Editorial

Randomised Controlled Trials (RCTs) in education research – *methodological debates, questions, challenges*

Introduction

This Special Issue of Educational Research Journal shines a spotlight on a key research design in the 21st Century in the field of education research: the randomised controlled trial (RCT) or 'true' experiment. The six papers included here explore the history and future of the design. They focus on challenges and opportunities, methodological developments and innovation, but above all they highlight the immense progress that has been made in rigorous evaluation over the last sixty years. They provide a historical background in the United States of America (USA), in Scandinavia and in the United Kingdom (UK), and cover the aspects of politics and methodology that have shaped the recent education science landscape. In a time when answers are sought to questions of efficacy and effectiveness of education policies and practices, RCTs have a special role. Uniquely among research designs, they are able to obtain unbiased estimates of the average effects of these policies and practices on children and young people's education and wider outcomes. Furthermore, they serve as a useful introduction to contemporary issues in RCTs for any education researcher who has an interest in them, but may have felt hindered by limited technical knowledge.

Randomised Trials in Education in the USA (Hedges and Schauer)

In their paper, Larry Hedges and Jake Schauer give a succinct overview of the history of educational experiments in the USA, from the first known randomised controlled trial (RCT), through five historical periods to the state of the art in the present day. This is an interesting and illuminating history, drawing together the research and educational contexts of each period, the methodological innovations and the political developments. Many issues they discuss have lessons for experimentation in education today.

One of the most striking developments highlighted in Hedges and Schauer's paper is the almost complete abandonment of experimental research in the USA for 20 years from the 1980s to the early 2000s. This move away from the 'experimenting society' of the 1960s and 70s was driven, largely, by the absence of treatment effects from randomised experiments of most of the interventions evaluated. However, rigorous evidence of no effect, or indeed of harm, of educational interventions is important scientific knowledge. Once we have evidence of this nature from RCTs, we can communicate the results so that policy makers and practitioners cease to promote their use, and funders and researchers move on to evaluate other promising interventions which may prove to be of benefit.

A rigorous RCT that shows no effect of an intervention is not a failure of the design used to evaluate it. Instead it gives us precious hard-won knowledge that the intervention is either ineffective or has negligible benefits in relation to its costs. The modern (post 2000) RCT may be as likely to find no benefit of an intervention as did a trial conducted in the 1960s and 70s, unless lessons are learned about the nature of readiness for evaluation of potential interventions. Interventions which do not have a strong rationale for evaluation using RCT design include those which are not yet fully developed and replicable, should a positive impact be demonstrated. Compelling rationales for evaluation are essential pre-requisites for evaluation and include the practical significance of an evaluation, or current widespread use without rigorous evidence of impact or strong theoretical

basis or evidence of promise from developer-led pre-experimental or quasi-experimental designs. Moreover, with an emphasis on avoiding pit-falls such as publication bias (i.e., the issue of trials with negative results being less likely to be published than trials with positive results), improving the rigour of the design of RCTs is critical, as poorly designed trials are more prone to finding positive effects. We should discourage policy makers and funders from the conflation of lack of positive findings with the use of the RCT design, as was done by Cronbach and colleagues.

A key lesson we can learn from history is to embrace 'null' or negative findings as enthusiastically as we tend to do with positive findings. Fortunately, as Hedges and Schauer describe, the impact and influence of the key educational funder of trials in the USA - the Institute of Education Sciences (IES) - from 2002, shielded from political interference, will ensure that the history of the 1980s will not repeat itself. Current and future generations of educational researchers will embrace experimental research design and use it appropriately to establish with scientific rigour which educational interventions work.

The Trials of Evidence-Based Practice in Education: A Systematic Review of Randomised Controlled Trials in Education Research 1980-2016 (Connolly et al.)

Connolly et al. have carried out a timely systematic review of education RCTs that has a specific focus on research practice. It contains information of critical importance to this Special Issue's intended international perspective. We learn that just over half (53%) of all the English language RCT reports identified were conducted in North America, with a little under a third (29%) undertaken in Europe. This reflects other papers in this Special Issue that concern the history of education RCT evolution. The development of experimentation in the USA, covered so lucidly by Hedges and Schauer, spans over 50 years of often intense activity. The two papers concerning the wider European picture (Dawson et al. and Pontoppidan et al) illustrate developments over a much more recent timescale, often within a decade.

Two of the main points Connolly et al. make are the importance of process evaluation and the need for trials to reflect upon the implications of their results for theory. These relate to a key finding of the one other paper that took an overview of several trials: the paper by Dawson and colleagues. The two papers present these conclusions from very different standpoints. Connolly et al. highlight the lack of process evaluation in two thirds of reviewed trials with nearly 40% failing to reflect on the implications of their findings for theory as a possible reason why criticisms of RCTs as being simplistic and atheoretical continue. Dawson et al. point to lack of sophisticated process evaluation leading to us not understanding why an intervention did not work. The authors in this paper do not refer to theoretical implications, their viewpoint reflecting the highly practical considerations of a funder under pressure to demonstrate positive effects.

Both the paper and the book (reviewed at the end of this Special Issue) by Connolly et al. are influenced by the trenchant paradigm wars of recent times. Connolly is a rare example of a user of RCTs who fully engages with opponents of their use. He creates measured arguments that incorporate and address their reservations. This approach is perhaps more useful than adopting a policy of 'going it alone', as many triallists do, since it endeavours to persuade others of the merits of RCTs in education. Furthermore, it potentially enhances trial design by, for example, encouraging more in-depth process evaluation derived from engagement with experts in other research designs.

Connolly et al. advocate the use of subgroup analysis as a route to finding out 'what works for whom', noting that only half of the studies in their review included a subgroup analysis. We would

agree this is an important element of analysis of an RCT but caution against over-interpretation. Using frequentist statistics, even when all subgroup analyses have been pre-specified to avoid 'data dredging', such analyses are vulnerable to the 'familywise error rate'. This is the notion that the probability of falsely concluding there is an effect increases with the number of hypothesis tests performed. Furthermore, randomisation tends not to occur within the subgroup itself and certainly does not allocate individuals to the subgroup. Therefore, any such subgroup evidence should be a spur for a future trial of the intervention *on that subgroup* rather than viewed as a definitive conclusion.

Methodological challenges in Education RCTs: Reflections from England's Education Endowment Foundation (Dawson et al.)

This Special Issue would have been incomplete without a contribution from the organisation that has revolutionised the education research landscape in England over the last seven years. Before the advent of the Education Endowment Foundation (EEF), RCTs in education were merely imagined by some enthusiastic proponents, who were only very occasionally granted the opportunity to run one. The first large scale pragmatic trial of a curriculum intervention in England evaluated 'Every Child Counts' (the last UK Labour government's flagship numeracy policy for pupils in primary school) (Torgerson et al, 2013), and was funded by the Department for Education, possibly partly due to pressure from a small number of MPs to use the design to evaluate government education policy. It was funded, designed and conducted just before the EEF was established. The paper by Dawson et al. describes the opening of the floodgates that ensued, with the concomitant challenges experienced by a new organisation funding and overseeing over 100 education RCTs in a very short period of time. The precise alignment of forces that resulted in this unprecedented change remains unclear, with the authors citing that their use was partly prompted by leading academics. This is undoubtedly true but there was also enthusiasm from civil servants. However, it would not have happened without the political will to allow evaluators enough power to use the most appropriate design for important questions of effectiveness in relation to policy and practice. By enabling researchers to randomise participants to receive an intervention or not enabled them to establish a research study with strong casual inference, rather than the alternative research in response to roll out and belated requests for evidence of an effect post hoc, as tended to be the practice before 2011.

Perhaps the most important point made in this paper is embedded within 'challenge 3': the realisation that when rigorous evaluations are performed using randomised experiments, the resulting effect sizes are much smaller than previous research in the same substantive area. Note the historical parallel in the USA described by Hedges and Schauer. 'Previous research' is an umbrella term that may include individual studies, meta-analyses, meta-meta-analyses (averaging the effect size from many meta-analyses) and highly influential books (for example Hattie's overview of reviews, 2008). This has implications for the choices that educators make in terms of which (if any) intervention to use and for the funding of RCTs in a politicised environment where positive effects matter. The latter point is addressed elsewhere within this editorial. Dawson el al. focus on the impact of these smaller effect sizes on research planning. A critique of Hattie (2008) is out of scope here and is not attempted by the paper's authors but the relevant point to note is his inclusion of non-randomised designs within his meta-meta-analyses, which may have led to an artificial inflation of effect sizes. His book was published only three years before the Education Endowment Foundation (EEF) was established as the main funder of education trials in the UK. Given its influence, it was perhaps inevitable that its inflated effect sizes would guide some EEF-funded evaluators towards inadequate sample sizes for their trials. One limitation of the EEF's earliest RCTs

was to leave independent evaluators powerless to change sample size, it having previously been determined by the intervention developer. This was swiftly rectified but did not stop some early trial planning being influenced by the grandiose claims of the very recent past regarding the effectiveness of certain types of intervention.

One intriguing aspect of the paper which rings true with the evaluation community in England is the acknowledgement by the authors that some of the interventions funded by EEF were simply not ready for trial. The responsibility for this problem is shared by: the developers, for seeking funding for something that has not previously been adequately tested for evidence that it *could* work; by the funders, for failing to look for sufficient evidence of preparedness; and by the independent researchers, for agreeing to evaluate something that did not demonstrate readiness during the trial preparation phase. The paper's authors propose checking relevant characteristics of an intervention before funding it, a simple solution that on occasion was missed by all three participant organisations.

All of the other challenges discussed in this paper could reasonably be addressed by a more substantial planning and review process before each trial starts. For example, the plan for recruiting and retaining schools for a trial could be peer reviewed. Sample size calculations and any planned process evaluation could be reviewed by a statistician and expert researcher, respectively. Whilst peer review is by no means perfect, it could have mitigated some of the issues documented in Dawson et al.'s paper. Furthermore, it is standard practice for healthcare trials to have an independently chaired expert Trial Steering Committee. Such scrutiny could help to address some of the challenges listed in this paper.

There is some overlap in the papers by Dawson et al. and Connolly et al. An interesting conclusion by Connolly et al. is that over three-quarters (80%) of RCTs in the study found evidence of intervention effects. The paper clarifies that this means statistically significant results and we presume this was widened to include secondary outcomes and more than one measurement time-point, if applicable. Given that the practice of pre-specifying a primary outcome is a relatively recent inclusion in English education RCT design, it seems logical to assume that most studies in the systematic review did not do this and therefore the reviewers looked across *all* measured outcomes for significant effects. A similar metric for English trials funded by the Education Endowment Foundation presented by Dawson et al. is interpreted in a different way, namely that their published RCTs have an average effect size of 0.1 standard deviations and that 71% demonstrated an improvement. It is not clear how many of these were statistically significant, but this result is more likely to be based on a pre-specified primary outcome, noting that not all EEF studies specify a single primary outcome. A statistically significant result is a higher, although by no means particularly meaningful (Colquhoun, 2014), bar for a study's result to cross than simply being a positive effect. On the face of it, therefore, it seems EEF's trials are less successful in demonstrating genuine positive effects than those contained within Connolly et al.'s systematic review. There may be good reasons for this that are nothing to do with the selection of ineffective interventions. EEF's trials are theoretically immune to publication bias since outcome pre-specification and publication are conditions of their funding. Studies in the wider literature are not immune. Furthermore, the EEF insists on independent evaluation, which is relatively uncommon elsewhere and may result in lower effects due to cherry picking of results by developers who are evaluating their own study, for example. This is an example of how the EEF is moving closer to funding the kind of research that attempts to reverse the 'replication crisis' - the notion that most published research findings are false; as described by Ioannidis (2005).

Randomised controlled trials in Scandinavian educational research (Pontoppidan et al.)

The paper by Pontoppidan and colleagues is a very valuable overview of the historical and current state of education experiments in Scandinavian countries (i.e., Denmark, Norway and Sweden). The Scandinavian countries (Denmark and Sweden, in particular) punch above their weight in their use of experiments in health care, and their large epidemiological datasets are the envy of the world. Recently, as Pontoppidan and colleagues show, their educational peers are beginning to develop a culture of experimentation in the field of education. The Scandinavian historical development of educational RCTs echoes the experience of American and UK education researchers. The common experience of Scandinavia, the USA and UK is that small numbers of trials, usually with small sample sizes, are run by researchers with either an interest in special populations (e.g., students experiencing dyslexia) or psychologists and economists, with educational researchers focusing on interpretative non-experimental methods. As with the UK and USA, it took government initiatives, sometimes spurred by change in government with a right wing administration (President Bush) in the USA, a left wing administration in Denmark and a right of centre administration in the UK to implement funding streams to drive forward the greater use of RCTs. Whilst the political leanings of these governments were very different, they did have this one key factor in common: they provided governmental funding to spur the growth of experimental research. However, there remains much to be done. The number of experiments remains small, with Sweden, in particular, producing fewer RCTs than the much smaller Denmark. However, the trajectory is positive and the numbers appear to be increasing, which should lead to improvements in evidence-based educational practice in the Scandinavian countries. Furthermore, the similarities between the three countries seem to be greater than the differences, so lessons learnt in one jurisdiction are likely to be applicable to the others.

Innovation, evaluation design and typologies of professional learning (Boylan and Demack)

Readers who require an explanation of randomisation and what it does, and does not, reveal about the effectiveness of an intervention should turn to the paper by Mark Boylan and Sean Demack for a useful account. The paper promotes the use of experiments to evaluate professional learning programmes whilst questioning their suitability for certain specific types of programme. On the way to reaching this conclusion, a novel typology of professional learning programme is developed. This classification is useful for any developer tasked with producing a theory of change for their intervention and for any evaluator tasked with measuring the impact and monitoring the implementation of a professional development programme.

This paper corroborates the point raised within three other papers of this Special Issue (Connolly et al., Dawson et al. and Siddiqui et al.) with respect to the importance of process evaluation. In particular, and because of the complex theories of change they are concerned with, the authors highlight the fact that, when an intervention has not proven to be effective, it is often not clear whether this was because it was not implemented properly or because the intervention itself was ineffective. Detailed process evaluation does indeed solve this problem. However, there is a school of thought suggesting that as we move from a small underpowered pilot trial, through to intervention delivery under ideal conditions (the efficacy trial) and end with a large-scale pragmatic effectiveness trial, we should eventually be in a position whereby the intervention delivery has already been demonstrated to be of high fidelity. A process evaluation under these circumstances becomes largely unnecessary. However, in practice, this is almost never the case since some aspect of delivery requires alteration to accommodate the scale of a large effectiveness trial and it becomes necessary to test this new formula for delivery with an extensive 'implementation and process evaluation' even if this has been done before.

The most interesting aspect of Boylan and Demack's paper concerns the notion that some types of professional development may not be amenable to evaluation using an RCT. They create a novel typology of professional development consisting of three categories: pedagogical, technical and curriculum. They indicate that an RCT design might well be suited to the evaluation of a technical professional learning programme: one that acts as a mediator and a means by which specific changes in practice are intended to occur. This is contrasted with pedagogical professional learning, involving experimentation (the term used here in its loosest sense) by the teacher rather than the implementation of a predesigned technique. Within this latter context, the authors cite a model where the causal relationship is reversed and teachers change their practice on the basis of a change in pupil outcome which they view positively. For pedagogical professional learning, an efficacy RCT is seen as having inconsistent requirements with the nature of the intervention such as uniformity versus diversity, adoption versus adaptation and fidelity versus variation. The key question for this Special Issue is whether the authors have happened upon a type of professional learning intervention that does indeed not lend itself to evaluation by means of a randomised experiment.

For a discussion of this point, we must first acknowledge that there are indeed types of intervention not amenable to evaluation using experimental design. Programmes that require systematic changes at more than one level of the educational system are difficult, if not impossible, to evaluate using this design, due to the difficulties of randomising at all levels. School funding arrangements and changes to the examination system are some extreme examples but there are more subtle ones: school leadership programmes or interventions that require systematic changes to the school timetable are, at least in England, difficult to evaluate using this design due to the autonomous nature of our school system which means that recruitment to randomisation would be unlikely to be achieved. Do these difficulties extend to professional development programmes, and more specifically to the pedagogical programmes cited by Boylan and Demack? Such programmes certainly have a more complex theory of change than a classroom-based intervention, having to cross the two divides of convenor to teacher and teacher to pupil. However, we are tempted to turn the discussion around and ask whether, ultimately, we are interested in improving pupils' attainment. If this is the case, then it is theoretically possible to randomise a large number of schools or teachers to receive (or not to receive) a pedagogical professional development programme, then to let it run its unpredictable course. That course could even include the reverse causality cited by the authors, since teachers in the intervention arm of the trial should presumably, on average, react differently to a pupil cue than those in the control arm. Attainment tests at the end of the trial would have to be accompanied by an in-depth process evaluation but they would surely still yield a useful causal conclusion? Of course, if a measure of attainment is judged too narrow to capture outcomes of such a diverse professional development programme then we naturally move away from experimental design. It is often forgotten that an ability to measure relevant outcomes well is a pre-requisite for embarking on an RCT.

The importance of process evaluation for randomised control trials in education (Siddiqui et al.)

Siddiqui et al. highlight the critical role of process evaluation to explain impact evaluation results by enhancing understanding of the context in which these results were obtained. In this respect, the answers to 'What works?' questions are complemented (and completed) by a range of implementation research questions, both logistical and methodological. For example, issues which have the potential to enhance or limit the impact, such as contamination between the intervention and control conditions, which could have led to dilution of effect, can be addressed in the write-up of the trial. These are, therefore, important for a full interpretation of the results and also in order to derive lessons for future research.

The paper makes an important distinction between fidelity to *design* and fidelity to *implementation*. The authors illustrate each of these aspects of process through an analysis of two recently conducted RCTs with embedded process evaluations, funded by the EEF. These were both 'aggregated' trials, in which the results of school-run trials were aggregated to produce an overall estimate of effect.

In their analysis of one of the trials, they describe a failure of fidelity to design in which 'switching' between groups was observed to have been done by one school after three pupils left the school who were in the intervention group. The school then switched three pupils from the control condition to the intervention condition, due to a prior belief towards more disadvantaged pupils likely benefitting more from the intervention. This was uncovered during the process evaluation. The impact evaluation analysed the participants in the group to which they were originally randomised (using an intention-to-treat analysis, where randomised pupils are included in the analysis regardless of whether or not they experienced the intervention). Although this would have minimised bias, it would not be able to prevent dilution of effect due to some of the control participants having received the intervention. Independent, concealed allocation (by a third party) would have minimised any potential undermining of the randomisation *at* randomisation (although this has been known to occur in at least one EEF trial where independent randomisation was undertaken but the school overrode the decisions), but could not have prevented this from happening after the start of the implementation period. Siddiqui and colleagues make some sensible suggestions to mitigate this possibility, but aggregated trials remain susceptible to the potential for lack of design fidelity, due to lack of experience and expertise in understanding the importance of the features and components of design being adhered to throughout. In this respect, the use of aggregated school-run trials is developmental and requires further training for practitioners in the science of robust experimentation, combined with very close monitoring through the embedded process evaluation procedures.

Synthesis of the results of the impact and process evaluations is critical, in not only explaining the context, highlighting any deviations from design or intervention delivery, but also in making sense of the results in relation to each other: an issue which becomes critical when the results do not agree or appear to contradict each other. For example, great care needs to be exercised when a rigorously designed, conducted and reported RCT finds no evidence of effect or evidence of harm of an intervention, but the stakeholders (participants, developers of the intervention) believe it to be effective.

Overview

The recent history of experimental research in the USA, the UK and the Scandinavian countries as described by Hedges and Schauer, Dawson et al. and Pontoppidan is remarkable. Moving from either antipathy to RCTs or apathy about their use - or a combination of both - all three geographical areas have, in recent times, moved relatively quickly into education experimentation in a big way. This will mean that school children in the future will have greater exposure to curriculum strategies and pedagogies firmly based on rigorous evidence of impact. Evidence of impact is only half the story, however, as Connolly et al., Dawson et al., Boylan and Demack, and Siddiqui et al. all highlight in their papers. Rigorously designed, conducted and reported process evaluations embedded within RCTs provide a critical role in explaining the effects observed in the RCTs.

The transformation in research design across several countries would not have been possible without the involvement of teachers. Professional development (in initial and continuing education) has a role to play in the education of teachers as future participants in the theory and practice of the

design. A commonality in the development of RCTs in the USA, the UK and in Scandinavia is that it was not normally initiated by teachers themselves, although the recent development of the use of aggregated trials described by Siddiqui et al. suggests that this may change in the future. However, the inclusion of professional development in research design may help prevent problems, such as the conflation of aspects of evaluation with the intervention being evaluated. The most pertinent example of this has been the tendency for schools randomised to the intervention group to drop out of testing if they cease to deliver the programme, possibly due to a lack of understanding of the importance of 'intention-to-treat' analysis.

As indicated earlier, EEF has made great progress in funding research that aims to avert the replication crisis that has unfolded across many areas of science. Aspects of research cited by Ioannidis (2014) as critical, which UK education RCTs have already embraced include large-scale collaborative research, trial registration, standardisation of definitions and analyses, and improvement in study design standards. An area that has been partially embraced is reproducibility practice, since a statistical analysis plan as required by the EEF, for example, is no substitute for pre-published analysis code. Areas which still need to be accommodated include replication culture, more appropriate (usually more stringent) statistical thresholds and training of the scientific workforce. Indeed, the Institute of Education Sciences in the USA set a good example in this latter regard by their programme of funded summer courses in education RCT design for established researchers, established in 2005.

Despite all these many positive developments, experimentation in schools remains under threat. There is a risk that, in the UK, for example, due to the explosion in the funding of RCTs, evaluators may increasingly find it difficult to recruit schools to take part in trials through competing pressures on their time. Researchers are powerless without the support of teachers and ceasing research activity is an easy way to cut costs when school budgets are eroded. Another risk for the future has already played out in the USA during the 1980s and 1990s, as described by Hedges and Schauer. It is easy for politicians and civil servants to assume a null result implies a problematic research design and then to use this as an excuse to stop funding trials. A concerted effort by the funding community, researchers and teachers alike should help to mitigate such risk in future and preserve the education RCT for as long as people feel the need to use interventions to help children learn.

References

Colquhoun D. (2014) An investigation of the false discovery rate and the misinterpretation of *p*-values. *R. Soc. Open sci.* **1**: 140216. http://dx.doi.org/10.1098/rsos.140216

Ioannidis JPA (2005) Why Most Published Research Findings Are False. PLoS Med 2(8): e124. https://doi.org/10.1371/journal.pmed.0020124

Ioannidis JPA (2014) How to Make More Published Research True. PLoS Med 11(10): e1001747. https://doi.org/10.1371/journal.pmed.1001747

Torgerson, C.J., Wiggins, A., Torgerson, D.J., *et al.* (2013) *Every Child Counts*: Testing policy effectiveness using a RCT, *Journal of Research in Mathematics Education*, 15 (2) 141-153

Ben Styles*, National Foundation for Educational Research, UK and
Carole Torgerson, Durham University, UK

*corresponding author: b.styles@nfer.ac.uk