Pre print version of:
Higgins, S. (2017) Room in the Toolbox? The place of randomised controlled trials in educational research, in A. Childs & I. Menter (Eds) *Mobilising Teacher Researchers: Challenging Educational Inequality* London: Routledge

Please check final published version in case of changes

5. Room in the toolbox? The place of Randomised Controlled Trials in educational research

Steve Higgins, Durham University

Introduction

This chapter considers the role of randomised trials in education as a necessary but not sufficient research design for drawing conclusions about effective educational practice. Using a toolbox metaphor, it identifies what kinds of questions are appropriate for different RCT designs in terms of supporting causal inference as part of a wider set of tools for educational inquiry. Trials have one key feature, randomisation, which uniquely addresses some aspects of potential bias in evaluative educational research. A review of the designs of the trials used in the Closing the Gap project helps to identify some strengths and weaknesses of randomisation, particularly in relation to the internal and external validity of the findings. A contrast will then be drawn between the large scale and micro-scale designs in 'Closing the Gap: test and learn' in response to Biesta's (2007) challenges about the democratic deficit in notions of 'what works' which he claims restricts opportunities for participation in educational decision making. By contrast it is argued that causal evidence is necessary, but not sufficient, for the normative professional judgments of teachers, and that teachers can be closely involved in the development and use of randomised trials.

Biesta (2007) critiques the idea of evidence-based practice and in particular the ways in which it has been promoted and implemented in education. He draws attention to a number of issues, in particular focusing on the dynamics between scientific and democratic control over educational practice and research and identifying the 'scientific' with a 'technological model of professional action'. This is not a necessary connection however, and I argue the dichotomy is a false one. It is possible to hold a 'scientific' view of causation at the same time as seeing education as a process of symbolically mediated interaction. Of course, the interpretation of the meaning of the findings of 'scientific' trials in terms of the impact of an intervention may be different as a result of this interpretation. A synthesis of these viewpoints puts greater emphasis on the internal validity of the trial findings in terms of answering the question "Was the intervention effective *there*, in terms of the outcomes measured in the trial?". The subsequent question of "And will it be effective in *my* school, for *my* pupils?" (in terms of external validity) requires, in my view, either extensive replication to understand the range of contexts where it can be successful or professional judgement and interpretation based on the limited inference from a single trial. This perspective is supported by a more rigorous understanding of what an 'average treatment effect' means in scientific terms (Deaton & Cartwright, 2016) and what can be inferred from the findings, in terms of the average impact in relation to the schools, teachers and pupils in

the trial, and how similar these pupils, teachers and schools are to the context under consideration for application in terms of 'evidence-use'.

Biesta (2007) examines three key assumptions of evidence-based education: first, the extent to which educational practice can be compared to the practice of medicine, second, the role of knowledge in professional action, particularly in terms of what kind of knowledge assumptions are appropriate for and relevant to professional practices that can be informed by research outcomes; and third, the expectations about the practical role of research implicit in the idea of evidence-based education. Perhaps unsurprisingly, I disagree with Biesta in some important respects on each of these issues, but most importantly his view that scientific knowledge diminishes democratic control over education and the decision-making of practitioners. By contrast, I argue that access to and engagement with 'scientific' knowledge is an essential condition for the democratic participation of teachers in making judgements about educational practice.

Biesta draws the conclusion that the notion of evidence-based practice is a limiting concept which not only restricts the scope of decision making to questions about effectiveness, but also that it restricts the opportunities for participation in educational decision making. He argues that we must expand our views about the interrelations among research, policy, and practice to keep in view education as a thoroughly moral and political practice that requires continuous democratic contestation and deliberation. On this point we agree, though I would go further and argue that the role of evidence is more crucial for practice here than for policy. Or rather that the role of evidence-based policy should be to support evidence-based (or more precisely evidence-informed) practice due to the variation in findings across educational trials and the challenge of interpreting average treatment effects from both single trials, as well as the pooled averages from research syntheses such as those found in meta-analysis.

These are serious claims against the use of rigorous inquiry and evidence in education that require further analysis in terms of whether these consequences are necessary and inevitable, or, if not, what can be done to mitigate the challenges of democratic participation in research by teachers and whether the sacrifice of causal inference could therefore be justified. The importance of causal claims is what I turn to in the next section.

Evaluation of impact in education

Impact evaluation in education usually means assessing the effects of policies and initiatives or other approaches to bring about intentional change in terms of valued educational outcomes for learners[1]. The aim of such research is to identify the impact so as to provide a retrospective assessment of whether the policy, intervention or approach was actually responsible for any changes in outcomes for learners (Higgins, 2017). The aims of the initiative will therefore determine the main questions for the evaluation (Rossi, Lipsey &

[1] Impact evaluation may, of course, also include the effects of change on educational systems or on the perceptions of those involved, rather than outcomes for learners.

Freeman, 2003). These are usually causal questions as policy makers, practitioners and researchers want to know whether the initiative has actually been responsible for any improvement.

Impact evaluation therefore tends to be summative rather than formative, in that the aim is to identify the effects of what has happened, rather than improve the effectiveness of a policy or intervention for the future. A key concept in any assessment of effectiveness or evaluation of impact or is therefore understanding the nature of any comparison being made, or the 'counterfactual' condition. We would ideally like to know what would have happened to pupils' learning both with and without the initiative taking place. This is not possible, of course, as a single student cannot both experience and not experience an initiative. We can't run a parallel worlds experiment in real life. So, different kinds of comparisons provide evidence for a stronger or weaker argument about the robustness of any causal claim in terms of whether an initiative has had an effect or not. The nature of the particular counterfactual or comparison in an impact evaluation affects what is a plausible explanation and a reasonable interpretation of the findings. More specifically, it affects the internal validity of the evaluation claims: what is the evidence that it has actually *worked*? Each of the approaches to impact evaluation in Table 1 (below, adapted from Higgins, 2017) seek to understand whether an initiative has achieved its aims or not. The strength of the claim weakens as the comparison is less capable of providing evidence that the change being evaluated is the actual cause of any improvement. The counterfactual comparison becomes less convincing the greater the threats to internal validity.

Table 1: Counterfactual comparisons and threats to internal validity in evaluative research designs

	Design	Counterfactual	Internal validity
Experimental	Randomised controlled trial (RCT)	Comparison of average outcomes from random allocated groups of students who are equivalent and either do or do not experience the change.	Provides a counterfactual which can infer causation. Controls for selection or allocation bias, regression to the mean effects and temporal effects; controls for both known and unknown characteristics which may influence learning outcomes (the majority of the time with a sufficient sized sample), except for the play of chance. Can control for the effects of innovation or novelty with an appropriate design (e.g. three arm trial with "business as usual" and "placebo" comparison). Provides a population average treatment effect (when the sample is randomly sampled from the population of interest and is sufficiently large).
	Regression discontinuity design (RDD)	Statistical model of average outcomes just above the cut-off in relation to the outcomes from all students, where students can be randomised around the cut-off.	Controls for selection and maturation effects by modelling the pre-post relationship at the cut-off point. This cut-off point must not be manipulable (i.e. the cut-off is arbitrary on all but the cut-off scale). Does not control for effects of innovation or novelty. Assumes pre-post relationship can be accurately modelled. Provides a local average treatment effect (i.e. inference may be limited to those around the cut-off point).
	Quasi-experimental design (QED)	Comparison of average outcomes from allocated groups of students who are non-	Provides a limited counterfactual which can infer limited causation. Does not control for selection or allocation bias, regression to the mean effects and temporal effects; does not control for and any

		equivalent and either do or do not experience the change.	unknown characteristics which may influence learning outcomes. Does not control for effects of innovation (unless more than one intervention condition is included). Provides a sample average treatment effect.
Observational	Natural experiments Matched comparison groups Difference in difference (regression)	Outcomes from similar students who do not experience the change.	Does not control for selection or allocation bias that is related to unobserved or unmatched characteristics. Groups must be sufficiently similar for analysis (matching). Does not control for effects of innovation.
	Time-series (e.g. single group design)	Outcomes from the same students, a number of times before and after a change (usually a minimum of three occasions).	Does not control for selection or allocation, other external change, or maturation and growth. Can provide limited causal inference if input and output variables correlate strongly (e.g. use of a particular approach in some time periods but not others).

Randomisation aims to take account of both known and unknown factors which may account for differences in groups, as opposed to matching, which controls for known factors (such as age, gender, socio-economic background, special educational attendance). Randomisation therefore aims to take account of aspects of the complexity of a context which may not be known in advance. Experimentation is a deliberate inquiry which makes intentional change and aims to identify the effects of that change. A further goal may be to identify and test a specific causal model or to validate *how* the change has been effective or which students benefited most. Approaches such as theory-based evaluation seek to do this by having a clear conceptualization or logic model which attempts to explain how the policy or intervention produces the desired effects (Fitz-Gibbon & Morris, 1996). In this approach factors or features of the theoretical model are included in the evaluation design so that any association can also be explored. This might include aspects of fidelity (tracking how faithfully those involved adopted the changes in practice) or measures which might indicate changes in participants' behaviours or the processes of the new practices being evaluated which are consistent with the theory and which would therefore be expected to be clearly correlated with successful outcomes. If 'evidence' is taken here to mean 'causal evidence of impact', then it seems clear that such evidence is necessary for decision-making in education. We need to know whether some things have been successful or not: whether they 'worked' as intended. In education, there are many factors which make this difficult to assess. I argue that this makes such designs to test causal claims more important, not less; though it is also important to understand the limit of the warrant of these claims. In the next section I turn to why it is particularly important in education to attempt to identify cause and effect.

Understanding what is happening in classrooms

One of the challenges in identifying and understanding learning classrooms is the complexity involved in the interactions between learners and the teacher (and other adults) in relation to intended educational aims and outcomes. Doyle (1977) characterised this in terms of "multidimensionality, simultaneity, and unpredictability" (p 52). He also described a range of the strategies which teachers use to deal with this complexity in terms of:

1. Chunking (the ability to group discrete events into larger units);
2. Differentiation (the ability to discriminate among units in terms of their immediate and long-term significance);
3. Overlap (the ability to handle two or more events at once);
4. Timing (the ability to monitor and control the duration of events); and
5. Rapid judgment (the ability to interpret events with a minimum of delay).

These strategies are all necessary to manage and cope with the complexity and hecticness of classrooms, but make it difficult to determine some aspects of cause and effect. This is because, as they are inducted into the profession, the ecological nature of teachers' learning (or their 'coping strategies': Pollard, 1982) means that their observations and experiences are filtered and interpreted often at an unconscious level or in relation to the immediate needs at hand. A novice teacher often struggles with behaviour and classroom control, and order can become an end in itself, rather than a means to an end (such as better learning). This makes teaching susceptible to a number of human biases in terms of interpreting and managing this complexity and highlights the challenge of validly and accurately identifying cause and effect. These inevitable and understandable biases also map onto aspects of research design and the approaches we can take to critically examine our understanding of cause and effect in classrooms. Whether we like it or not, we all form 'personal theories' (Cole, 1990) about how teaching relates to learning. They are one of the main reasons that I argue that studies with strong causal warrant are sometimes necessary, but never sufficient, in educational research. It is all too easy to interpret practices and processes which reinforce our existing beliefs about cause and effect in the classroom and which bolster our personal theories without sometimes checking that they are actually achieving what we believe.

A number of natural biases make it difficult for any individual to judge the accuracy of their perceptions when identifying cause and effect in classrooms. Many of these can be dealt with through systematic data collection and analysis, but some require further control for possible bias or misinterpretation (see Table 2).

Table 2: Some biases affecting causal claims

Natural Bias	**Description**	**Example**	**Design control**

Anchoring bias	A tendency to rely on, or "anchor", one piece of information in making decisions (often the first piece of information acquired on that subject).	Noticing particular pupils' response to an intervention (e.g. boys and technology) and attributing increased engagement and outcomes for some, as evidence of success for all.	Systematic data collection of relevant data. Effective comparison group (or counterfactual comparison).
Confirmation bias	Likelihood of finding, remembering or interpreting information so as to confirms existing beliefs or hypotheses, and/or finding less salience in alternative possibilities.	Seeing some pupils respond positively to a 'learning styles' intervention and interpreting this as evidence of the effectiveness of 'learning styles' rather than an increased range of teaching strategies being used, and/or with greater choice and responsibility for learning taken by pupils.	Systematic collection and analysis of relevant data.
Innovation bias	Tendency to favour change and see the positives (similar to confirmation bias): c.f. *status quo* bias.	Introduction of a digital technology which is successful, but which takes more teacher-time and is more expensive (i.e. is less efficient).	Systematic analysis of relevant data. Effective comparison group (or counterfactual comparison).
Maturation bias	Most pupils improve over time.	Pupils' reading improves after the introduction of reciprocal questioning, but hard to determine the extent to which the new approach was responsible.	Effective comparison group (or counterfactual comparison).
Selection bias	Pupils are chosen in relation to an expectation about how they are likely to perform.	Pupils are 'triaged' by schools according to how close they are to the C/D grade or Level 3 to 4 borderline, or more challenging pupils are rejected from a support programme.	Randomisation or independent allocation to control for known and unknown factors.

A 'scientific' approach attempts to control for other possible explanations for improvement. It aims to create a fair test of the claim that an intervention or approach is successful. Most potential biases can be addressed through rigorous and systematic data collection and analysis. Most require more than this. First, so as to provide a strong case for causal validity, an effective comparison needs to be made (the 'counterfactual' condition, as discussed above). Second, the groups need to be equivalent in terms of both known and unknown factors which might explain any improvement. This can partly be achieved through matching pupils or creating equivalent groups that are as similar as possible in terms of the factors which might explain any differences in outcomes (e.g. current level of attainment, age, sex, free-school meal or special educational needs and disability status). The advantage of randomisation is that, on average, it controls for both known and unknown differences. These measures, do not, of course guarantee that research adopting these principles will

always be more accurate, but rather that, assuming that the imposition of the research design framework does not change the context in a way which alters the causal conditions, they will provide a more robust and accurate answer to a causal question.

The Design of the Macro-trials

The aim of the design of the large-scale school trials in Closing the Gap: test and learn was to conduct evaluative research of interest to schools which addressed just such causal questions, but to devolve much of the responsibility for the management of the trial to the schools themselves. Accordingly, the design team (see Chapter 4 for more details about the design and rationale for the macro-trials) identified a long-list of potential approaches then consulted school teachers and leaders about interventions and current gaps that were priorities for them through online surveys and focus groups. A shortlist of 17 interventions suitable for trialling was selected on the basis of:

- the evidence of promise from research;
- the availability of suitable outcome measures and suitability for testing impact as a distinct or discrete approach;
- the manageability within the project's timescales and resource levels, and likely demands on participating schools; and
- the appeal to schools based on criteria identified by them (to increase likelihood of take-up and relevance of findings).

At this stage the aim was to end up with seven pools of schools of roughly similar size allocated to these interventions as groups for randomisation. A smaller pool was identified in terms of sample size for one intervention because of the higher cost and the greater demands made on schools and because of the intention to test this approach over two years.

Participating teaching schools and their partner schools were sent details of the seven interventions and asked to rank them according to their preferences and to identify any which did not meet their current needs or which were not suitable for their context. This was so that the approaches were not tested in conditions which were not appropriate or where they would not be selected by schools. The key aim was to identify a match of schools to possible interventions for subsequent randomisation, which in turn aimed to minimise selection bias as far as possible within the constraints of the project so as to make the comparison a fair test.

The final design phase aimed to help establish the conditions for successful implementation for research and involved:

1) establishing (sometimes, negotiating) with the intervention providers the detail of their provision so that the intervention was replicable;
2) producing broad descriptions of each intervention;
3) devising protocols for managing the interventions in for schools to use and to improve the comparability of implementation;
4) providing advice and guidance to the College's Closing the Gap team on other features of the programme particularly the selection, design and logistics of testing, and the management of randomisation;

5) documenting the process and creating guide resources for programme managers.

This step was to ensure that what was tested in each context was replicable, both across the schools in the project, but also for subsequent adoption (should the approaches be shown to be successful). A wait-list design was used so that schools allocated to control groups in the first year could carry out the intervention in the second year (should it be successful). This was for ethical reasons and for equity in terms of the resources and support offered to the schools involved. A minimum of 40 schools was needed in each group to ensure adequate power for analysis so as to increase the likelihood of a conclusive result (whether positive or negative), together with the use of the same assessments which met the requirements for validity and reliability across the interventions.

A number of strategies were put in place to reduce the risk of drop-out and attrition, including providing schools with opportunities to make their own choices about the intervention groups they were allocated to and particular target groups of pupils (within the overall design and randomisation constraints) as well as making the benefits of and commitments to being a control school clear. Pupil groups were selected according to a protocol based on vulnerability (such as FSM, Looked after, Special Educational Needs and Disability) and low achievement in specific areas (such as literacy or mathematics). Within this framework schools could select target pupils themselves, so the team developed a Pupil Identification Tool to structure the process and base it on criteria from the analysis of appropriate test and descriptive data to ensure common processes across schools and to reduce the risk of re-introducing selection bias. The research design therefore tested use of the approach in schools. This is an effectiveness question in that it aimed to answer the question: is this approach effective in schools in typical conditions? An alternative would have been to try to answer an efficacy question: is this an effective intervention? However, this would have meant more rather strictly controlling the protocols for use to ensure consistent processes (see Table 3 below).

One of the interventions, Research Lesson Study, was distinctive as it required development to enable trialling for Closing the Gap. It was therefore offered as a pre/post test pilot with a group of 20 schools in the first year, and, subject to a successful pilot, as a full trial in the second year.

It is important to be clear about what kind of question a trial is answering as this affects the design and interpretation of the findings. The model the Education Endowment Foundation (2016) uses is shown in Table 3.

Table 3: Trial Stage

Trial stage	**Description**	**Inference**
Pilot study	Conducted in a limited range of schools (e.g. three or more) where an intervention is at an early or exploratory stage of development. More fine-grained data used to develop and refine the approach and test its feasibility in schools. Initial indicative data collected to assess its potential to improve outcomes.	Is/ *is not* feasible and has/ *does not have* indications of promise

Efficacy trial	Aims to see whether an intervention can work under ideal or developer-led conditions across a range of settings (e.g. ten or more) schools. Has an impact evaluation to identify effect on attainment and a process evaluation to identify the elements of effective practice.	Has been effective/ *has not been* effective under *ideal* conditions.
Effectiveness trial	To test the whether an intervention can work at scale in a large number (e.g. 40 or more) schools, where the developers are no longer the only deliverers. Has an impact evaluation to assess the effect on attainment and a process evaluation to identify the challenges and solutions to roll-out. The cost of the intervention at scale will also be calculated.	Has been effective/ *has not been effective* under *typical* conditions at scale.

For most interventions, standardised tests for either literacy or numeracy were appropriate progress measures as these mapped directly onto the curriculum and were the focus of the intervention. Where interventions were cross-curricular, the team recommended literacy and numeracy tests. It was, however, recognised that some interventions would have additional outcomes, and that schools and researchers would need to collect other evidence about complementary outcomes. This aimed to acknowledge the importance of a wider range of outcomes from education, but also the constraints under which schools work and the value of success in the current assessment system for individual pupils.

The team recommended that each series of pre-tests in all intervention and control schools for a particular intervention needed to be completed within a four-week window and that pre-testing should occur before randomisation and target pupil identification and any training as the first stage of the interventions, so as to avoid allocation bias. Post-testing was recommended at the end of the academic year as it would make the interventions easier to compare; would allow a focus on sustained or longer term benefits; and be easier to manage logistically. It was also recommended that pre- and post-testing should be conducted on the trial participants who were in the control/wait-list groups (and not just the 'intervention groups') to provide an effective comparison group or counterfactual.

These measures were all put in place to provide as fair a test as possible of interventions and approaches that schools would be likely to adopt and to let them run as closely as possible to the way schools would manage them were they not in a trial, so as to see if there was any overall average benefit for these interventions, compared with what schools normally did (the counterfactual).

The Design of the Micro-trials

One of the goals of the Closing the Gap: test and learn project was the development of understanding of educational research methods among the research leads from the 206 participating lead schools. A key strand in this was the 'Early Adopter of Teacher-Led RCTs' programme (or 'Early Adopter' programme: see Chapter 6). This involved inviting proposals for small-scale experimental research studies (teacher-led randomised controlled trials or

quasi-experimental 'micro-trials'). Additional training and support was provided to schools that were successful, and there were no requirements regarding the focus and content of the studies or the nature of the measures used, other than they had to be of professional interest to the proposer and amenable to experimental inquiry.

The teachers were all encouraged to choose an area of their practice which they wanted to improve based on a hunch or hypothesis about what might be successful and then to design a study to test whether or not this improved outcomes for pupils as rigorously as was practical in a school setting. Most of the teachers were familiar with some contemporary approaches which aim to provide quantitative estimates of effect (such as Hattie's (2008)) 'Visible Learning' or the Sutton Trust-Education Endowment Foundation's 'Teaching and Learning Toolkit' (EEF, 2017). The teachers designed the experiments, undertook them, usually collecting pre- and post-test data, analysed the data (with help from the CfBT Education Trust team (now Education Development Trust): see Chapter 6). Support included identifying the advantages and disadvantages of between-subject versus within-subject designs, choosing and designing tests to ensure validity and reliability, and the benefits of pre- and post-test designs and when to use them. They also wrote up their research studies in a poster format and presented findings at a conference, echoing Stenhouse's (1981) notion of "systematic and sustained inquiry, planned and self-critical, which is subjected to public criticism and to empirical tests where these are appropriate" (p. 113), but differing from more usual action research-based approaches (which many of the schools were already involved in: see Chapter 1) by using small scale experimental trials which sought to control for possible allocation and maturation bias (see Table 2 above) as well as other possible threats to internal validity.

The range of areas of inquiry and approaches to evaluate these varied considerably across the micro-trials (see Chapter 6 for more details), but the research design was tailored to the inquiry question, so as to produce as robust an answer to the question (usually an impact question), given the constraints of one or two teachers pursing the investigation within their own professional context. There is a long history of classroom investigations and teacher self-study (see, for example, Loughran, 2004) though rarely including small scale trials with randomisation (for some exceptions see Coe, Fitz-Gibbon & Tymms, 2004 and Gorard, Siddiqui & See, 2016). In health, such approaches are now advocated to help researchers and practitioners understand whether interventions are having intended effects, when and for whom they are effective, and what factors moderate an intervention's effect, so as to develop of more effective 'just-in-time' adaptive interventions (Klasnja *et al*., 2015). Similar potential has been recognised in educational settings.

'Scientific' knowledge and the democratic deficit

The 'Closing the Gap: test and learn project shows that experimental trials with randomisation which involve schools and teachers in selecting the focus for experimental inquiry and in managing and conducting the process of the trials themselves are both feasible and acceptable in schools in England. This in itself is sufficient to counter Biesta's (2007) claim that 'scientific' approaches *necessarily* create a democratic deficit in

educational research. As Churches, Hall and Brookes argue in Chapter 2, the programme overall has shown that schools and teachers have the capacity to engage in research through both large-scale multi-arm trials and small-scale experimental studies: both macro- and micro- trials. This involvement appears to have increased interest in and discussion of research and research findings. Contrary to some of the beliefs expressed at start of the programme, schools were not resistant to the use of control groups, or to the use of quantitative approaches and in statistical analysis or to the use of randomisation as a method to improve the internal validity of research in schools. Importantly, the project has also shown that teachers and schools can take a more active role in the delivery of randomised controlled trials, as the Education Endowment Foundation has also discovered with their 'aggregated trials' model (e.g. Siddiqui, Gorard & See, 2015; Gorard, Siddiqui & See, 2016).

However, it must also be acknowledged that this approach is not a panacea for education research. Testing approaches using trials methodology requires a well-formulated question of the form "on average, does intervention or approach X improve Y outcomes for pupils, when compared with Z (usually either 'business as usual' or an active comparison of an alternative intervention or approach)?". The design and analysis of trials is not without its challenges (Xiao, Kasim & Higgins, 2015). Whilst these kinds of questions are undoubtedly important for the profession (and for policy decisions), not all important educational questions can be formulated in this way, or are amenable to causal experimental inquiry of this kind. I see the different types of educational research methods as a toolbox which needs to be matched to a particular educational inquiry question. Randomised experimental trials have a particular function and are best suited to questions of causal impact – was approach X responsible for effect Y? They are like a chisel which has been designed and developed for working wood, but, as a tool with a particular function, chisels are not useful for hammering nails, sawing wood to size or bolting or fixing items with screws. Similarly, randomised experimental trials are not sufficient to identify or understand the complex causal processes which lie behind effects. They help with identifying the 'what' in causal investigations but not the 'how' or the 'why'. However, to understand the 'how' and the 'why' also presupposes that the 'what' actually works, so I'd argue in all causal inquiries they have a role.

Evidence from trials cannot and should not determine what ought to be taught, but once the content of the curriculum is agreed and the broad pedagogical values of a school or system have some consensus, then evidence about the relative effectiveness of different approaches to achieve these goals are essential in informing the decisions educators need to make in the best interests of their pupils. Not all educational inquiries are causal or about effectiveness, however. Biesta is justified in his critique that education is also about values and that the aims of education and the nature of the curriculum are also essential areas for discussion, identification and clarification, perhaps rather more that the current policy discourse allows. I would set the goals broader here too in terms of the range of tools we need in the educational research toolbox, drawing on other disciplines and research traditions in psychology, sociology, history and economics for example, for developmental studies, cohort studies, capturing the lived experiences of teachers and learners, these are all important traditions with methods matched to the focus of a specific research question.

'Scientific' knowledge about causes and effects in education is an essential tool for the professional educator, however. It is necessary, but not sufficient for professional decision-making. Not to be open to evidence from research with strong causal inference is problematic as it implies professionals are limited to opinion and judgement, with only limited knowledge about the effectiveness of what they do in relation to specific ends (e.g. reading or proficiency in aspects of mathematics). If teachers are to understand the effects of what they do, then engaging in and with educational research which strong causal inference is a necessary part of developing as an effective teacher. The benefits outweigh the costs. In terms of Biesta's (2007) "democratic deficit", I am perhaps more concerned at the issues of power which lie behind control over aspects of the education system, in terms of pedagogy, curriculum and assessment. If researchers and practitioners collaborate to develop understanding of cause and effect in education, both at the level of classrooms and at the level of schools, this may help redress the political balance of power between policy and practice. The issue for me is therefore not whether randomised experimental trials are possible or desirable in education, but rather when this particular research tool provides the best answer to an educational question and then where and how to use the technique to best effect.

References

Biesta, G. (2007). Why "what works" won't work: Evidence-based practice and the democratic deficit in educational research. *Educational Theory*, 57(1), 1-22.

Cole, A. L. (1990). Personal theories of teaching: Development in the formative years. *Alberta Journal of Educational Research*, 36(3), 203-222.

Coe, R. Fitz-Gibbon, C. and Tymms, P (2000) *Promoting Evidence-Based Education: The Role of Practitioners* Round table presented at the British Educational Research Association Conference, Cardiff University, 7-10 September 2000. Available at: http://www.leeds.ac.uk/educol/documents/00001592.htm

Deaton, A., & Cartwright, N. (2016). *Understanding and misunderstanding randomized controlled trials* (Working paper no. 22595). Cambridge MA.: National Bureau of Economic Research.

Doyle, W. (1977). Learning the classroom environment: an ecological analysis. *Journal of Teacher Education*, 28(6), 51-5.

Education Endowment Foundation (2016) *EEF evaluation: A cumulative approach* London: EEF. Available at: https://v1.educationendowmentfoundation.org.uk/uploads/pdf/EEF_evaluation_approach_for_website.pdf

Education Endowment Foundation (2017) *The Sutton Trust-Education Endowment Foundation Teaching and Learning Toolkit* London: EEF. Available at: https://educationendowmentfoundation.org.uk/resources/teaching-learning-toolkit

Fitz-Gibbon, C. T., & Morris, L. L. (1996). Theory-based evaluation. *Evaluation Practice*, 17(2), 177-184.

Gorard, S., Siddiqui, N. and 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.

Hattie, J. (2008). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement.* London: Routledge.

Higgins, S. (2017) Impact evaluation: a case study of the introduction of interactive whiteboards in schools in the UK in R. Coe, J. Arthur, M. Waring & L. V. Hedges (Eds) *Research Methods & Methodologies in Education* (2nd Edition) London: Sage.

Loughran, J. J. (2004). A history and context of self-study of teaching and teacher education practices. In J. J. Loughran, M. L. Hamilton, V.K. LaBoskey & T. Russell (Eds) *International Handbook of Self-study of Teaching and Teacher Education Practices* (Volume 12) (pp. 3-7). Dordrecht: Kluwer.

Klasnja, P., Hekler, E. B., Shiffman, S., Boruvka, A., Almirall, D., Tewari, A., & Murphy, S. A. (2015). Micro-randomized trials: An experimental design for developing just-in-time adaptive interventions. *Health Psychology*, 34(S), 1220.

Rossi, P.H., Lipsey, M.W. and Freeman, H.E. (2003) *Evaluation: A Systematic Approach.* Thousand Oaks, CA: Sage.

Siddiqui, N., Gorard, S., & See, B. H. (2015). Accelerated Reader as a literacy catch-up intervention during primary to secondary school transition phase. *Educational Review*, 68(2), 139-154.

Stenhouse, L. (1981). What counts as research? *British Journal of Educational Studies*, 29(2), 103-114.

Xiao Z., Kasim, A., Higgins, S.E. (2016) Same Difference? Understanding Variation in the Estimation of Effect Sizes from Educational Trials *International Journal of Educational Research* 77: 1-14.