditions of implementation are relevant to the real-world implementation). This distinction links also with that between internal validity of a trial (unbiased comparison, fidelity etc.) and its external validity (i.e. generalisability to the real world). Some authors see a trade-off between these two types of validity (Streiner, 2002). This suggests that the efficacy of complex interventions cannot be simply evaluated, since striving for internal validity would have to be balanced with real-world applicability. This is particularly relevant to educational interventions that have to involve practical arrangements that are relevant to schools and also ethical, although such arrangements could be seen as a threat to internal validity. Vaughn et al. (2016), for example, discuss a similar dilemma:

> '…after students were identified with significant reading comprehension problems and were randomized to treatment and comparison conditions, the schools decided to provide their own interventions to students in the comparison condition. Since students were so far behind, it was unethical to ask them not to provide the intervention particularly since the study was scheduled throughout the 4th grade year' (p. 16).

Therefore, in the case of effectiveness trials, where internal validity claims have to be balanced with the complexities of the real word, exploring context is crucial to understanding why a programme works or not, and what is actually being compared in a programme trial.

This discussion is questioning the current belief that randomised control trials (RCTs) provide a secure gold standard 'for identifying what works'. Cartwright (2007) for instance writes that RCTs, even when they have good internal validity, can put severe constraints on the assumptions that a target population has to meet to justify the applying of RCTs generalisations. She goes on to argue that other methods might provide more reliable information, but this would have to be judged on a case-by-case basis. This argument is also echoed by Hammersley (2015), Scriven (2008), Shaffer (2011) and others. Although Hammersley (2015), for example, notes that RCTs do have many advantages, he argues that by themselves they can give little information about causal mechanisms and the conditions under which these operate. This becomes particularly problematic in the social world, where information about causal mechanisms is less reliable or predictable. Hammersley also

discusses measurement difficulties in social experiments, and a standardisation of treatment issue – for instance, almost inevitably a teacher will adapt their teaching or behaviour in response to their students, since all social science treatments involve some kind of social interaction (Dunn et al., 2007 e.g. examine this in relation to psychotherapy treatments). Concerns about the relevance of RCTs in evaluating social 'real life' interventions have also been raised by other authors (indicatively Goodman et al., 2018; Thomas, 2016), and it is a matter broadly acknowledged by researchers involved in large-scale RCTs (see for instance Humphrey's et al., 2016 guidance on process evaluation for EEF trials).

The issue that this paper raises is whether null RCT findings can by themselves be a definitive indicator of programme failure. Research has taken different approaches in interpreting null findings, for instance by acknowledging *compensation effects* (control schools compensating for their treatment status assignment by increasing the amount of additional teaching) an in Patel et al. (2017); or implementation failure (Harachi et al., 1999). However, in any case null RCT findings call for a deeper exploration of the mediational factors affecting how an intervention works (e.g. teachers' self-efficacy and motivation), as well as of the context in which an intervention was evaluated. With reference to the latter which is the particular focus of the paper, such an exploration is expected to shed light onto the contextual factors that serve as facilitators or barriers to successful programme implementation, in order to better understand why a programme works for some but not for all, and identify areas for revision and future development. These issues are explored drawing on the evaluation findings of the Integrated Group Reading (IGR) programme, and particularly on teacher cases studies that illustrate some of the different ways the programme was used in real classrooms.

The IGR intervention

The IGR programme is a targeted tier 2 intervention designed to be delivered four times a week for 30 minutes as part of a whole class group reading session. It adopts a multi-perspective approach to the teaching of reading, integrating (analytic) phonics, story-telling for oral language development, word games, and elements from Paired Reading and Reading Recovery. For details about IGR visit: http://www.integratedgroupreading.co.uk/. The teacher teaches an IGR group of four pupils identified as in need of tier 2 support twice a week, and the teaching assistant (TA) works with the group in-between the teacher sessions for consolidation (four sessions in total). Teacher and TAs have discrete yet interconnected roles, with the teacher keeping the main role. During the teacher-led IGR sessions, the rest of the class is working in various groups independently or with a TA on various reading-related activities. Organising for IGR can be challenging for classes with many reading groups; teachers are encouraged to seek a variety of solutions to ensure all pupils have access to the teacher's time (e.g. involving other staff).

The findings of the IGR trial

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 for two years (2015-2017). The project involved a clustered RCT with a process evaluation and found that participating children in schools using IGR made the same degree of progress in reading accuracy and comprehension, compared to similarly struggling children in control schools: a ratio gain [note 1] of 1.6, seen as 'modest impact' (Brooks, 2016) across phases 1 and 2 (or 11.5 months of progress in 7 months of implementation). However, there was great variation in the way teachers implemented the IGR programme and the gains they recorded (Norwich et al., 2018). This finding (and the recognition that the IGR programme

was a complex intervention implemented in real classrooms) drew a lot of attention to the context in which IGR was implemented. The crucial question was whether the null RCT findings meant that IGR was not an effective programme; we explored this issue using a number of teacher case studies.

Aims and methods

As part of the IGR process evaluation [note 2], we also conducted a number of teacher case studies. We expected that high fidelity of IGR teaching would be associated with greater reading gains for the IGR groups – yet, this was not confirmed in all cases. The selection of cases was based on a combination of fidelity and mean IGR group reading gain scores. Using the fidelity and reading scores, different combinations were selected to represent teachers, presented in table 1. These are the questions we tried to address:

For cases where there was a match between IGR fidelity level and reading gains:

a. For cases with high fidelity and high reading scores

   Is IGR teaching and/or some other factors related to higher than mean reading gains?

b. For cases with low fidelity and low reading scores

   Were low reading gains because IGR was used poorly or some other factors that were related to lower reading gains?

For cases where there was a mismatch between IGR fidelity level and reading gains:

c. For cases with high or medium fidelity and low reading scores

   Why are the pupils not getting higher reading gains, even though the teacher was using IGR at an average/high level?

d. For cases with low or medium fidelity and high reading scores

   What other factors despite low or medium fidelity teaching were related to higher reading gains?

Insert table 1 here

The instruments used in the selection and analysis are presented below:

Single Word Reading Test (SWRT) (Foster, 2007): SWRT was used to select cases for the case studies as it was administered by trained visiting researchers to the four IGR pupils. SWRT is an assessment of reading accuracy.

Hodder Group Reading Test (HGRT) (Hodder Education, 2000): delivered by the class teacher using detailed instructions, and then returned by post to the research team for scoring. HGRT is an assessment of both reading accuracy and comprehension.

Fidelity

All teachers received IGR-related programme training and support and were expected to be able to use the approach with good fidelity. A fidelity index was devised using a 3-point scoring system to evaluate the quality of IGR teaching [note 3]. Phase 2 observation data were more detailed, so the case studies reported are all from phase 2. The same teachers were independently scored by the programme and evaluation teams using the same observation notes from programme support visits; the scores correlated highly (0.8).

Self-efficacy questionnaire

Teaching self-efficacy was 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 & Rankin-Erickson (2013) and Tschannen-Moran & Johnson (2011) (Cronbach's alpha for phase 1 training day was 0.91).

CMO questionnaire

Following a realist evaluation framework (Pawson & Tilley, 2004), an IGR *programme theory* was constructed and in turn used to devise a context, mechanism and outcomes (CMO) questionnaire [note 4]. Only outcome (10-item bipolar scale) scores are reported in this analysis.

Teacher interviews and observations

All teachers discussed in the case studies were individually interviewed about their experiences of using the programme, and 4 out of 7 were also individually observed by the evaluation team.

Table 2 sets out the broad factors and data sources used in the teacher case studies analysis.

Insert table 2 here

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 also sought from parents. Anonymity and confidentiality has been applied to every aspect of the project, and schools/ individual participants had the right to withdraw at any time.

Case studies

The case studies are presented in terms of the structure presented in table 2. Teacher and IGR group characteristics are presented in tables 3 and 4.

Insert table 3 here

Insert table 4 here

Teacher 1 (T1): High fidelity; High gains

School characteristics

The teacher was serving in a suburban school in the West Midlands with 31.4% of pupils in receipt of Free School Meals (FSM). The school had strong literacy provision, and keen leaders who supported the programme – e.g. the Assistant Head (responsible for literacy) offered to read with one of the non-IGR reading groups (and did this for both participating classes), so that all pupils had equal reading time with a teacher. This happened every week, all year round, and the teacher noted in his interview that: '*if I wasn't in this school and didn't have that available, I would have found it really difficult to fit in every child reading to me each week*'.

The local literacy adviser was very supportive of the programme and visited the school several times to make sure that the teachers were using the programme with good fidelity.

IGR organisation

During group reading, there were 5 groups plus the IGR group. The teacher was concerned about the noise (IGR pupils were particularly enthusiastic in playing the games), so on a rotating basis one of the groups was working just outside of the class with a TA – this was usual practice in the school. This was possible because two TAs were available: one was working with the group out of the class, and the other was overlooking the class during the teacher-led IGR sessions.

IGR teaching

The teacher was observed by the programme team to have good fidelity to the programme (2.8/3); there was no independent observation of his IGR teaching. He appreciated the multi-perspective approach of IGR but reported that it took him some time to get used to the programme delivery.

Mean group reading scores, and reported outcomes:

For the 4 IGR pupils, the ratio gain was 2.4, often seen as 'useful impact' (table 5).

Insert table 5 here

The teacher also reported in his interview positive pupil outcomes in reading expression, accuracy, comprehension and confidence. He particularly noted how the safety of the IGR group helped the self-esteem of one particular pupil with language difficulties who was not worried about reading in the small group setting. These findings were further supported by the teacher's CMO questionnaire responses in which all programme outcomes were rated highly positively (table 3).

Teacher 2 (T2): High fidelity; High gains

School characteristics

The teacher was serving in a Greater London school with particularly high FSM percentage (43.6%). The school had strong literacy provision and dedicated school leaders. E.g. the Deputy Head was actively involved in the programme procedures, consistently supported the participating teachers, and was interested in exploring ways of using the programme after the completion of the project.

IGR organisation

There were four other reading groups in addition to the IGR, doing comprehension or spelling tasks, reading one-to-one with a TA or reading for pleasure. The teacher was particularly concerned about the time devoted to IGR teaching (especially closer to the SATs), and was worried that other pupils were missing opportunities to read individually with herself or a TA. She had explored other options, e.g. reading with the IGR group in assembly time for one of the sessions, but found such options difficult to implement consistently. She also felt initially that 30 minutes was too long for the rest of her Year 2 class to work meaningfully, and that class pupils were getting distracted and interrupting the teacher. However, these issues improved during the year.

IGR teaching

The teacher was reported by the programme team to have good fidelity to the programme (2.8/3), and there was no independent observation of her IGR teaching. She reported in her interview that she felt that IGR fostered the enjoyment of reading, and used creative ways (e.g. games) to support pupil learning.

Yet, she was worried that the programme focused more on accuracy rather than comprehension which is particularly important for the SATs. Despite her concerns, the teacher could understand why IGR had to stay simple (being a remedial programme), and tried to practice comprehension and inference outside IGR teaching.

Mean group reading scores, and reported outcomes

For the 4 IGR pupils, the mean ratio gain (table 5) was 2.1, often seen as 'useful impact'.

The teacher reported positive outcomes for the IGR group, especially in relation to their confidence and oral language skills. She also noted that they were working towards meeting the expectations for the end of Key Stage 1, but were still behind their classmates. She rated highly all outcomes in her CMO questionnaire at the end-of-the-year review meeting (table 3).

Teacher 3 (T3): High fidelity; Low gains

School characteristics

The teacher was serving in a school in the South West, with 26.3% FSM percentage. Throughout the programme implementation, the project team had minimal communications with the school leaders.

IGR organisation

The teacher had one TA available for her class, and during teacher-led IGR, used to send one group out of the class (in an area still visible from the class) with the TA. In her interview, she explained that this was common practice in the school, and she found it difficult to have two adult voices in the class. She also noted that she was able to have good control of the remaining groups in the class when she was teaching IGR – the independent observation confirmed this, as her classroom was one of the most settled classrooms with few distractions from non-IGR pupils.

The teacher had wondered whether this organisation would be acceptable for the evaluation and had sought guidance when she had started using the programme. The project team advised her to feel free to explore the different options for organising her classroom. In her interview, the teacher wondered whether IGR could work better in TA-led pull-out sessions, as pupils would be able to work without other distractions; however, she did not use this organisation during the trial.

IGR teaching

The teacher was observed to have generally good programme fidelity by the programme team (2.7/3) with some notable weaknesses especially in the collaborative reading element. The independent observation indicated that the teacher was trying to be faithful to the programme, but used the programme strategies in a mechanical way. Her approach was also reflected in her body language: she was facing directly the two weaker pupils (to support them as far as possible), and was sitting far away from the two stronger members of the group who remained unengaged. This worked against developing the good group dynamics that are central to IGR. This can also be related to her way of using storytelling that was observed to be closer to an inference-based questioning (often associated with group reading), failing to engage pupils in the story of the new book. Yet, these weaknesses were not captured well by the fidelity index used (discussed in detail later on), and thus were not reflected in her (high) fidelity score.

The teacher reported that she felt IGR supported pupils' vocabulary and language but was worried that there was not enough focus on comprehension to prepare the pupils for the SATs. She was then planning to include more opportunities for comprehension practice closer to the summer term.

Mean group reading scores, and reported outcomes:

For the four IGR pupils, IGR had little or no positive impact (table 5).

However, the teacher reported positive outcomes in her interview and CMO questionnaire (table 3), especially in relation to the children's confidence: '*most of them are confident in the small group and can make mistakes*' without having to worry about other pupils' judging their reading.

Teacher 4 (T4): Low fidelity; Low gains

School characteristics

The teacher was serving in the same school as T3.

IGR organisation

Pupils were organised into 5 reading groups (including the IGR group). The teacher reported in her log that 2 of the groups worked with TAs (two TAs were often available) and 2 on independent work. She noted that Year 2 children needed to be trained to work well independently for 30 minutes, but this was not a matter of concern. She was only worried that '*for me to get around the other groups of my class takes me two weeks as opposed to one*'. Despite her concerns about the teacher time spent with the IGR group, she kept to the suggested organisation for the duration of the trial.

IGR teaching

The teacher was observed by the programme team to have poor fidelity to the programme (1.8/3), particularly in relation to the storytelling aspect. During the independent observation, the teacher read part of the book instead of narrating it, and failed to deeply engage the pupils in the new story; pupils seemed to enjoy the session but were not enthusiastic. The programme team reported that she gradually improved and managed to use the approach effectively towards the end of the year.

The teacher noted that she felt that IGR was useful because of '*the structure and the pace. It means that we get through a lot of books and material […] and it's focused me on SATs*', suggesting a misunderstanding of the remedial nature of the programme.

Mean group reading scores, and reported outcomes

For the IGR group, the difference between SWRT and HGRT is considerable (table 5); the high HGRT gains seem to be due to low baseline scores, suggesting possible difficulties in using HGRT in time 1, whereas the SWRT scores show little progress. Note that HGRT was delivered by the class teacher, whereas SWRT by a trained visiting researcher.

She reported in her interview that pupils in the IGR group were progressing well, and rated highly most programme outcomes in the CMO questionnaire (table 3).

Teacher 5 (T5): Low fidelity; Medium high gains

School characteristics

The teacher was serving in a South-West school with a very low 8.8% FSM percentage. The school run a School Direct Teaching Programme supported by a local University, which also meant that trainee teachers were assisting in the participating classrooms. The school offered additional phonics and TA-led pull-out spelling sessions every morning and afternoon to weak readers (10min sessions), including some of the IGR pupils. Throughout the year, Friday was dedicated to comprehension and all pupils had activities tailored to the reading they did during the week.

IGR organisation

The teacher had 6 reading groups in her classroom (including the IGR group), but since she had a TA and a trainee teacher available to work with the rest of the class, she did not experience any issues. This is from her log entries: '*IGR now well established in classroom, no concerns; my only concern is the length of time non-IGR groups have to work independently*'. As she was teaching a Year 2 group, 30 minutes were often too long for class pupils to work independently.

She particularly appreciated the IGR organisational model, as she felt that it gave structure to the group reading organisation to the benefit of all pupils: '*the longer we're doing it, the better it is, and the more it's benefiting the whole class rather than just the IGR four*'.

IGR teaching

The teacher was observed by the programme team to have poor fidelity to the programme (1.8/3), and there was no independent observation of her IGR teaching. What the programme team noted though was that the low fidelity scoring was because aspects of the routine were missing due to adjustments required to slow down the pace of the lesson to be more relevant to the pupils' needs. However, the elements that were present were taught with high fidelity and attention to detail. The programme team noted that some of the changes were appropriate for the group, but this was not captured by the fidelity index.

The pace of the lesson was an issue for the teacher who noted in her log that: '*my concern is, it does seem to be a little rushed to cover 2 books a week*'. She tried then to use only one book per week, but it took her some time to feel comfortable with the programme routine for one book.

The teacher appreciated the IGR approach that allowed her to focus more on the pupils' reading: '*[In previous years] I used mainly my own material and sentence work. The children weren't always reading as such, they didn't have the book there all the time'*.

Mean group reading scores, and reported outcomes

The IGR group progressed at different rates, and especially one child progressed ahead of the others affecting the mean (table 5). This group also scored higher compared to other groups in the study.

The teacher also reported positive outcomes in her interview and CMO questionnaire (table 3).

Teacher 6 (T6): Medium fidelity; Low gains

School characteristics

The teacher was serving in a West Midlands school with 14.2% percentage of Free School Meals. There were no direct communications between the school leaders and the project team.

IGR organisation

The IGR group read one book (instead of two) per week, but not in the suggested way. As a result, the group had only one TA session (as opposed to two) every Friday inside the class during assembly. This was partly to give time to the TA to read with all other reading groups every week, and when the teacher was asked to restore the number of TA sessions, she argued that '*it would mean the rest of the class are not heard; it would have a knock-on effect'*.

The teacher also noted in her log that without an experienced TA it is difficult to implement the programme properly: *'I'm still having to manage the rest of the class. […] My TA is still young and she's not that competent yet to managing another 24 children'.*

IGR teaching

The teacher was observed by the programme team to have moderate fidelity to the programme (2.4/3) – which represents the mean fidelity score across teachers. The independent observation confirmed that not all aspects of the routine were present (resulting in lower fidelity scoring) as the teacher used to split her routine for one book per week without following the suggested procedures. The aspects of the routine that were present were mostly used with good fidelity, with some weaknesses observed in the storytelling and collaborative reading element.

The teacher appreciated the multi-perspective approach of IGR to reading, particularly the games and the phonics element of the programme: '*Last year we tended to just use books, so it's nice to have the different games and the phonics'*. Yet, she noted that future versions of IGR could include made-up words for pupils who have failed their Year 1 phonics screening.

In addition to IGR, all children had interventions for oral language support (e.g. *Black Sheep*), vocabulary (e.g. *Word Aware*), and one of the children had also individual TA-led phonics sessions.

Towards the end of the year, the teacher wrote in her log: *'I am worried that IGR pupils are not having comprehension written activities in preparation for SATs; this could affect their performance'.*

Mean group reading scores, and reported outcomes:

For the IGR group, the ratio gains suggest 'modest' impact of IGR (table 5).

The teacher reported positive outcomes in her interview and CMO questionnaire (table 3). She also described IGR as a nurture group: '*It's such a small little nurture group, it's really developed those lacking in confidence to become more confident in saying something'*.

Teacher 7 (T7): Medium fidelity; Low gains

School characteristics

The teacher served in the same school as T2, and job-shared her class with another teacher (not T2).

IGR organisation

Pupils were organised into 6 reading groups (including the IGR group). The teacher reported in her interview that in order to have an IGR group of 4 pupils, the other groups had to be larger. This was because this school had particularly large classes (35 pupils), compared to other schools in the study. Having larger groups allowed the teacher to read with all pupils every week – yet, she still had to read with one of the (non-IGR) groups in assembly time.

IGR teaching

The teacher was observed by the programme team to have average fidelity (2.4/3) with some weaknesses in the storytelling and collaborative reading aspects. The independent observation noted particularly the calm manner of the teacher, and the enthusiasm of the group. There were also interesting dynamics in the classroom, with the calm nature of the teacher being complemented by a particularly lively teaching assistant who took over the rest of the class during teacher-led IGR.

However, the programme team noted that the teacher-TA dynamics did not work as effectively with the job-sharing teacher, and there were some personality clashes.

The teacher was positive about the IGR approach to reading and noted that it was particularly suitable for EAL pupils: '*the books are [more accessible to the children] than a lot of the books that we've got in school. The games repeating the language […] are really good for the EAL children'*.

Her main concern in relation to the impact of IGR on the class pupils was that there was limited TA time to read with them, as almost all of them were EAL and could benefit from one-to-one support.

Mean group reading scores, and reported outcomes

The programme did not have any impact on the IGR pupils (table 5).

The teacher did not complete a CMO questionnaire. In her interview, she reported that she believed that the IGR pupils became more confident because of the structure of the IGR approach.

Discussion

Seven teachers with diverse profiles were examined as individual cases representing different combinations of fidelity scores and reading gains. The above analysis suggests the following:

T1 and T2: high gains can be attributed to the quality of IGR teaching and other supportive factors

T3: low/no gains can be attributed to specific aspects of IGR teaching – this case also indicates that the fidelity index might not give due weight to the quality of collaborative reading activity

T4: the low gains can be attributed to the low level of IGR teaching and organisational issues

T5: the good gains can partly be attributed to other programmes operating during IGR, and partly to IGR teaching, as the fidelity index could not capture appropriate changes in pace. The IGR pupils had also higher initial reading scores compared to other groups in the study

T6: low gains can partly be attributed to implementation, class organisation and TA-related issues and partly to not using the full sequence and number of IGR sessions

T7: low gains were not only associated with medium fidelity of IGR teaching but also with large class, job-share and inconsistent IGR approach between job-sharing teachers

These cases are now discussed overall to explore the relationship between teaching fidelity and reading gains. There will also be some discussion of the way fidelity was calculated in the study.

Match between level of IGR fidelity and reading gains (T1, T2 and T4)

These case studies represent a relationship which is consistent with the original hypothesis that using the IGR programme with good fidelity corresponds with higher reading gains for the IGR group (and the vice versa).

The main factors associated with high fidelity and high gains (from T1 and T2) were the enthusiasm of teachers and pupils, the involvement of the school leaders and local literacy adviser, the ethos of the school (as reflected on the approach to literacy provision), the way the IGR model fitted with pre-existing reading organisational arrangements, the understanding of the theory and rationale behind the programme and the presence of additional to IGR input for the IGR pupils. The main factors associated with low fidelity and low gains (from T4) were an unclear understanding of the theory and goals of the programme, concerns about its organisational model, difficulties in

implementing aspects of the programme (e.g. storytelling) thus failing to engage and inspire pupils, and lack of evidence about the school leaders' involvement.

The extent to which the IGR organisation model fitted well with existing arrangements in schools seemed to be particularly relevant. Where IGR could be incorporated into the existing arrangements with few changes, then the programme implementation was smoother and there were few practical problems – this is, e.g., the case of T1 (and T5 from the next section). The implications of the IGR arrangements in such cases could be as limited as having to train class pupils to work independently. Yet, where more changes were needed then a few issues arose, especially in relation to the teacher/ TA time that the other reading groups were getting during the year; this is the case of T4 and applies also to most teachers in the next section.

The extent to which the IGR programme fitted with pre-existing arrangements in schools could be related to Streiner's (2002) argument about a trade-off between efficacy and effectiveness trials. Effectiveness trials that can fit well with pre-existing teaching and staffing arrangements can have greater applicability to the real word, although they might have to sacrifice stricter adherence to fidelity to achieve this. The IGR trial is more towards this side of the continuum. On the other hand, efficacy trials designed with stricter requirements in place (e.g. re fidelity, control group teaching) can present more practical and organisational challenges for schools (that might intimidate them), and their findings would be less applicable to the real world.

Mismatch between IGR fidelity and reading gains (T3, T5, T6 and T7)

These teacher case studies present a mismatch between fidelity and reading gains which does not confirm the original hypothesis that using IGR with good fidelity corresponds with higher reading gains for the IGR group. These cases could be organised into two groups: the first includes teachers with high or medium fidelity scores and low reading gains (T3, T6 and T7); and the second involves the single case study of a teacher with low fidelity scores and relatively high gains (T5).

The first group of case studies involves teachers who despite their adherence to the programme procedures, taught groups that did not record good progress on the standardised assessments used. For T3, this was mainly due to her mechanical approach to the programme implementation that became a barrier to using the programme strategies effectively and engaging pupils at a group level. This can be seen as a reminder that interventions have to be understood pedagogically by teachers and not simply delivered. T6 and T7 had both organisational difficulties related to their TAs. T6 had a less experienced TA who was less comfortable to manage the rest of the class during teacher-led IGR, so the teacher had to pay attention to both her class and IGR group. T7, on the other hand, had a confident TA who took over the rest of the class during IGR, but this did not work well with her job-sharing teacher, so resulting in tensions. As a result, T6 and T7 had a less smooth experience with the IGR organisation during the year. In addition to this, they were both concerned about the limited teacher time available for the class pupils and both tried to find alternative solutions (reading in assembly time, organising larger reading groups). These issues suggest that, although both teachers tried to implement the IGR model, this was only possible to a certain extent, as there was a less good fit between IGR and pre-existing arrangements. These issues might explain the low gains of their IGR groups despite the average or good fidelity of their IGR teaching.

On the other hand, T5 represents the case of a teacher with poor fidelity to the programme, but relatively high gains for her group. T5 received a low fidelity score mainly because aspects of the routine were missing as a result of her attempt to slow down the pace of the lesson using one book per week. Her decision was in response to the needs of her group; it is indicative that the aspects of

the routine that were present were taught with good quality and attention to detail. The gains of her group suggest that a programme needs to be flexible to adjust to particular circumstances – a matter more broadly associated with the evaluation of classroom-based teaching interventions as this quote notes:

> 'On one hand, investigators need to ensure consistent treatment implementation [and on the other] teachers need some leeway to adapt interventions within particular classroom contexts' (Conroy et al., 2008, p. 211).

This teacher's case points also to issues relevant to the measurement of teaching fidelity in the study. Fidelity was calculated from an index that was developed after the programme team observations had taken place (and before knowing the statistical results). Since the programme team observation notes were made for support (and not data collection) purposes, the observation data available were very varied in style and quality. In addition, there were some independent observations from the evaluation team but these were conducted with a focus on organisational decisions and challenges, so they could shed little light on the fidelity of teaching.

The fidelity index also had weaknesses, as it failed to distinguish between omissions and strategic changes in pace that were teacher responses to pupils' needs – the most notable example of alterations was the use of only one book per week. This resulted in lower fidelity scores for some teachers, as for example T5 (whose IGR group showed progress). A more sensitive instrument should be able to distinguish between good and less good teacher decisions, rather than expecting adherence to procedures that might not be relevant to the pupils at the time of the observation. This matter is also echoed by Moore et al. (2015) when discussing the issues associated with understanding fidelity in relation to a complex intervention:

> 'Fidelity is not straightforward in relation to complex interventions. […] Strict standardisation may be required and controls put in place to limit variation in implementation. But some interventions are designed to [or will] be adapted to local circumstances' (p. 3).

A related issue was that teachers who used the programme mechanically (such as T3) scored highly on the fidelity index, although they failed to engage in depth with the strategies involved.

Future studies might consider using more sophisticated ways of measuring fidelity, e.g. the approach presented in Vaughn et al. (2016) using a sample of audio-recorded lessons and a rating scale. Programme fidelity as a notion can also be re-examined as it is possible that good teaching decisions (in response to children's needs) may not be in all cases consistent with an RCT protocol, suggesting that personal judgement and strict adherence to an RCT protocol might come into tension.

Implications of the case studies analysis

The analysis suggests 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). There was a variety of local factors (IGR group, teacher, school or broader curriculum) that can be seen to affect programme implementation and outcomes. Despite this complexity, the teacher cases can point to particular combinations and interactions that may be associated with successful or less successful results. Based on the analysis in this paper, figure 1 below presents a process model of the main aspects relevant to the IGR programme implementation that seem to affect IGR group reading outcomes. This process model might be useful for future research and development work on IGR. The case study analyses also show that 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 using all elements of the programme, such as storytelling). 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. In addition, future research can use designs, such as single case experimental designs with multiple cases or quasi-experimental designs which can be more adaptable to school conditions than large scale RCTs. It is also relevant to consider how teachers adapt IGR to local circumstances while still adopting IGR principles. A step towards this direction would be to examine how schools and teachers continued (or not) using IGR after the end of the trial, as the programme materials stayed in schools. This focus on the local context highlights also the important role of teachers in educational research:

> 'Teachers are not going to be given a recipe for 'what works' from research; by its nature, educational research cannot provide certainty of outcome. What it can achieve is to provide reasonable warrant for decisions that must be taken by teachers, in full knowledge of the circumstances in which they work' (Winch et al., 2015, p. 210).

In other words, teachers have the situated understanding to interpret the knowledge produced by evaluation research and use it accordingly. This does not make evaluation research less relevant, but recognises the significance of understanding context in programme evaluation.

Insert figure 1 here

Conclusion

This paper examines whether null RCT findings are a secure indicator of programme failure in relation to the evaluation of the IGR programme. This question was explored drawing on the recent IGR evaluation findings and a number of teacher case studies from the process evaluation that accompanied the RCT. The conclusion drawn is that in addition to the statistical effects sizes evaluators ought to pay attention to the context in which a programme is implemented, especially when it comes to complex interventions evaluated in real classrooms. This has been illustrated by the IGR teacher case studies that showed how the same programme can be implemented differently in local circumstances with different outcomes. The ways in which IGR was used reflect various factors including how teachers experienced the pressures of the curriculum, their attitude to the IGR approach to reading, the school ethos and the resources and support available. These case studies of IGR implementation point therefore to how IGR use can be enhanced to result in more significant reading gains. They also imply that in relation to programme evaluation, it might be too simple to seek to answer whether a programme works or not for all and under any 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 policy makers would likely opt for, but it can 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 future development.

Note 1: *Reading progress divided by the duration of the intervention*

Note 2: *For the IGR process evaluation, see Koutsouris et al. (2018)*

Note 3: *For a copy of the fidelity index, email the corresponding author*

Note 4: *For a copy of the CMO questionnaire, email the corresponding author*

Funding details: 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

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

Cartwright, N. C. (2007) Are RCTs the gold standard?, *BioSocieties*, 2, 11-20.

Conroy, M. A., Stichter, J. P., Daunic, A., & Haydon, T. (2008) Classroom-based research in the field of emotional and behavioral disorders: Methodological issues and future research directions, *The Journal of Special Education*, 41(4), 209-222.

Clarke, P. J., Snowling, M. J., Truelove, E., & Hulme, C. (2010) Ameliorating children's reading-comprehension difficulties: A randomized controlled trial, *Psychological Science,* 21(8), 1106-1116.

Duff, F. J., Hayiou-Thomas, M. E., & Hulme, C. (2012) Evaluating the effectiveness of a phonologically based reading intervention for struggling readers with varying language profiles, *Reading and Writing,* 25(3), 621-640.

Dunn, G., & Bentall, R.P. (2007) Modelling treatment-effect heterogeneity in randomized controlled trials of complex interventions (psychological treatments), *Statistics in Medicine*, 26(26),4719-4745.

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

Goldacre, B. (2013) *Building Evidence Into Education*, (London, Department For Education – DFE.)

Goodman, L. A., Epstein, D. & Sullivan, M. (2018) Beyond the RCT: Integrating Rigor and Relevance to Evaluate the Outcomes of Domestic Violence Programs, *American Journal of Evaluation*, 39(1), 58-70.

Hammersley, M. (2015) *Against 'gold standards' in research: On the problem of assessment criteria*. Paper given at 'Was heißt hier eigentlich "Evidenz"?', Frühjahrstagung 2015 des AK Methoden in der Evaluation Gesellschaft für Evaluation (DeGEval), Fakultät für Sozialwissenschaften, Hochschule für Technik und Wirtschaft des Saarlandes, Saarbrücken, Germany, May. Retrieved from http://www.degeval.de/fileadmin/users/Arbeitskreise/AK_Methoden/Hammersley_Saarbrucken.pdf

Harachi, T. W., Abbott, R. D., Catalano, R. F., Haggerty, K. P., & Fleming, C. B. (1999) Opening the black box: using process evaluation measures to assess implementation and theory building, *American journal of community psychology*, 27(5), 711-731.

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

Humphrey, N., Lendrum, A., Ashworth, E., Frearson, K., Buck, R., & Kerr, K. (2016) *Implementation and Process Evaluation (IPE) for Interventions in Education Settings: An Introductory Handbook* (London, Education Endowment Foundation).

Koutsouris, G., Norwich, B., & Stebbing, J. (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.

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

Norwich, B., Koutsouris, G. & Bessudnov, A. (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*. Available online at: http://www.integratedgroupreading.co.uk/evaluation-project/ (accessed 1 June 2018).

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

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

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

Scriven, M. (2008) A summative evaluation of RCT methodology and an alternative approach to causal research, *Journal of Multidisciplinary Evaluation*, 5(9), 11-24.

Shaffer, P. (2011) Against excessive rhetoric in impact assessment: Overstating the case for randomised controlled experiments, *The Journal of Development Studies*, 47(11), 1619-1635.

Stame, N. (2004) Theory-based evaluation and types of complexity, *Evaluation*, 10(1), 58-76.

Streiner, D. L. (2002) The 2 E's of research: efficacy and effectiveness trail, *Canadian Journal of Psychiatry*, 47(5), 553-556.

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.

Tschannen-Moran, M., & Johnson, D. (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., & Fletcher, J. M. (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.

Winch, C., Oancea, A., & Orchard, J. (2015) The contribution of educational research to teachers' professional learning: Philosophical understandings, *Oxford Review of Education*, 41(2), 202-216.

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

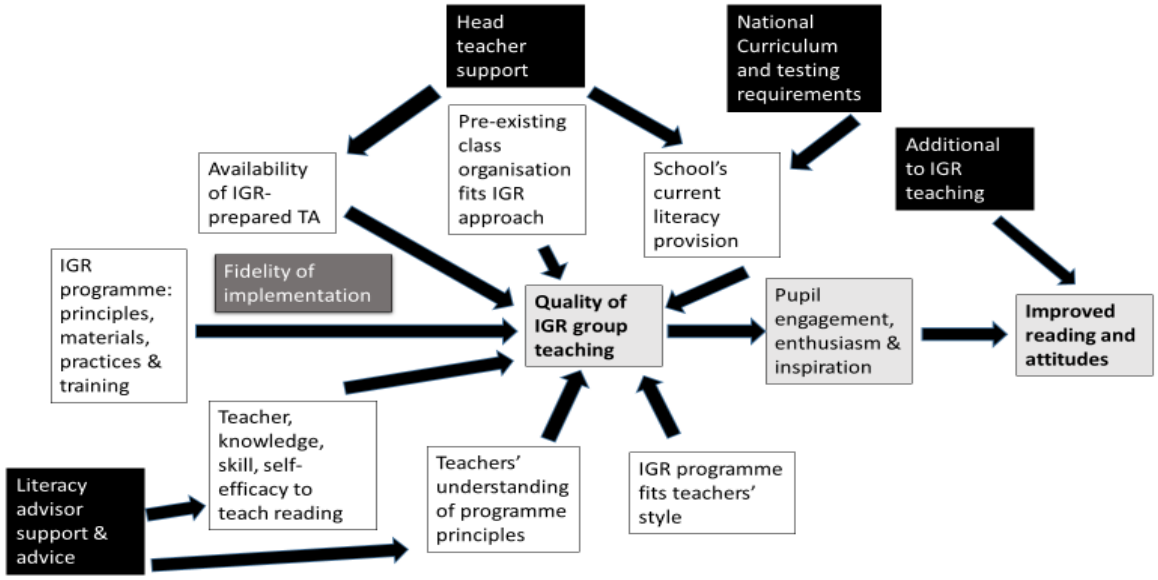

Table 1. Teacher cases based on combinations of fidelity and mean group reading gains (SWRT)

	Fidelity scores – mean 2.4			**SWRT scores – mean difference (t2-t1) 3.7**	
Teacher name	High – **above mean**	Middle – **around mean**	Low – **below mean**	High – **above mean**	Low – **below mean**
T1	✓ 2.8			✓ 10.5	
T2	✓ 2.8			✓ 12.5	
T3	✓ 2.7				✓ 0.2
T4			✓ 1.8		✓ 1.2
T5			✓ 1.8	✓ 4.8	
T6		✓ 2.4			✓ 3.5
T7		✓ 2.4			✓ -1.5

Table 2. Framework used in teacher case studies

Factors	Data sources
1. Teacher characteristics Education, years of experience, age, job-sharing, self-efficacy scores etc.	Demographic and self-efficacy questionnaires, information from programme team and independent visits
2. IGR pupil characteristics Year group, Pupil Premium, gender, EAL, SEN, reported issues	Demographic questionnaire, teacher interviews, the log
3. School characteristics Area, FSM, literacy provision, school leaders and support, other information etc.	Publicly available demographic data, information from programme team and independent visits
4. IGR organisation IGR organisation, attitude towards IGR organisation	Independent observations, teacher interviews, the log
5. IGR teaching **a.** Brief summary of IGR observations	Programme team and independent observations
b. Attitude towards IGR teaching, reported issues, pupils' engagement response to IGR, using IGR with other pupils	Programme team and independent observations, teacher interviews, pupil interviews, the log
6. Mean IGR group reading scores, and teacher reported outcomes	SWRT and HGRT mean gains, reported outcomes from interviews and CMO

Table 3. Teacher characteristics – including self-efficacy and CMO outcomes scores

Teacher	**Gender**	**Age**	**Qualification**	**Experience (years)**	**Self-efficacy t1 (mean across teachers 7.1/9)**	**Self-efficacy t2 (mean across teachers 7.8/9)**	**CMO outcomes mean (-2 to 2)**
T1	M	31	PGCE	5	8.5/9	8.3/9	2
T2	F	-	PGCE	-	5.3/9	7.6/9	2
T3	F	39	PGCE	5	8.0/9	8.4/9	1.6
T4	F	47	PGCE	5	6.7/9	8.0/9	1.5
T5	F	53	B.Ed.	10	6.6/9	7.9/9	1.8
T6	F	56	B.Ed.	34	8.9/9	8.9/9	1.5
T7	F	38	PGCE	12	7.6/9	-	-

Table 4. IGR group characteristics

Teacher	**Year Group**	**Boys**	**Girls**	**English as an additional language (EAL)**	**Pupil Premium**	**SEN school support**	**Education Health and Care plan (EHCP)**
T1	Y3	4	0		0	1	Speech Language and Communication Needs (SLCN)
T2	Y2	3	1	3	1	1	0
T3	Y3	2	2	0	4	0	0
T4	Y2	2	2	0	4	1	0
T5	Y2	1	3	0	2	3	0
T6	Y2	2	2	0	0	1	0
T7	Y3	2	2	3	0	4	0

Table 5. Mean IGR group ratio gains

Teacher	HGRT ratio gain	SWRT ratio gain	
T1	**2.4**	**2.3**	**Ratio gain:**
T2	**2**	**2.2**	
T3	**-0.6**	**0.9**	Reading progress in months divided by the duration of the intervention (7 months)
T4	**3.2**	**1.2**	
T5	**2.9**	**1.5**	
T6	**1.8**	**1.3**	
T7	**0.3**	**0.8**	