The evidential basis of ‘Evidence Based Education’: An introduction to the special issue

Adrian Simpson
School of Education
Durham University

In the last 15 years or more, there has been a significant shift towards what has been called “evidence-based education” (EBE). While the phrase may seem benign – after all, who would want to base practice on anything other than evidence – proponents often appear to accept only particular meanings for the term ‘evidence’. A strict, fixed hierarchy of more or less acceptable study designs places randomized controlled trials (RCTs) and meta-analyses of these at the top as a ‘gold standard’.

Figure 1 illustrates the growth in the proportion of reports of RCTs and meta-analyses; their centrality to policy decisions is illustrated by the fact that the US “what works clearinghouse” for education research, which describes itself as a “resource to help you make evidence-based decisions”, only endorses RCT designs. In the UK context, Coe (2004) equates ‘evidence-based’ with RCTs and their derivatives. This appears to support Sampson’s (2010) argument that ‘evidence’ is often taken exclusively to mean the results of experimental designs.

Of course, many have a more nuanced view: Davies (1999), often cited as one of the papers most influential in the establishment of ‘evidence-based education’, explains at length the need to tie research method to research question. That is, there are no fixed hierarchies of evidence. Instead Davies argues that evaluations comparing interventions on well-defined outcomes may be best undertaken with RCTs and quasi-experimental designs, but that questions about processes, meanings and consequences will need other methods, such as surveys, observations, ethnographies and interviews.

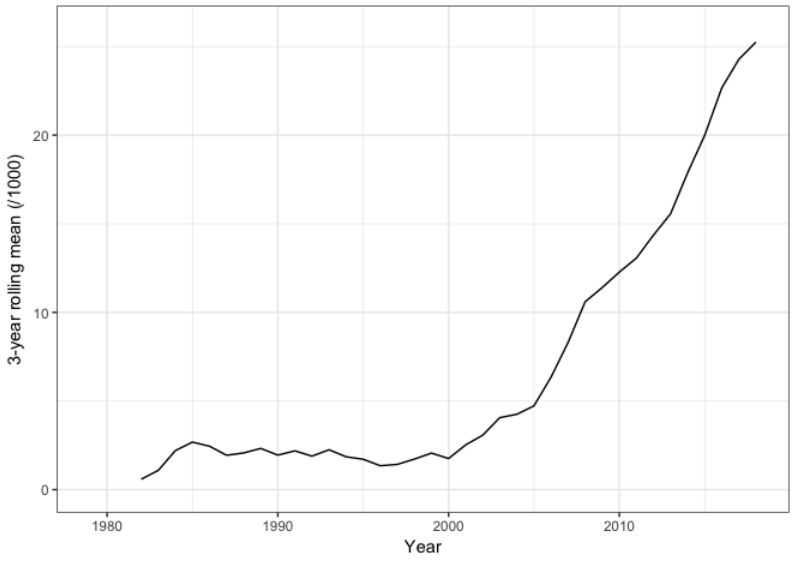

Figure 1: Growth of RCTs and meta-analysis[1]

[1] The number of reports listed on the ERIC database with the keywords ‘school’ and at least one of the words ‘randomised’, ‘randomized’ or ‘meta-analysis’ as a proportion of reports with the keyword ‘school’.

However, governments are increasingly encouraging or requiring policy makers and schools to use the more restricted notion of 'evidence' to direct decision making. As part of the "No Child Left Behind" Act, the US congress mandated the use of a restricted set of research designs for most education research grant schemes; in the UK, government encourages the use of EEF evaluation tools (grounded in RCT design principles) for schools accessing over £2bn of funding to support the learning of disadvantaged pupils. An examination of the funding recently provided by the main sponsors of education research in the UK suggests around 90% of mainstream resource is directed at RCTs.

In this context, it is important to critically explore the evidential basis of 'evidence-based education'.

There have been critical perspectives on EBE approaches before. For example, Biesta (2007) argued that the focus on particular forms of evidence narrows policy making down almost exclusively to questions about effectiveness; Eacott (2017) argued that there has been a lack of criticality, particularly of key works and key figures in the EBE movement; and Shahjahan (2011) argued that the EBE movement can be placed within a particular colonial discourse.

A key argument has been that the often cited analogy with medicine is fundamentally flawed (e.g. Hammersley, 2005) and, moreover, that within medicine this rigid notion of a fixed 'gold standard' research method is increasingly questioned. While provided certain strong assumptions are met, RCTs and meta-analysis may have a role in deciding if one intervention is better than another on reducing stomach ulcers, when one looks at producing public policy guidelines and working with individual patients one needs to account for the complexity of social situations. A stomach ulcer does not have consciousness and agency to decide how to respond to an intervention, while different patients do have that freedom. One patient's ulcer's response to proton pump inhibitors is not impacted by another's cholesterol level's response to statins, but one patient's experience of being treated by a particular doctor may have a profound effect on another patient's decision to seek or respond to treatment advice, even from another doctor in that practice. When we move from treating the body part to treating the patient, evidence from even well designed RCTs may not be enough. Complex systems in which conscious agents interact and which co-evolve with other systems are not susceptible to randomised controlled trials or their derivatives and the UK's Medical Research Council is now recognising the difficulties of using methods designed for simple, non-interacting situations for complex social ones (see Greenhalgh & Papoutsi, 2018). In education, there are few obvious analogies of body parts: all the components in education have agency, consciousness and are part of a complex, networked social system.

While accepting many of these existing critiques, this special issue has a slightly different focus: it provides a number of different perspectives on the extent to which the narrow notion of 'evidence' in 'evidence-based education' is a basis for educational policy making; the role it plays in policy arguments and how proponents of the current EBE approach respond to their critics.

The first paper in the special issue gives an overview of the concerns. Phillips argues that there are two camps: those who support evidence-based education (the 'tough-minded') and those that are sceptical about it (the 'tender-minded'). To this latter camp, he attributes a belief – which he counters – that educational processes are not suitable for causal investigation of the form espoused by those in EBE. Within the former camp he distinguishes those who see randomised controlled trials (or, in his terms, randomised field trials) as

directly generating policy from those who see it as one potential part of a more complex policy argument. With tongue only slightly in cheek, we might see these latter people as critics of EBPACPIE ('evidence based policy as currently practiced in education'): those who can see a place for talking of causes and effects, but see only a partial role for the results of RCTs in policy arguments, rejecting the current vogue for direct application of 'what worked' evaluations to classroom practice.

Within this special issue, the reader will see papers which fit Phillips's notion of the 'softer' side of the tough-minded and perhaps those Phillips might assign to the tender-minded camp.

Across this special issue as a whole there is a recognition that, even in the tough-minded camp, the argument from research to policy is a chain with many links. Kvernbekk's paper outlines a possible structure of such arguments in terms of Toulmin's model in which the result of research studies play a subsidiary role. Testing the validity of an argument means testing each of these links, not focussing only on the internal validity of the RCT.

This issue brings together papers which test the strength of EBE policy argument chains at different levels. At the level of particular links, papers explore the links forged from individual RCTs, links joined together by effect sizes and in meta-analysis and links which join the RCT sample to a different population. Other papers look at the chain as a whole, how the whole argument is structured and why certain forms of EBE argument have such apparent political power

Standing back to look at the chain as a whole, Kvernbekk explores the way in which practitioners might reason about the evidence generated from 'evidence-based education' (particularly from RCTs). She takes care to focus her attention on only one particular purpose for education (qualification) to which she argues RCT evidence is most obviously applicable – though readers will find other authors in this issue who disagree that EBE's 'gold standard' method has much value even for determining what might work for practitioners even here. Exploring different models, she concludes that a particular form of Toulmin's argumentation scheme may have the best qualities to make it the most practitioner friendly, provided they understand where different pieces of knowledge and evidence (including RCT evidence) sit within the argument. Despite coming down in favour of one model, the analysis of each will give readers a clear summary of the key ways in which evidence plays a role in the 'practitioner tale' which might be built on each model, the questions each can answer and the practicality of each for the practitioner making decisions about teaching approaches.

Joyce's and Cartwright's papers separately tackle the issue of how results from RCTs can become of use to practitioners. Even the rare educational RCT which might meet all of the necessary assumptions for drawing a causal conclusion draws it only for the sample studied. Joyce and Cartwright both contend that transporting that conclusion to another situation almost always requires forms of argument beyond the RCT.

Joyce tackles this key issue by focusing on the representativeness of the sample and setting. Provided some fairly stringent and rarely met assumptions hold, we can conclude from a well conducted RCT that some intervention led, on average, to better performance on some measure than some other treatment for some sample of participants. However, for a practitioner to argue that this provides grounds for believing this intervention will lead to better performance requires an appeal to the extent to which the study situation is representative of the practitioner's. While Joyce notes that some EBE advocates acknowledge

the importance of context in some circumstances, it tends to be downplayed when they make claims for evidence of general effectiveness. Some evidence aggregators do enable practitioners to identify studies which took place in contexts which are in some ways similar to the practitioners' own classrooms. Joyce argues, however, that unless we have other ways to identify what factors are causal or support an intervention's effectiveness, there is little justification for concluding the study context is representative in the right ways and therefore that the intervention has a good chance of working for the practitioner.

It seems inconsistent that those who insist that observational and quasi-experimental methods are inadequate because they fail to construct groups matched on unknown causal factors, don't appear to recognise that the samples in field RCTs similarly fail to match the populations to which results may be applied. If two groups being 'similar enough' is not sufficiently strong for one link in an argument chain, surely it is not sufficiently strong for any other link.

Towards the end of her article, Joyce argues that a solution may be found in understanding how an intervention works – what mechanisms are in play, in what contexts, which result in the intervention leading to the desired outcomes. RCTs, in the form usually met in EBE, do not provide this and Joyce argues that only by combining RCTs with other research methods can we develop the knowledge of underlying mechanisms that can be useful for practitioners.

Cartwright's article complements Joyce's by examining EBE's focus on 'rigour' within an RCT at the apparent expense of its usefulness. While advocates of EBE appear to argue that they are informing the educational community about interventions which will be effective, their naïve view of the role rigour plays in their argument means they may well be misinforming the community. By focusing carefully on what characterizes RCTs, Cartwright distinguishes what can be inferred *about* the study and what can be inferred *from* the study (or multiple studies) about future practice. The excessive emphasis placed on RCTs undermines the credibility both of the argument for effectiveness and that the claims derived are useful in practice. To draw a conclusion about the effectiveness of an intervention for students not in the original study requires chains of argument beyond the RCT and forms of argument which normally will not be rigorous (in the EBE sense). Moreover, even to draw a conclusion *about* the study from the RCT design requires assumptions, many of which will not be credible in real educational settings: controlling for unintended, post-randomisation differences between treatment groups is likely to be impossible and arguments for their unimportance take us outside the supposed bounds of the rigour claimed for the method.

However, all is not lost. At the heart of her article Cartwright has practical suggestions for how teachers can make decisions about practice. They mean giving up the certainty promised by EBE's 'rigour', but Cartwright shows that any claimed certainty was illusory anyway. The article proposes visualising the kind of policy arguments she has in mind as 'bird's nest' of interrelated ideas, results, analyses and reasoning.

One obstacle for this suggestion, however is detailed in Cowen's paper: bureaucracies can't deal with bird's nest arguments. Cowen takes a different perspective on EBE, explaining why its apparent certainty and its basis in simple measures may be particularly attractive to policy makers. Introducing the reader to a model of how bureaucracies work, Cowen explains how the current approach to evidence-based policy in education might be well suited to modern bureaucratic systems. By purportedly being 'scientific' (which Wrigley & McCusker's paper disputes) the apparently simple messages from RCTs can travel around the bureaucracy with

less transmission error or manipulation. That is, Cowen is arguing that EBE's attractiveness is precisely its lack of nuance, network of provisos or complex theory explaining how something works. Moreover, it shifts the onus away from a need to make major structural changes at the top of a bureaucratic hierarchy towards changes which schools and teachers might implement *within* existing structures.

Among other things, Simpson's paper explores how key players in that bureaucracy react to the identification of errors in their argument. Noting that one of the core measures in the EBE decision making process, effect size, is not fit for purpose, Simpson highlights how some of the defences thrown up in response follow a familiar pattern: listing assumptions which are not checked (but are often prima facie absurd), arguing that the conclusions from their flawed arguments should stand until they are proved wrong, or claiming that mere awareness of the flaws is sufficient for them to continue with the same arguments. At worst, they fall back on arguing that they know about the criticisms, so can now exclude them from further consideration.

The final main paper, from Wrigley & McCusker, takes the reader back to a wide view of the research-policy argument chain used by the EBE movement. They argue that EBE is merely adopting some of the language of science without adopting methods which would be appropriate for a scientific endeavour in the kinds of complex social situations common in education. They argue that the randomized controlled trial (RCT) methods which may play a role in determining the effectiveness of a fertilizer in agriculture or a drug in pharmacology are inappropriate for determining effective policy for classrooms. As noted above, others have argued equally cogently that the RCT as the canonical grounding for policy arguments may also be inadequate in medicine (Marchal et al., 2013): for example, even if we are convinced that a treatment was on average more effective than an alternative for a particular sample drawn from a particular population in a particular context, we need to know much more before we can draw a conclusion that the treatment will be suitable for another person (or even that it won't be actively harmful for that person). Moreover, Wrigley & McCusker point to the centrality of theory in science and its near absence in EBE. They note the crucial role of understanding 'why', at least in the sense of positing a mechanism through which a policy change might affect an outcome. Most often ignored by the EBE practitioners, at best it is sidelined to a 'logic model' which seems to never be re-examined should an RCT indicate the suggested mechanism might be ineffective in the trialled context. By exploring one particular EBE summary (sports participation in the Educational Endowment Foundation's 'teaching and learning toolkit'), Wrigley & McCusker note how the mechanisms by which interventions may be effective and the contexts in which those mechanisms work are hidden from practitioners, even though (as Joyce's paper also notes) these are more likely to be valuable guides to what might work in practitioners' own contexts than unwarranted averages of statistical parameters.

Finally, Dylan Wiliam reflects back on the issue as whole, drawing the papers' arguments, together with his extensive experience in educational research, to consider the implications for the field. He brings out some examples of mistaken inferences which have resulted from a naïve view of RCT evidence in both class size reduction and ability grouping. He concludes, boldly but (as the remainder of the issue demonstrates) justifiably, "the entire project of evidence-based education can never be successful".

That, of course, is not to ignore the value which might come from combining careful quantitative studies (including RCTs when appropriate) with other forms of evidence which

help us understand the mechanisms of teaching and learning and how they can be influenced. It involves understanding that there is not one fixed hierarchy of evidence when it comes to policy decisions. Transporting results from one context to another cannot involve simply implementing the same intervention elsewhere, but should involve understanding why (and in what contexts) different approaches may lead to different outcomes – it is the mechanism which is transportable, not the intervention.

To be a critical consumer of research is not a matter of keeping up with a growing list of 'what works' statements (which are, anyway, at best statements of 'what worked, compared to something else, on average, for a particular outcome, for some people in one context'). Nor is it a matter of knowing the current rank ordering of interventions (which are, anyway, ranked on the basis of a mistaken understanding). Wiliam argues, instead, that it is to have a broad understanding of alternative mechanisms in the context of a deep sense of our classrooms and our schools.

Acknowledgements

It is appropriate to end this introduction with some acknowledgements. The editors of *Educational Research and Evaluation* were kind enough to encourage me to develop this special issue and provided support throughout. The reviewers were uniformly excellent: normally responding quickly, but never at the expense of extensive and insightful engagement with the papers. They were critical in the most positive sense and helpful to both the authors and to me in improving the clarity of the papers and the issue as a whole. Finally, thanks to the main paper authors and the reflections author, Dylan Wiliam: each has engaged with the theme thoughtfully, bringing a different perspective to the evidential basis of 'evidence based education' which has strengthened my understanding of the issues and I hope will add to the understanding of each reader of this special issue. To all, my heartfelt gratitude.

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.

Coe, Rob. 2004. What kind of evidence does government need? *Evaluation & Research in Education 18*(1-2): 1-11.

Davies, P. (1999). What is evidence-based education?. *British Journal of Educational Studies*, *47*(2), 108-121.

Eacott, S. (2017). School leadership and the cult of the guru: the neo-Taylorism of Hattie. *School Leadership & Management*, *37*(4), 413-426.

Greenhalgh, T., & Papoutsi, C. (2018). Studying complexity in health services research: desperately seeking an overdue paradigm shift. *BMC Medicine*, 16(1), 4-9.

Hammersley, M. (2005). Is the evidence-based practice movement doing more good than harm? Reflections on Iain Chalmers' case for research-based policy making and practice. *Evidence & policy: a journal of research, debate and practice*, 1(1), 85-100.
Chicago

Marchal, B., Westhorp, G., Wong, G., Van Belle, S., Greenhalgh, T., Kegels, G., & Pawson, R. (2013). Realist RCTs of complex interventions–an oxymoron. *Social Science & Medicine*, *94*, 124-128.

Sampson, Robert. 2010. Gold standard myths: Observations on the experimental turn in quantitative criminology. *Journal of Quantitative Criminology 26*(4): 489-500.

Shahjahan, R. A. (2011). Decolonizing the evidence-based education and policy movement: Revealing the colonial vestiges in educational policy, research, and neoliberal reform. *Journal of Education Policy*, *26*(2), 181-206.