# Title: The importance of process evaluation for randomised control trials in education

Nadia Siddiqui*, Stephen Gorard and Beng Huat See
School of Education, Durham University

Corresponding author:
nadia.siddiqui@durham.ac.uk

## Abstract

**Background:** Educational interventions are often complex, and their outcomes could be due to factors not focused on in the impact evaluation. Therefore, educational evaluations using a randomised control trial (RCT) design approach need to go beyond obtaining the impact results alone.

**Purpose:** Process evaluation is embedded in the evaluation design in order to enhance contextual understanding of the outcome results achieved from an RCT. However, in the context of evaluation studies, reporting on the fidelity to the research design protocol is also important and can be undertaken by the same process evaluation approaches as used for studying the mechanism of interventions.

**Research design and method:** This paper reports on two RCTs in which school staff led the trials themselves in their schools – in an aggregated trial study managed and conducted at secondary school cluster-level, with expert advice from us, as independent evaluation advisors. The interventions implemented for evaluation were two highly structured programmes that targeted improvement in the literacy attainment of pupils who were at risk of failing to achieve the expected levels. Assessing the effect on literacy performance was the primary objective. However, the research design included methods for understanding the process of the interventions being implemented. Our main findings on the feasibility of aggregated trials led by schools are informed by our process evaluation, which included participant observations of the training, session observations of the interventions being implemented, and interviews with pupils, teachers and school leaders. Data on fidelity to the interventions were collected, and, using similar techniques, we also evaluated the feasibility of the research process led by school leaders and the possible barriers and challenges of RCT management by schools. We included information on randomisation procedures, perceptions of pupils and teachers involved in the study, and the programme website resources.

**Results:** The primary outcome results of the trials showed promise from the interventions, in raising disadvantaged pupils' reading scores during transition from primary to secondary school stage. The process evaluation revealed considerable potential benefits from involving school leaders as evaluators, and the paper describes a way forward here. However, there were some indications that there was not full compliance with the randomisation process and this might have resulted in initial imbalance of pre-tests scores between the treatment and control groups in one trial.

**Conclusion:** Process evaluation cannot answer research questions regarding impact outcomes, but for a general understanding of outcomes it is important to report how the impact results have been achieved. This is where rich, in-depth data of the kind we describe can assist.


Key words: *Process evaluation, aggregated trials, school leaders, fidelity, in-depth data, randomised control trials*

## Introduction to process evaluation

Educational interventions are complex, as there are often several active elements that create effects on the primary outcomes. An analogy used to describe impact evaluations alone is that of a 'black box'. In other words, the results are wrapped up in a box that needs to be opened in order to develop a systematic and deeper understanding of the achieved effects (Harachi et al. 1999). Impact assessment can answer research questions such as the level of association between variables and the magnitude of 'effect' sizes. The context in which impact results are achieved can reveal wider knowledge about the achieved outcomes (Oakley et al. 2004, 2006). This contextual knowledge helps to understand the mechanism of variables involved in producing the outcomes (Moore et al. 2015), and so aids explanation and further research.

An example from education could be an impact assessment of an educational intervention on students' academic outcomes. We can assess through a proper design-based approach whether students have received (or not) any impact from the implemented intervention compared with an unbiased control group. We can also estimate the magnitude of impact, and whether the effect is positive or negative on the desired educational outcomes (Nelson et al 2012). But these outcomes then need contextual information for interpretation and for making further judgments and decisions about wider implementation. This also means considering a wider set of information to be included as a matter of course in RCT designs, so that the explanation can be built on what works (or fails to work), how it works for further implementation and how respondents involved (or controlled) in the process react or perceive the effects of intervention (Gorard et al. 2017). An impact evaluation might address a question like "Does this intervention work as intended?" or "How much better does the treatment group do as a result of the intervention?" Studying the process, on the other hand, would traditionally address questions such as:

- How well is the intervention implemented in the treatment group?
- How practical are the training and resources for the intervention?
- Are there any barriers to implementation?
- Was the intervention subverted in any way?
- Were there any unintended consequences?
- Is the intervention fully developed for implementation at a larger scale?
- Did the participants appreciate the intervention?
- Did this appreciation vary between different groups of students?
- What happened to the control group during the period of the intervention?
- Is there any evidence of diffusion such that the treatment became more widely used (outside the treatment group)?

In evaluation studies, this second element is widely known as the 'process evaluation', 'implementation evaluation' or 'formative evaluation'. However, the idea of studying the context extends to all research designs such as randomised control trials, efficacy or effectiveness studies, cohort studies, quasi-experimental studies, cohort approaches, and beyond. Systematic data collected on various aspects of any study can generate explanation of the reliability and validity of achieved results (McMillan 2007, Humphrey et al. 2014). However, the emphasis here remains on robust research designs where process evaluation is only a component (Morris et al. 2016). It is sometimes said that research design is the 'soul' of any good research project, while fair reporting of the process adds 'flesh and blood' to the main structure of a research report.

Evaluation studies report on the process of the intervention implemented, including details of the settings and how closely the treatment of intervention was followed in relation with the prescribed method or protocol. According the UK Medical Research Council, process evaluation is used to understand the functioning of interventions or treatments, specifically where multiple and complex components interact and/or intersect with each other and produce change in expected outcomes (Moore et al. 2015). These guidelines for embedding process evaluation in the protocol for research in health and medical sciences are now increasingly adopted for evaluation studies in education and social sciences.

Fidelity to the treatment (and, where appropriate, to the research process) are key aspects of the process evaluation reported in well-conducted RCTs. The reports, published by the Education Endowment Foundation, an education charity established in 2011 that funds RCTs in England, focus particularly on reporting outcomes of the process evaluation of the intervention process. One of their five criteria for judging the security of the findings touches upon the validity of the impact outcomes in the light of process evaluation (EEF 2014). There is existing evidence that fidelity to the treatment is correlated to some extent with the impact outcomes in education trials (Topping 2017, Durlak et al. 2011, Davis 2014). However, there are currently few robust trials in education where additional issues of fidelity to the research process are systematically assessed and meticulously reported.

There is a clear distinction between assessing the fidelity of different processes (i.e. fidelity to the treatment and fidelity to the research design). Both are important for research findings. Fidelity to the treatment concerns assessing implementation in relation to the protocol of the intervention, and fidelity to the research design is about how carefully the research design protocol was observed. The aggregated trials reported here were led by school leaders; therefore, reporting on the fidelity to the research design protocol was one of the main objectives of the evaluation study.

In design-based projects, studying the context for intervention delivery and of the research design is an integral part of supporting the achieved results. However, the methods used depend on the researcher's choice, time, availability of resources and access to respondents. Whatever method(s) is selected for collecting data on process, the aims would largely be to generate information on intervention implementation, the amount of treatment offered and/or received by the participant (termed 'dosage'), fidelity to the protocol of the treatment and research process, limitations of implementation, assessment of the Hawthorne effect (participants' awareness of being in treatment group) or of John Henry's effect (participants' awareness of being in the comparator group), perceived impact, and evidence of diffusion of the treatment to the control group (Gorard and Siddiqui 2017, Gorard 2013).

**The structure of this paper**

In this paper, we explain the concept of aggregated trials in schools, and the protocols we followed for the process evaluation in order to gain more information than usual on the feasibility of conducting aggregated trials. The two RCT studies are discussed separately at first. In each example, we have outlined the RCT evaluation design, summarised the impact findings, and those achieved from the process evaluations. The main conclusions drawn are based on the general findings from these two studies, which have wider implications for research practice and recommendations for research schools and school research leads.

We were commissioned to oversee evaluations of two popular literacy interventions that are widely used in UK schools (Fresh Start and Accelerated Reader). The two educational evaluation projects were actually run day-to-day by the participating schools themselves. Our role was to help with design advice, explain how trials work to the school leads, monitor events, aggregate the eventual results, and assess how good schools were at running their own interventions with evaluation. We provided school leads with an initial workshop on conducting small-scale trials so that they could understand the process and efficiently generate some basic results. There was another later workshop on analysing and presenting results. The randomization was at pupil-level in each school, conducted by schools, and reported to us. At the end, we aggregated the results from all the participant schools for our main evaluation report.

The primary research objective was to assess the impact of each literacy catch-up intervention on pupils' reading ability. However, we were also interested in estimating the feasibility of conducting aggregated trials directly managed by school leads. Evaluation of research practice was an integral component of our research design that generated information on the needs, challenges and advantages of involving school leads in independently conducting the trials in their schools. We focused on gathering information on school leads' experience of conducting research, teachers' potential bias in selection, randomization and group allocation, and the fairness of procedures involved in testing. Guided by what we achieved in the evaluation of research practice, we could interpret the effects of interventions with caution and, at the same time, realize the potential of carrying out more school-based aggregated trials. We comment on the advantages and disadvantages of school leaders conducting randomised control trials, and make our recommendations on the development of the aggregated trials approach in education. We also make recommendations as to whether these two interventions were fully developed for scaling-up, including bigger samples and conducting larger effectiveness trials for each.

*Ethical considerations*

Ethical approval was provided by the Durham University Ethics Committee. The project was conducted in accordance with the School of Education Code of Practice on Research Ethics and in line with the British Educational Research Association's 'Revised Ethical Guidelines for Educational Research' (2004). There was assured anonymity and confidentiality for all participants. No individual pupil, teacher or school is identified or identifiable in the evaluation reports or in any published documents. Schools and individual organisations obtained opt-out parental consent for activities and to be part of the evaluation.

## Aggregated trials

The concept of aggregated trials is relatively common in medicine, epidemiology and the health sciences (Chen et al. 2016). There is clear scope for attempting a similar approach to generating results in education. This could be achieved by involving schools and teachers as participant-researchers and data collectors. Aggregated trials could be a way forward in evidence-based education. However, before implementation on a wider scale, issues such as the feasibility of schools conducting experiments, teachers' awareness and practical knowledge of conducting fair experiments and other factors need to be assessed. Little is known about such teacher-led aggregated trials, and our study is among the first in the UK.

The background to the trials discussed in this paper is as follows. Individual secondary schools and small networks of schools applied to a national funder (EEF) to conduct evaluations of two literacy interventions. The numbers of schools in each application were grouped and the funder proposed two aggregated trials with pupil-level randomisation. The schools in each trial had to agree to certain similarities in their timing, design and outcome measures, and to be overseen by independent evaluators. As independent evaluators, our role was partly to assist with the evaluation where needed, to aggregate the findings from all of the networks in each trial, and partly to assess how good schools are at running their own evaluations. If the schools were successful in managing trials, then this opens the way for large-scale, even national, ongoing trials of the kind proposed for general practitioners in health care (Goldacre 2012).

None of the schools had implemented their chosen interventions before, and all had high proportions of children performing below the expected Key Stage 2 literacy levels (an expected level of attainment for 11-year-old pupils at the end of primary schooling in England). The interventions were aimed at improvement in literacy outcomes for children who found reading and comprehension difficult during the transition stages from primary to secondary school years, and so may have had further difficulties accessing the more varied curriculum at secondary school. In both projects, the research designs involved pupil-level randomization to treatment and control groups, standardized-tests as outcome measures, Key Stage 1 reading results (performance in national assessment at the end of Year 2, when pupils are 6 to 7 years of age) or a standardized pre-test as baseline measures, and an external process evaluation.

One intervention was Fresh Start (FS) phonics reading, which is a highly structured approach towards teaching literacy through systematic phonics. In FS, the individual letters are sounded out within words, and these sounds are then blended to form the pronunciation of the word, and so to 'read' it. The 44 basic sounds, used as building blocks, are taught first, rather than the letters. When writing, the combination of sounds is said aloud and then converted to letters and written on the page (Brooks, 2003). It is implemented as a literacy catch-up approach for pupils identified to be at risk of failing to achieve expected literacy levels. The targeted sample included 433 Year 7 pupils (ages 11 to 12) who were randomised into treatment and control groups (for full details of this trial, see Gorard, Siddiqui and See 2016).

The other intervention was the Accelerated Reader (AR) programme, which is a computerized intervention that assesses students' initial reading level and then suggests texts for reading that best match with a student's abilities and interest in reading. The programme also assesses the performance of students regularly, and at each phase increases the reading challenges. The teacher's role is to motivate students to read, and support the process of regular reading and using the AR programme. AR was implemented here as a literacy catch-up approach, again for pupils identified as at risk of failing to achieve expected literacy levels at Key Stage 2. The sample included 349 Year 7 pupils who were randomised into treatment and control groups (for full details of this trial study see Siddiqui, Gorard & See 2016).

The main impact outcome was assessment of the effect of each programme on disadvantaged pupils' literacy scores on a standardized test called the New Group Reading Test (NGRT). We planned the studies as efficacy trials, lasting 20 weeks. Following workshops on conducting trials in schools, school leaders identified the groups of pupils who were at risk of not attaining the expected literacy levels. A pre-test was conducted with all identified pupils, after which pupils were randomised by the schools to treatment and control groups. The control group would receive the same treatment in all schools after the 20 week trial. School leaders conducted the randomization as explained to them in the workshop and sent us the results of

randomization and group allocation. The schools organized delivery of teacher training programmes, pre- and post-tests, implementation of the programme in the schools, and arranging and recording pupils' attendance in the sessions (which, in both of the interventions, were outside the mainstream classroom).

**Outline protocol for process evaluation**

The process evaluation concerned both how the given intervention was conducted, as should be standard, and how the schools managed the impact evaluation component. For convenience and due to constraints on time, access and available resources, we used a variety of research methods to collect the data. For each trial, we conducted participant observation in sessions that were training teachers about the intervention, observed 15 complete sessions of the actual intervention in operation in different schools, interviewed staff, students, and school leads, used teachers' logs as secondary data to assess the regularity of the sessions and pupils' attendance, and followed the Fresh Start and Accelerated Reader developers' website materials. Also, the intervention protocol documents were read thoroughly, in order to understand the mechanism of the interventions.

*Observations*

We observed teacher training as participants, where that was possible. FS training consisted of two days of delivery, while AR training was one day. In the training, around 100 experienced literacy teachers and 40 teaching assistants participated. We shared our protocol for observation of the training with school leaders, which involved informing all stakeholders of their leading role and the independent evaluators' involvement in the process. The school leads introduced the evaluation team members as participant observers. The evaluators participated in all training activities along with teacher trainees. The school leads made all teacher trainees aware that the evaluation advisors were not inspectors and were not visiting for school monitoring purposes. The school leads ensured that all teacher trainees felt secure and free in sharing their feedback and experience of training with the evaluator – which was noted anonymously in terms of participant and school.

The guidelines followed by the participant observers involved establishing a rapport with the teachers so they did not feel they were being judged. Our guidelines for observations allowed them to ask questions about the evaluation and evaluator's role. Whenever required, we gave sufficient explanation of the evaluator's role. In the training sessions, we became participating members of the group and performed the training activities in the same way as other teachers. This helped us in understanding the mechanism of interventions that yielded overall outcomes. In order to keep to the process of discovering teachers' views, we were intentionally informal in asking teachers and teaching assistants their experiences and views of the training, during breaks and after the training. This method yielded in-depth information on teachers' understanding of the intervention and its possible effects on pupils. We asked teachers about the perceived relevance of the programme for their teaching context and recorded detailed field notes, which were shared with the other evaluation team immediately after the observation was completed.

Two main reports were developed, which included detailed field notes, narratives of teachers and the general experience of the evaluators as participant observers. The notes were coded into major themes and the final reporting involved incorporating the observed evidence on the feasibility of implementation.

We observed the actual implementation of the interventions at classroom level. These observations led to records of the events in the classroom during the implementation of the intervention. The teachers were informed of our planned visits for observations and were made aware that the purpose of the observations was for research and not to assess their teaching performance. The classroom observations followed the guidelines where the evaluation advisors conducted non-intrusive observations. The teachers could introduce the evaluation advisors to pupils and inform them of the reasons for their presence in the classroom. We recorded detailed notes of the events during the session and asked questions, if needed, after the session was over. We recorded the use of teaching materials and details on pupils' participation, engagement and attentiveness during the session. We specifically recorded our comments on the actual classroom session in relation to the actual protocol of the interventions.

For each trial, there were 15 observations of classroom sessions. The in-depth data from the classroom sessions were mainly useful concerning the fidelity of the treatment protocol in actual settings. These 15 sessions gave insights into the feasibility of implementing both interventions in schools and into pupils' responses to the intervention, as well as how keen they were in engaging with these interventions. Treatment fidelity and pupils' engagement with the intervention can have important effects on outcome results and therefore the findings achieved from observations were incorporated as evidence in the research reports. However, we are cautious about the status of information gathered in the presence of evaluation advisors. The settings could have been different or rather more natural in the absence of evaluation advisors. However, classroom observations were deemed the best possible compromise here. We have presented the evidence after the triangulation of the achieved information, which is another way to enhance the reliability of process evaluation data.

*Interviews*

In the schools, evaluation team members conducted face-to-face interviews with 14 staff members, 20 students and 10 project leaders per trial. These interviews were semi-structured conversations about participants' experiences, general feedback and the perceived effects of the programmes. We set interview guidelines and followed a standard protocol for collection of the data in the interviews, but allowed the conversation to follow a natural course. This permitted us to find out about aspects of the intervention of which we were not aware. The interviews with school leaders involved questions about the process of selecting the target group of students. In the interview, we asked for detailed information on the process and timing of randomisation, and who was involved in the process. We asked about the challenges of conducting RCTs in schools, for feedback on the programme, training received and on teachers', parents' and pupils' responses to the programme. We also asked the staff members about the assessments conducted before and after the intervention.

Face-to-face interviews with school leads involved inquiry about their roles as trial managers and school or research leads and their general experience of being part of the research study. School leads were trial managers in their schools and were responsible for the implementation of the intervention programme. We conducted these interviews during the trial. School leads' role and engagement in the research process were crucial to the impact findings. Therefore, we maintained contact with them through the regular exchange of emails and by providing regular advice on the trial management procedures.

The interview data were partly transcribed and the evaluator's reports were generated based on

the interview process and field notes. We triangulated these sets of information and incorporated the information achieved from the regular exchange of emails and telephone conversations. Interviews with school leads were one of the richest sources of data about the advantages and challenges of the aggregated trials. Based on the recurring themes in the interviews, we assessed the trustworthiness of the impact findings and made recommendations for aggregated trials in education for future studies.

Teachers and teaching assistants who had been trained for the programmes and who were involved in the implementation were also interviewed. The interviews were again semi-structured, based on a scheme of questions but permitting interviewees to tell their stories. We included questions on the context of teaching, and the challenges of working with pupils who were at risk of failing in literacy. We asked for feedback on the quality and relevance of the training and the challenges of intervention implementation. The interviews were recorded as field notes and parts of the interviews were transcribed. We used these data to assess the perceived impacts of the programmes on pupils, and teachers' experiences of the intervention implementation.

*Other resources*

The programmes under evaluation had considerable resources on their websites for schools and parents. We compared the information from attending the teacher training with the information available on the programme websites. The programme websites were important sources of information for the understanding of the intervention aims and pedagogy. The website resources were also informative about the feasibility of treatment as promoted by the developers in relation to school contexts.

We also examined pupils' session attendance records maintained by teachers and school leaders. We obtained session attendance data in order to consider the amount of treatment (dosage) given to pupils. Treatment dosage was one of the explanatory variables in the main outcome findings of the study.

**Brief findings from the Fresh Start phonics reading programme**

The control group had higher pre- and post-test scores than the treatment group (the pre-test 'effect' size was -0.36, and the post-test 'effect' size was -0.19). Using the progress scores from pre- to post-test, the intervention suggested a positive impact on reading comprehension (+0.24). The same 'effect' size occurred when only Free School Meal (FSM)-eligible pupils were considered (FSM is a measure of family poverty or relative disadvantage). The imbalance at pre-test stage was large and should lead to some caution about the validity of the final results. If students at high risk of failure were somehow selected for the treatment, and their scores were easier to improve, this could appear as a bigger change in their performance as a result of intervention. We do not think this happened here, as the lower attaining half of all students did not make more progress than the rest.

Imbalance between the groups, even after blind randomisation, occurs sometimes by chance – it is, after all, a key component of what chance means. Nevertheless, we gained some useful insights in our interviews with school leads, who conducted the randomisation of the targeted group. We found that a few of them were slightly biased towards selecting the most disadvantaged students for the intervention immediately. The school leads explained in the

interviews that they believed some students were at high risk of failing, and so needed immediate intervention. Therefore, they felt that they deserved more chance to receive the treatment than the rest. A similar phenomenon has been observed in medical trials involving practitioners (Kendall 2003). One of the school leads also told us retrospectively about their decision to swap students from the control to the treatment group, after three students in the treatment group left a school. Although baseline equivalence of the treatment and control group did not suggest any difference, this is poor practice and could have compromised the overall result. Cases were analysed in terms of the group to which they were randomised, whether they received the intervention or not. The message of selecting pupils highly at risk of failing in literacy was also emphasised in the training by the intervention developers in FS. The strong requirement that 'random means random' explained in the workshop was not enough here – probably because the FS trial also involved the FS developer (whereas the AR trial had no developer involvement and had no such problem with randomisation).

Therefore, our process evaluation suggested among other things that imbalance between the randomised groups in FS may not be due to chance. In our workshops for teachers, we made school leads aware of unconscious bias and the consequence of biased group allocation. However, it is possible that the initial workshop could not effectively convey the message or give sufficient examples of random selection procedures. We could probably introduce the methods of blinding the group allocation process by an independent party, which we did not suggest to the schools here. In future, we would suggest, teachers' unconscious biases need to be given emphasis and practical rules given to teachers to avoid the problems of deliberate imbalance in grouping.

During our participant observation of the teacher training, we found that the intervention had specialist resource material, which covered a systematic plan of learning decoding followed by reading comprehension. The intervention teaching style is a core element of this intervention, which encompasses teacher's passion, praise for pupils and a dynamic pace for the lessons. The classroom management and teacher-pupil communication techniques are prescribed in the training and teachers' handbook. This suggests that this intervention cannot be successfully implemented on its own, without training or the teachers' handbook.

The schools involved teachers and teaching assistants in receiving the FS training. There were concerns that FS-trained teachers could implement the same knowledge and practice when working with students who were in the control group. All school leaders assured us that they would prevent such diffusion - but we found, for example, a FS phonics chart displayed in a classroom which had students from the control group. Although only a relatively small slip, awareness of the treatment among the control pupils can reduce its apparent effectiveness. The teacher reported that it was simply a visual aid used for teaching students in the treatment group during other sessions.

Pupils on FS were observed during the sessions and some of them were interviewed after the sessions. The Fresh Start teaching style kept almost all pupils thoroughly engaged throughout the sessions. The selected pupils who were struggling to read at age 11 seemed to have enjoyed the practice of phonics through various mnemonics. Each session included one FS teacher and one FS-trained teaching assistant, so the pupils were seen to be given a lot of support and individual attention, which they might not have received in the whole group. In the interviews with pupils, many reported that they preferred coming for FS sessions rather than going to other lessons. Schools generally received positive feedback from parents when they were informed about participation in the FS intervention. However, according to the interviews with teachers,

there were also parents who raised concerns about their children's participation in such a basic programme. In a sense, for the pupils at risk of failing in literacy, FS was appropriate; therefore, there was an advantage for them, but for those whose reading was perhaps a little more advanced it could be seen as patronizing.

**Brief findings from the Accelerated Reader programme**

In this trial, 349 pupils were assessed on Key Stage 1 results in English to establish a baseline equivalence, and at the post-intervention stage a standardized test was conducted as the main outcome. The treatment and control groups were balanced on Key Stage 1 results in English with an 'effect' size of zero. This also meant that the post-test differences between the two groups would give us a truer estimate of the impact of AR. At the post-test stage, the final effect size was +0.24, with results at least as good for FSM-eligible pupils.

In general, the trial was well-managed by the school leads. However, as in the FS aggregated trial, this one also had some issues in the process of pupil randomization. We found during our school visits and interviews with the school leads that one of them allocated entire pre-set groups into treatment and control without pupil randomisation. Although the aggregated results showed zero effect size at the baseline, this school did not adhere to the protocol of pupil-level randomisation. The challenges reported by the school lead were issues in the school timetables that could not allow flexible arrangements to randomize students and implement AR in different groups once they had arrived at school for pre-testing. However, they had not raised this objection at the training session. Pupil-level randomisation is ideal for experiments. However, we found that in real-life settings there are barriers to adhering to the randomisation process and general threats of diffusion. Schools may subvert the process, usually with good intentions.

The interviews with teachers provided useful feedback on their experiences of using AR. A general response was that this intervention had a very different approach towards teaching literacy and therefore needed a separate allocation of staff and student time and other resources, such as technology. According to teachers' feedback, their overall preference was to use AR as an additional or supplementary approach rather than for the mainstream teaching of literacy.

We interviewed pupils and received feedback from those who received AR. According to some pupils' feedback, reading was not generally their chosen leisure activity. Some students reported that reading books was what they had been asked to do by teachers - to read and take quizzes. Some students enjoyed the technology component of the AR programme such as taking quizzes on tablets, reading electronic books and browsing on the internet.

In some groups, teachers incentivized the reading progress for the whole groups. According to the teacher, this was effective for the engagement of those students otherwise not showing interest in reading books. We could not exactly measure through this small scale aggregated trial whether reward and motivation were core effective elements of the AR or if it was the overall approach. However, the process evaluation drew attention to further paths of inquiry to assess whether the teachers' strategies to stimulate students' interest and engagement in reading were at least partly responsible for the impact.

**Discussion**

Potential threats to validity in the aggregated trials were discerned and reported as part of the process evaluation. It is hard to envisage how else they could have been discovered. These included possible diffusion of the intervention to the control group (as exemplified by an intervention resource displayed in a general classroom), moving a few cases between the groups to which they were originally allocated, and using class rather than individual randomisation for practical reasons. Threats can compromise the overall results of trials, although these are all relatively minor examples, and merely modify our assessment of the security of the results downwards. The teacher training for trial management can strategically emphasise the importance of overcoming or avoiding these threats by giving school leaders various examples of effective randomisation techniques and demonstrating useful methods to tackle practitioner bias.

The process evaluations expanded our knowledge about the achieved impact results, feasibility of implementation and the challenges that need addressing before conducting aggregated trials in education. These two trials showed considerable promise for future aggregation of smaller trials run by schools themselves. Aggregated trials are useful and have various merits in education research. Involving school leads in the research process was associated with no school dropout, and very little pupil dropout after randomisation. Of course, this is probably also due to the fact that all schools had both treatment and control groups. In other studies, dropout is a general threat to the findings of the study but, if school leads are given control to manage a research process, they are more likely to utilize their available means and access to pupils in preventing dropout. Schools are good at getting permission for trials, giving them the impetus they need for implementation, and at keeping records of attendance. There was a general enthusiasm among school leaders to engage in the research process. The workshops were very well attended, with full participation by school leads. Some school leads even managed to conduct school-level analyses for their own use and information. We found this very encouraging and it has led us to promote basic research skills for robust evaluations and for this content to be incorporated into teacher training courses (as an alternative to the very weak types of "action research" commonly suggested).

We found that the pupil selection bias in randomisation by school leads was the biggest challenge in conducting aggregated trials led by teachers. However, we think that this challenge and others can be addressed in training and workshops: teachers and schools do need training and reminding about all aspects of managing a trial. If teachers were given more examples of conducting small-scale experiments and shown a variety of ways to avoid bias in pupil selection, then teacher bias could be addressed to some extent. The key is to understand the importance of gaining secure knowledge for the longer term, rather than issues of treatment in the short term (in this wait-list design).

Impact studies investigate feasibility, efficacy and effectiveness of the applied interventions and approaches. However, new interventions and approaches emerge from details of the process evaluation itself. As in the case of these two aggregated trials, what works for student improvement could be as simple as giving pupils appropriate challenges, or simply rewards or individualized attention (Gorard et al. 2017).

Based on the formative data, we wrote two trial reports that were comprehensible to practitioners, school leaders and general readers. The impact results were reported along with in-depth details of the context in which the research was conducted. Independent evaluations are not only about reporting impact results but also providing sufficient details to interpret the

findings and make appropriate judgment about the trustworthiness of the final results.

## Conclusion

Process evaluation provided a great deal of information on the mechanism of two complex interventions. Structured interventions, regularly implemented in small group settings separated from the mainstream group, were the most important mechanisms that seemed to have an impact on pupils' literacy attainment during the transition phase from primary to secondary school. The effectiveness of these active elements of interventions could be assessed in future evaluation projects.

Process evaluation helped us in judging the feasibility of aggregated RCTs where school leaders managed the evaluation process. The major implications of these two evaluation studies encourage the concept of 'research schools', where teachers can be researchers who can implement research design-based evaluations and, at the same time, can develop evidence-based teaching practice. As in case of these two aggregated trials, we found that the school leads' role in research is underestimated. However, based on the information collected from the evaluation of their research practice, we have put forward recommendations on how to make school leads' judgement in the research process more reliable.

A large volume of 'soft data' was collected through various means but the views and perceptions that emerged from soft data are not conclusive. This could be because these are participants' perceived impacts rather than actual impacts and there could be biased opinions and judgements. The need to report this information is, though, important, as it clarifies the impact results in a wider context of research and practice. However, it is no substitute for the impact results themselves.

Further implications of this study (and others we have conducted) also suggest that face-to-face observations in schools are essential because the actual implementation and fidelity to the treatment and to the research process cannot be assessed only through on-line surveys or telephone interviews. In other words, it needs to be seen. Methods implemented from distance to the actual research settings are still far too common and 'comfortable' for researchers in evaluation studies. However, there could be serious concerns regarding the reliability of information derived from these methods. We recommend that any process evaluation framework must incorporate recording first-hand information, which should be independently gathered from research sites.

## References

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

Chen, R., Desai, N. R., Ross, J. S., Zhang, W., Chau, K. H., Wayda, B., ... & Krumholz, H. M. (2016). Publication and reporting of clinical trial results: cross sectional analysis across academic medical centers. *British Medical Journal*, 352, i637.

EEF (2014) Classification of the security of findings from EEF evaluations, https://v1.educationendowmentfoundation.org.uk/uploads/pdf/Classifying_the_security_of_EEF_findings_FINAL.pdf

Davis, D. (2014). *Fidelity of Implementation, Teacher Perspectives and Child Outcomes of a Literacy Intervention in a Head Start Program: A Mixed Methods Study*. University of Nebraska: http://digitalcommons.unl.edu/cgi/viewcontent.cgi?article=1203&context=cehsdiss

Durlak, J. A., Weissberg, R. P., Dymnicki, A. B., Taylor, R. D. & Schellinger, K. B. (2011), The impact of enhancing students' social and emotional learning: A meta-analysis of school-based universal interventions. *Child Development*, 82: 405–432.

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

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

Gorard, S., Siddiqui, N., & See, B. H. (2016). An evaluation of Fresh Start as a catch-up intervention: a trial conducted by teachers. *Educational Studies*, 42(1), 98-113.

Gorard, S., See, BH and Siddiqui, N. (2017) *The trials of evidence-based education*, London: Routledge

Gorard, S. and Siddiqui, N. (2017) There is only research: the liberating impact of just doing research, *International Journal of Multiple Research Approaches*, 10, 1, 1-115

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.

Humphrey, N., Lendrum, A., Ashworth, E., Frearson, K., Buck, R., & Kerr, K. (2014). Implementation and process evaluation (IPE) for interventions in education settings: A synthesis of the literature. (Available at: https://educationendowmentfoundation.org.uk/public/files/Evaluation/Setting_up_an_Evaluation/IPE_Guidance_Final.pdf )

Kendall, J. M. (2003). Designing a research project: randomised controlled trials and their principles. Emergency Medicine Journal, 20(2), 164-168.

McMillan, J. H. (2007). Randomized field trials and internal validity: Not so fast my friend. *Practical Assessment Research & Evaluation*, 12(1).

Moore, G. F., Audrey, S., Barker, M., Bond, L., Bonell, C., Hardeman, W& Baird, J. (2015). Process evaluation of complex interventions: Medical Research Council guidance. *British Medical Jounral*, 350, h1258.

Morris, S. P., Edovald, T., Lloyd, C., & Kiss, Z. (2016). The importance of specifying and studying causal mechanisms in school-based randomised controlled trials: lessons from two studies of cross-age peer tutoring. *Educational Research and Evaluation*, 22(7-8), 422-439.

Nelson, M. C., Cordray, D. S., Hulleman, C. S., Darrow, C. L., & Sommer, E. C. (2012). A procedure for assessing intervention fidelity in experiments testing educational and behavioral interventions. The Journal of Behavioral Health Services & Research, 39(4), 374-396.

Oakley, A., Strange, V., Stephenson, J., Forrest, S., & Monteiro, H. (2004). Evaluating Processes A Case Study of a Randomized Controlled Trial of Sex Education. *Evaluation*, 10(4), 440-462.

Oakley, A., Strange, V., Bonell, C., Allen, E., & Stephenson, J. (2006). Process evaluation in randomised controlled trials of complex interventions. BMJ (Clinical research ed.), 332(7538), 413-416

Siddiqui, N., Gorard, S., & See, B. H. (2016). Accelerated Reader as a literacy catch-up intervention during primary to secondary school transition phase. *Educational Review*, 68(2), 139-154.

Topping, K. (2017). Implementation fidelity and pupil achievement in book reading: variation between regions, local authorities and schools, *Research Papers in Education*, 1-22.