For whom does 'what works' work? The political economy of evidence-based education

Nick Cowen, New York University School of Law

CENTRE FOR HUMANITIES

CHESS

ENGAGING SCIENCE AND SOCIETY

For whom does 'what works' work? The political economy of evidence-based education

Nick Cowen

Nick Cowen[1]
New York University School of Law
40 Washington Sq. South
New York
NY 10012
nick.cowen <at> nyu.edu

Abstract

What role does scientific evidence play in educational practice? Supporters of evidence-based education (EBE) see it as a powerful way of improving the quality of public services which is readily applicable to the education sector. Academic scholarship, however, points out important limits to this applicability. I offer an account inspired by Tullock's theory of bureaucracy that helps explain EBE's influence despite these limits. Recent configurations of EBE are an imperfect solution to two imperatives where policymakers are at an informational disadvantage: (i) guiding professionals working in the field and (ii) evaluating evidence from academic researchers. EBE, especially in the form of RCTs and systematic reviews, offers a way of filtering a complex range of research to produce a determinate result that is transparent to policymakers. However, this impression of research transparency is misleading as it omits theoretical background that is critical for successfully interpreting the results of particular interventions. This comes at a cost of relevance to the frontline professionals whom this research evidence is supposed to inform and help.

Introduction

In April 2018 Schools Minister for England Nick Gibb provoked a social media cascade when he tweeted his consternation at a poll which revealed that most teachers still believe that individual students have unique 'learning styles'. This is despite such a theory being bereft of research evidence (Newton & Miah, 2017; Pashler, McDaniel, Rohrer, & Bjork, 2008). Among the critical responses, some pointed out that Gibb (2017) accepted the evidence-based refutation of learning styles while ignoring compelling evidence that selective grammar schools (still supported by his Conservative government) were socially divisive while achieving no increased educational attainment (Gorard & Siddiqui, 2018). The claim was that Gibb was being hypocritical and selective in his deployment of research evidence, using it only to criticise teachers while ignoring demonstrated societal harms caused by the government's policy.

While this *tu quoque* response to Gibb is not a persuasive defence of learning style theory, this controversy demonstrates an underlying concern with evidence-based education (EBE). Being selective about one's use of evidence is easy enough. We all do that and benefit from being called out on it too. But increasingly policymakers do more than determine what evidence to acknowledge and which to ignore. They now develop policies defining what sort of research evidence is worth producing in the first place. What I propose here is that there are features of EBE that make it attractive, or at least uncontroversial and safe, to policymakers. This is because it avoids casting judgement on structural features of the education system that policymakers, rather than teachers, are in a position to change. At the same time, the attractiveness of this kind of research for policymakers comes at a cost of relevance to teaching professionals.

My argument proceeds as follows. I begin by describing evidence-based education's relationship with evidence-based policy (EBP) in other domains. I explain how the strength of evidence drawn from EBP is strictly limited in the absence of theories establishing the causal mechanisms through which experimentally tested interventions achieve their outcomes. These limits apply to education research as much as other domains. Using the emergence of EBE in the US and the UK as a point of departure, I then introduce a general explanation for policymakers' attraction to EBP and willingness to overlook its weaknesses: its relative compatibility with existing bureaucratic accountability mechanisms. Having set out this general account, I indicate what a more productive conception of EBE could look like in terms of content and scope if the strictures of explicit government involvement and oversight could be loosened.

The limits of the evidence-based policy paradigm

What is evidence-based education (EBE)? Such a label can represent a broad range of approaches (Davies, 1999; cf. Lucas, 2017). As commonly understood and as promulgated by its proponents, however, EBE refers centrally to the development of school policies and classroom practices based on systematic reviews of experimental research evidence (Boaz, Ashby, Young, & ESRC, 2002; Slavin, 2002; Young, Ashby, Boaz, & Grayson, 2002). The paradigm experimental research design is the randomized controlled trial (RCT). The ideal systematic review is a meta-analysis of such trials, or those with similar research designs, that provide a simplified estimate of a treatment effect.

The principal inspiration for EBE is the evidence-based medicine movement, although the experimental paradigm of EBP has also advanced within the field of behavioural economics (Behavioural Insights Team, 2011; Halpern, 2016; Thaler & Sunstein, 2009) and international development (Banerjee & Duflo, 2012; Duflo, Glennerster, & Kremer, 2007; Reddy, 2012). Key scholars associated with the origins of EBE include Hattie (2009) and Marzano (2005). Experimental research designs have made important contributions to education research in the

past (Inglis & Foster, 2018). What makes this recent revival of experimental research unique is that it is advanced primarily on the grounds that it is useful for establishing appropriate policies and for testing specific interventions (Eacott, 2017). In addition, the impetus comes not only from educational academics but also from policy researchers who see education as one of several sectors that are ripe for reform using what they see as a superior method.

In terms of organizations, the most prominent are the What Works Clearing House[2] in the United States and the Education Endowment Foundation[3] (EEF) in the United Kingdom. The EEF is part of the wider What Works Network in the UK, a group of government-backed research centres that aim to apply experimental evidence to a widening range of public policy areas (Alexander & Letwin, 2013; Goldacre, 2013). The most prominent source of EBE in the UK aimed at teachers and school leaders is the EEF's 'Teaching and Learning Toolkit' (Higgins et al., 2016) that summarizes and compares various approaches along the dimensions of cost, effectiveness and certainty of evidence.

Many proponents present this enterprise of bringing these research techniques to bear in new policy areas as being straightforwardly scientific, an advance on other approaches to public policy that rely on evidence that lack appropriate rigor. The classic slogan is that RCTs are the 'gold standard' because it is the only approach that can definitively prove a causal relationship between an intervention and an outcome (Sanders & Halpern, 2014; Coe, 2004; Goldacre, 2013; cf. Cartwright, 2007).

There is some truth to this characterization but its implications are more limited than its proponents tend to suggest. A successfully conducted RCT provides an unbiased estimate of the impact of a particular intervention on individuals, in one place, at one time, for a given study population (Deaton & Cartwright, 2018; Joyce & Cartwright, 2018).[4] All other approaches require imputing a causal relation between intervention and outcome based on a theory or model that attempts to exclude other possible causes.

How useful is this feature of an RCT likely to be in policy terms? Paradoxically, precisely what makes RCTs attractive within context reveals their limited usefulness outside of it. An RCT makes sense when you want to isolate the effect of the intervention in circumstances where you cannot control for other factors. In a controlled laboratory environment, randomization is unnecessary as you can observe and manipulate all the features of an experiment until you know precisely how each factor affects the outcome. In the world outside of the laboratory, however, interventions are made where there are limitless varying factors (and interactions between factors) that could change the outcome. So it seems to make sense to use a research design that keeps these factors the same while varying the intervention. That produces a compelling result (almost a guaranteed causal inference if a statistically significant difference between treatment and control is found). However, the researchers still do not know whether that causal effect will survive a change in

---

[2] https://ies.ed.gov/ncee/wwc/
[3] https://educationendowmentfoundation.org.uk/
[4] This is assuming the experiment is conducted in a way that does not introduce confounding variables, which itself is far from trivial. Unless a treatment is triple-blind, placebo controlled and based on an intention-to-treat analysis, which is often very hard to build into a social policy intervention, it is almost inevitable that some confounding factors will have been introduced. Deciding that it is safe to ignore such factors itself requires theoretical assumptions. In addition, an unbiased estimate can still turn out to be inaccurate, especially if it relies on a small sample. Under ordinary tests of statistical significance, we expect 1 in 20 unbiased estimates to reject the null hypothesis erroneously, and these errors can survive in the literature if such results are subject to publication bias even if all individual studies are unbiased.

environment (Cartwright, 2013). To make such claims, they need a broader account of how the intervention worked, not just to demonstrate that it worked. These claims cannot be established through experimental testing (Cartwright & Munro, 2010).

This concern for establishing what proponents of EBP call 'external validity' is what prompts the next step in the EBP paradigm: meta-analysis. The idea here is to test the same intervention in multiple but supposedly comparable cases then statistically synthesize the results. If the intervention works reliably to produce an effect within some acceptable level of variance, then supposedly you can be more confident that the effect survives variation in circumstances that people implementing the intervention are likely to encounter.

But can meta-analysis solve the problem of generalising the effect of an intervention? Again, not without a theory of how the intervention works and, therefore, an understanding of the kind of environment within which can it can be suitably applied. Without theoretical guidance, the possible combination of contributing factors is infinite. The population of interest is potentially open-ended and ill-defined. On the other hand, the population that can be practically sampled even for the largest meta-analyses is likely to be limited in ways that re-introduce the dreaded bias that RCTs were meant to remove. What sort of populations and environments are accessible for randomized testing? Often peculiar ones with the institutional and physical infrastructure that allows researchers to work *in situ* over a period of time. Access to study sites can depend a great deal on personal relationships and trust between researchers and potential participants. The study population will therefore have more of the properties of a convenience sample than a random sample. You can apply randomized treatments within the possible study population but that is very unlikely to be the case for the whole population of interest (Fortin, 2006).

Common challenges with implementing EBP are often characterized as problems in ensuring the scalability or fidelity of interventions that are already known to work based on supposedly secure evidence. They may be better characterized as part of the underlying problem already identified: even a systematic review or meta-analysis of many high-quality RCTs provides little epistemic warrant for applying to a new population. It is hard to know when to trust the results of EBP out-of-sample (Deaton & Cartwright, 2018). When implementation disappoints, we do not know whether it is because the implementation was flawed or because the intervention's capacity to work has broken down in the new context. In other words, the combination of meta-analysis and RCTs do not solve the fallacy of reasoning from induction.

A plausible objection to these concerns is that recognizing these problems is all well and good, but there are no viable alternatives unless we want to dispense with an empirical approach to policymaking altogether. However, once the limited value of RCTs in terms of practically establishing causal relations for real-world application is understood it then becomes equally justified to choose research whose strengths are along different methodological dimensions (Cartwright & Hardie, 2012; Deaton & Cartwright, 2018; Donmoyer & Galloway, 2010). This is especially true for studies that can sample a general population of interest more effectively.

Quantitative research designs, including surveys and observational studies may permit causal inferences when combined with other background knowledge. Regression analyses that take observational data and model causal relations between independent and dependent variables, while controlling statistically for other variables, may indicate plausible causal effects. Studies based on quasi-experiments, natural experiments, instrumental variable estimation, synthetic controls, differences in differences designs may all provide more useful estimates of the impact of

public policies than RCTs and meta-analyses depending on the domain under examination. On the qualitative side, in-depth surveys, participant observation, interviews, case studies, event studies, process tracing and analytic narratives, may all contribute to establishing the likely effects of a policy approach. Indeed, very often, it will be a combination of some of these approaches and forms of evidence that will credibly identify causal mechanisms. While the assumptions within these research designs may turn out to be more complex and sometimes less accessible to non-experts, they may ultimately offer more accurate accounts of the results of policy and practitioner decisions.

The bureaucratic knowledge problem

I have identified some limits to EBP's experimental paradigm and its underlying reliance on other aspects of scientific knowledge to make plausible claims. Given these limits, what has attracted the extension of EBP to policymakers, especially in the domain of education? In the United States EBE rose to prominence through President George W. Bush's No Child Left Behind Act (Cochran-Smith & Lytle, 2006; Eisenhart & Towne, 2003; Thomas, 2012). NCLB introduced new federal funding for school districts aimed at improving the performance of disadvantaged students. It was accompanied by various new accountability frameworks including a requirement that schools in receipt of funds use scientific methods to evaluate their use of government resources. In the United Kingdom, EBE's emergence was linked to the 2010-2015 Coalition Government's policy of introducing a pupil premium aimed similarly at raising educational attainment amongst disadvantaged pupils. Critically, the pupil premium is not intrinsically linked to EBE as such (Marshall, Rabindrakumar, & Wilkins, 2007). The pupil premium was originally conceived as a more transparent way of ensuring additional funding reached schools with disadvantaged students. But as is almost inevitably the case when resources are expanded under contemporary public finance regimes, more spending implies a greater need for accountability. Major and Higgins (2015, p. 16) position the Education Endowment Foundation's work as a solution to the 'high autonomy high accountability' regime that schools face, where school leaders are expected to achieve improved outcomes but are not told explicitly by policymakers how to pursue that objective.

How does EBE contribute to policymaking objectives? Applying Tullock's (2005) model of bureaucracy to this question can help outline a theory of what EBE's function might be. Tullock is a major contributor to public choice, the analysis of collective decision-making from a methodologically individualist standpoint (Mueller, 1976; Ostrom & Ostrom, 1971). It is an attempt to explain collective outcomes from the choices individuals make given the constraints and opportunities they face. Several scholars have applied public choice analysis to the behaviour of public sector organizations. My reason for using Tullock's model as a point of departure is its compatibility with a wide range of agent motivations. Niskanen (1968, 1994) controversially models bureaucratic firms as budget-maximizers in a way that parallels wealth-maximizing agents in private markets. Dunleavy (2002) models individual civil servants as at least partly self-interested and career driven. In contrast to these approaches, my central concern is not the problem of individual or group interests diverging from the stated aims of a public service. Although public choice approaches are sometimes associated with rigid self-interest assumptions (Mueller, 1976), my account here does not posit self-interest or opportunism as a core problem that bureaucracies face. Instead, I am interested in a broader problem that separate individuals face when attempting to cooperate even in pursuit of a common good (Buchanan, 1999). This is the epistemic challenge of getting people with limited knowledge and bounded rationality to coordinate their activities consistently across time and space using fallible communicative and

cognitive capacities to produce collectively beneficial outcomes (Ostrom, 1993; cf. Lucas, 2017; Meadowcroft, 2005; Pennington, 2011; Cowen, 2017).

The eyes of the sovereign

This is the problem that Tullock identifies for bureaucracies. In his simple model, bureaucracies are essentially composed of a pyramid structure: what he calls the sovereign, a generic (classically absolute) ruler of an organisation, and her subordinates at various levels in a hierarchy. He outlines what he calls a 'standard' model of bureaucracy as follows:

> The lower levels of the structure receive information from various sources. This information is then passed along upward through the pyramid. At the various levels, the information is analyzed, collated, and coordinated with other information that originates in separate parts of the pyramid. Eventually, the information reaches the top level where the basic policy decisions are made concerning the appropriate actions to be taken. These decisions are then passed down through the pyramid with each lower level making the administrative decisions that are required to implement the policies sent from on high (Tullock, 2005, p. 149)

Tullock acknowledges that this standard model is not considered descriptive by social scientists. It is, nevertheless, a sort of paradigm of how a bureaucracy *ought* to work, at least as conceptualized by ordinary citizens who hold the sovereign accountable for the actions of government agencies. In Britain, this somewhat stylized fiction of ultimate sovereign control over the bureaucracy is recognized in the notion of collective ministerial responsibility for all departmental decisions (Flinders, 2000; Palmer, 1995). For education in England and Wales, the best candidate entity to think of as the sovereign in practice is probably the Prime Minister and her Cabinet of Ministers, with the Secretary of State for Education and Minister of State for Schools making most decisions with the implicit assent of their Cabinet colleagues. In the United States, legal sovereignty is deliberately dispersed and checked between the states and the branches of the federal government. Nevertheless, so far as the federal bureaucracy is concerned, the sovereign is the President of the United States with the Secretary of State for Education acting as their first subordinate in the area of national education. The challenge that the sovereign faces is a kind of principal-agent problem, which are ubiquitous throughout hierarchical organisations (Bergman & Lane, 1990; Eisenhardt, 1989). There are equivalent challenges for headteachers with respect to school staff and council officers with respect to local authority-controlled schools. These challenges tend to more substantial the larger and more complex the organisation the sovereign or principal is attempting to manage.

Tullock identifies a problem that prevents sovereign control from obtaining in practice. Whenever information travels up the pyramid or instructions are sent down the pyramid, some information is inevitably lost through errors in interpretation. In addition, at each stage, time and energy constraints mean that some information is deliberately omitted. The sovereign cannot gain a synoptic grasp of all the relevant information nor provide perfectly detailed instructions for all her subordinates (McCaughey & Bruning, 2010). As a result, Tullock suggests that the key objective for the sovereign cannot be to supervise the output of the whole structure but to delegate powers and resources in such a way that subordinates take decisions that the sovereign would ideally hypothetically take herself if she had the capacity.

How does the sovereign attempt to make it so that her will is carried out? Tullock suggests some solutions that are familiar to anyone with experience of contemporary public sector management frameworks including in education (cf. Meadowcroft, 2003). He notes the strictly limited role of

charismatic leadership, a commonly discussed managerial technique for motivating subordinates. Indeed, the obvious limits of personal charisma is precisely why organizational hierarches have to be established. More promising is the splitting up of particular tasks through formal rules and limited delegated powers. This constrains subordinates from acting with discretion that deviates from the sovereign's will. Random inspections can help gain a sample observation of the typical conduct and output of subordinates; this can act as a feasible substitute for constant oversight. In addition, schemes such as 'judgement by results' (2005, p. 205) are supposed to establish substantive measurable outcomes (or targets) for the subordinates to achieve. Results-based accountability is then presumed to provide a greater degree of autonomy than a framework based only on procedural rules.

These approaches also have well-known trade-offs. Rigid rules constrain discretion that might be abused but could also make subordinates unable to cope with unanticipated scenarios. Thus they can produce decisions that the sovereign did not intend. Senior managers may have a bias towards expecting perfection and so are disappointed when inspections turn up weaknesses in procedures that they imagined would be implemented with absolute fidelity. It is often not clear what level of fidelity to explicit procedures is reasonable to expect. This is one reason why pre-announced inspections are sometimes used *in lieu* of random inspection as they are easier both for inspectors and subordinates to control, even if they do not measure typical activity as accurately. Meanwhile, judging by formal outcome measures alone can encourage administrators to engage in practices (for example, 'gaming' the system) that impact negatively on the underlying objective that the sovereign wills (Campbell, 1979; Goodhart, 1981). Of course, it is also possible for a less benign sovereign to prefer a system to be gamed if perceptions of effective provision are more important to them than the underlying reality.

Evidence-based policy as an additional strategy for sovereign delegation

With this conception of the sovereign's challenge in mind, we can understand attempts to advance EBP as a more advanced mechanism for efficient delegation. In the case of complex public policy areas, the sovereign struggles not just with the problem of ensuring subordinates carry out the ends of the organization. She may not even be aware of how those ends might be best achieved in principle.
One solution is to consult recognized experts in the relevant field of public policy, perhaps even to appoint them to policymaking positions. In the latter case, the sovereign delegates not only implementation but also the policymaking decisions to subordinates. Rather than deciding the policy to be pursued, the sovereign creates a specific subordinate position tasked with establishing the policy and a communication channel between the delegated policymaker and the implementers. With this channel in place, a policymaker need not be formally a superior of the implementer. Policymakers and implementers could be on the same rung, hierarchically incommensurable rungs of a bureaucratic structure, or even in a separate agency. Nevertheless, the will of the sovereign is that implementers follow the guidelines set by the designated expert policymakers.

The sovereign faces further problems when attempting to delegate to experts:

1. The experts often disagree with each other about the effectiveness of different policy proposals and the sovereign may be incapable of evaluating which expert is correct or of appointing the correct one.

2. The advice offered by the experts may be difficult for the sovereign's subordinates to interpret and impossible to implement in practice.

3. The experts may have objectives and motivations other than passively advising the sovereign on how best to achieve her goals (Pritchett, 2002). They may have value commitments to a policy goal different from which they are being explicitly asked to achieve.

A key concern is that by the time information travels up through a bureaucracy to the sovereign (or, at least, her senior subordinates) and back down into the administration as instructions for implementation, it will necessarily have lost much of its accuracy, nuances and caveats. It may, in addition, be distorted along its path by people with different motivations and perspectives from the sovereign.

Given these challenges to expert delegation, we can now clarify one of the attractions of RCTs and systematic reviews as a preferred form of EBP. Both impose a set of priority rules on scientific evidence that are, or are at least supposed to be, resistant to manipulation from those generating and presenting the evidence. An RCT, when properly conducted, involves specifying preferred outcome measures prior to testing a treatment or intervention. Randomization prevents the researcher from pre-selecting specific individuals to receive the intervention or selecting alternative outcomes after the study has been completed. This means that a favourable or negative outcome of the study is not pre-determined by those directly carrying it out. Similarly, a systematic review should specify a search protocol prior to undertaking the study and analysis. In neither case is the researcher able to guarantee data that will point in a particular direction in advance of carrying out the study. In other words, this is a procedure for a sovereign wary of potential bias amongst her own experts and advisers.

However, these procedures are not, in fact, immune to researcher bias (Every-Palmer & Howick, 2014; Ioannidis, 2017). Whether a particular treatment is found to 'work' may depend on discretionary factors. Scientific researchers can select hypotheses that they consider worth testing. For example arguably one of the reasons that Cognitive Behavioural Therapy gained support in mental health policy is that its proponents were among the first psychological therapists to adopt experimental trial methods (Hofmann, Asnaani, Vonk, Sawyer, & Fang, 2012). Researchers can use preferential controls (such as 'wait list' compared to 'treatment as usual') in order to increase the apparent impact of an intervention. If they are particularly keen on an intervention turning out to have positive effects, then they may be able to specify multiple outcome measures in the hope that at least one of them will be favourable. Similarly, if experts have some idea of the shape of a scientific literature before undertaking a systematic review, they may be able to specify a search protocol that manages to include studies that are favourable while excluding those that are unfavourable.

Nevertheless, the need to state criteria and aims at the outset makes researchers' interpretation of the evidence somewhat more transparent. Gough et al. (2013, p. 8) note the advantage of systematic reviews compared to reliance on more intuitive approaches as follows:

> Existing research can… be reviewed by academic experts, either individually or in groups ('expert panels'). Experts have many specialist skills; but relying on expert opinion can be risky if that opinion is not supplemented by a systematic review of the research. The

authority or reputation of experts may obscure their ideological and theoretical assumptions, the boundaries of their knowledge (and consistency of depth of knowledge within those boundaries), and possible flaws in their methods of synthesizing knowledge.

The advantage of formal criteria for the sovereign is that she does not have to explore the details of research evidence, pass judgement on specific interventions, or evaluate the approach and conduct of individual experts. Instead, she requires that policymakers discover 'what works' according to EBP criteria and that the bureaucracy disseminates the policies that pass through this filter. The problem is that this same criteria does not necessarily align with practical usefulness in the field. In order to be successful in a public sector setting with bureaucratic oversight, the process must at once furnish guidance to policy implementers while also producing observable feedback that is both transparent and pleasing to the sovereign.

Approaches that appear to work from the perspective of field practitioners but whose outcomes and procedures are opaque ('lacking transparency') or difficult to measure formally will likely be excluded. Thus, we might expect field practitioners to defend their preferred practices and even resist EBP when it runs against their own experience. Meanwhile, both bureaucratic subordinates and field professionals have to remain focused on the sovereign and her formal accountability mechanisms for fear of penalty. They may well aim to use research evidence, even on their own initiative, but continue to prioritize commands emerging from elsewhere in the bureaucracy for which they are more directly accountable.

One question, to which I can only offer tentative hypotheses limited to England, is 'why now?'. Sovereigns always have to contend with managing subordinates and contending with experts. So this raises the questions why exactly EBE in this form has emerged this decade. Part of the answer is that the scale of a nationalized and centralized public sector, and the resulting demands on government managers, encourages more refined and abstract forms of delegation. While experimental research has always been an important part of scientific inquiry, it was Cochrane (1972) who initiated the case for putting evidence based on randomization at the forefront of policy development. The context in which he made the case was an observed mismatch between resource inputs and associated outputs in the National Health Service, at that point still a new but vast public sector organisation. Evidence-based medicine promises increased efficiencies that are supposed to work at scale.

To a somewhat lesser extent than the NHS in the middle of the 20th Century, the education system in England and Wales has recently undergone a series of centralizing reforms that have taken powers away from local authorities and given them to the Department for Education along with some central government agencies, and several Academy Trusts that report directly to the DfE (Lawn, 2013; Whitty, 2008). These centralizing reforms that have taken place over a series of governments were prompted by new theories of governance that put more effective and equitable education at the core of attempts at broader economic and social reform (Crafts, 1996). Policymakers want to get more out of the education system than before. This has generated more challenges of management and responsibility for ministers and this has made them receptive to new forms of governance.

The attraction of novelty also offers some explanation. New governments tend to displace incumbents by pointing out their failed policies so there is a strong imperative to introduce innovative reforms that make a decisive break with previous administrations. Prime Minister David Cameron was attracted to the new science of behavioural economics which promised the

possibility of producing better public policy outcomes without spending additional resources, particularly useful during a time of austerity. He established the Behavioural Insights Team as a Downing Street policy unit. Behavioural economics, reflecting its origins in psychology, makes substantial use of experimental research designs so the methods and attitudes of its practitioners resonates with the experimental EBP paradigm. The advantage that experimental EBP has over other paradigms is that it is relatively simple and abstract. This allows its practitioners to move quite rapidly between policy areas without as much (apparent) need for domain specific knowledge. What began as a narrow policy approach focused on behavioural changes, especially around personal health and finance, ended up easily extending to a range of public sector reforms where the role of experimental evidence could conceivably to contribute. It was this initial impetus that eventually led to the emergence and expansion of the What Works Network.

Implications for evidence-based education

The previous section offers an account of how EBP in general fits into government accountability frameworks that, I believe, resonates with the emergence of EBE in the US and UK. EBE is an example of academic research playing a role within a top-heavy bureaucratic infrastructure that prioritises accountability and transparency to remote decision-makers. This can end up distorting research evidence, and how it is presented, received and interpreted. The question I turn to now is what this account means for identifying the weakness and a tentative approach to reforms of EBE itself.

In essence, the current EBE framework produces results and conclusions that are not as helpful as they could be for schools and teachers. These problems are recognized within the sector. Major and Higgins (2015) acknowledge that some EBE research from the EEF, particularly its emphasis on feedback, has ended up justifying additional pressures on teachers from inspectors that are probably not ultimately productive. Nevertheless, they emphasize a continuing need 'to ensure schools use their pupil premium effectively and avoid shallow compliance' (2015, p. 18). This implies a continuing link between EBE and the systems of accountability that oversee teachers and schools.

What does this emphasis on compliance mean in practice? What it amounts to is a strong preference for quantitative measures of discrete school-level and classroom-level interventions. This is illustrated by the design of the current major output in the UK, the EEF's Teaching and Learning Toolkit (Higgins et al., 2016). It summarizes and compares the expected effects of various interventions that have been tested experimentally. As Simpson (2017) argues, these comparisons, based on widely different measurements of varied populations, are likely to be deeply flawed. But the Toolkit is 'useful' insofar as it provides a series of legitimized approaches that school leaders can try to implement with the resources dedicated to disadvantaged students. It allows for something with some evidence behind it to be implemented along with a fairly rapid construction of a rational narrative that can be fed back into administrative and inspection regimes.

On my account EBE could become substantially more useful if this link with administrative accountability was severed, or at least loosened, along with the reliance on experimental evidence. This would have several advantages. EBE would be less associated with the spurious precision that comes with its current quantitative form of presentation. Instead EBE could be used to pin down the broader principles of what constitutes effective practice. Without being tied to RCTs, researchers could observe the context, background and sheer variation of existing teaching practices without seeking to generate an exogenous shift in practice for the sake of the clarity of a

research design. This would permit a more detailed understanding of what 'business as usual' is like in contemporary school settings and what explains the prevalence of existing practice. This would help with understanding the factors that help make or break a proposed change in practice, including the role of policymakers in the process. Perhaps most importantly, a new EBE approach would allow the questions pursued by researchers to reflect more directly the practical needs of teachers as they see it in their circumstances. Sometimes those questions will be best answered by experimental evidence. Very often it will be through some other approach.

What I propose is hardly alien to the education research community as a whole, although, it has a precarious relationship with the contemporary practice of EBE. There are several practitioner-active academic researchers, as well as teachers who engage deeply with academic research, who make use of a range of scholarly evidence, including relevant experiments, to develop their approaches. Deans for Impact (2015) shows how theoretical principles derived from cognitive psychology can be applied to teaching practice. Christodolou (2016), drawing on similar foundational assumptions, takes on the issue of feedback in the context of the English school system. She deals with the puzzling failure of *Assessment for Learning*. AfL was a government program for rolling out feedback strategy based on strong evidence from a range of scholarly sources, including experimental evidence. It commanded strong support among policymakers and a great deal of the teaching profession. It was successfully implemented at least to the extent that teachers in England now provide a great deal more feedback than before, and more than teachers in other countries. Yet the theorized improved student outcomes did not materialise.

Christodolou uses a range of evidence to argue that part of the reason for this is a failure to differentiate formative and summative assessment, a related distortion of priorities driven by performance management and a more fundamental failure to understand the role of feedback in learning. She argues that feedback only works in a context where complex skills are broken down into simpler manageable tasks that, through practice, become part of a student's repertoire. Through this appraisal of a promising but failed approach, she is able to offer some practical evidence-informed steps that teachers can use to give feedback in an effective manner.

Ashman (2018) takes a similar theoretical approach but applies it more broadly to several English-speaking countries and to a wider array of issues. In addition to covering assessment, he discusses evidence-informed approaches to classroom management, explicit teaching, lesson planning as well as contemporary controversies surrounding new technology and phonics. Critically, Ashman places the research evidence about effective classroom approaches in the context of theoretical debates that have come to influence existing policy and practice. So rather than simply being a series of topics on which the evidence for particular interventions falls one way or the other, he offers some idea of how his account differs from previous government-supported practices. This historical orientation is practically significant because it is often unclear to teachers, from looking at something like the EEF Toolkit, how exactly the evidence-informed approach is expected to differ from their existing practice (Cowen, Virk, Mascarenhas-Keyes, & Cartwright, 2017).

Unlike experimentally constituted EBE, these approaches proceed on the premise of an underlying theory of how students learn. Such theories are controversial and might turn out to be wrong, or in need of substantial refinement. However, this enterprise has the benefit of working towards teaching approaches that are credibly generalizable while furnishing information for application in context. They are trying to establish not only what works, but why and under what conditions. It does not select or weight evidence on what amounts to arbitrary methodological grounds or avoid taking a theoretical stance altogether.

As Cowen *et al.* (2015; 2017) argue, this approach reflects how experienced teachers already interact with EBE research. Teachers recognize the limited applicability and explanatory features of EBE as it is currently constituted and use it to generate new ideas of their own that they think would be worth implementing. This is very different to faithful implementation of experimentally validated approaches which is how EBE supposedly works in theory. Teachers often apply EBE through the lens of theoretical frameworks that they have found to resonate with their own experience in the classroom. Since teaching professionals are using EBE research in these ways anyway, insofar as the overall objective is to help teachers and school leaders perform better, the research design and presentation should reflect the way research is used in practice.

The scope of evidence-based education research

So far, I have suggested how a reformulated EBE can better support teachers and schools. The other part of my critique suggests that the current EBE framework builds in an overly narrow scope of the sort of questions worth asking by education researchers.

The issue of scope takes us back to the initial charge of evidence selectivity levelled at Nick Gibb from his authoritative position as Schools Minister: his criticism of teachers ignoring evidence relevant for their practice combined with his own avoidance of evidence regarding more pervasive source of inequity in the education system. The claim that individual students do or do not have different learning styles relates to the way teachers and schools organize lesson formats. This claim supposedly can be tested experimentally and rejected as a potentially effective practice. The nature of experimental methods means an almost exclusive focus on interventions at the scale of schools and classrooms. By contrast, the claim that selective grammar schools help or harm students, although likely much more relevant to the issue of closing the gap between advantaged and disadvantaged students, is a structural question that reflects policy decisions rather than school leadership and teacher decisions. RCTs cannot test these sorts of questions (Krauss, 2018). Insofar as answering these questions is critical for the government's stated objective of increasing equity and fairness in education, then EBE as it currently stands displays a dearth of ambition.

How would reducing the role and priority of experimental evidence improve EBE? First, EBE could take into greater account the role of decisions and priorities of policymakers in determining educational attainment. The way the EBE framework is set up at the moment means that judgements about systemic resource allocation are almost ruled out of the discussion. The point of departure for EBE is that there are given resources available. In the UK, this includes the pupil premium aimed at reducing educational disadvantage. The task of EBE is to work out how to use those resources effectively. For example, Major and Higgins (2015, p. 16) describe one of their key messages as 'It's not what you spend, it's the way that you spend it… that's what gets results'.

This avoids an important conclusion of academic research in education: resources matter. More funding going into schools leads to predictably improved results (Holmlund, McNally, & Viarengo, 2010). Another way of looking at this is that schools already have much of the know-how to generate better educational attainment when they are given additional resources. This could be considered obvious and trivial. But recognizing that schools and teaching professionals can 'do more with more' is relevant as it implies a trade-off. Resources and effort devoted to funding central education agencies, or to setting up experiments throughout the school system, could instead be used by professionals on the ground to improve education outcomes with their pre-existing knowledge of what works in their particular contexts. This problem bites especially in circumstances where teaching professionals leaving the sector is a prevailing problem, and where

we have strong (though, naturally, non-experimental) evidence that teacher attrition is a significant barrier to student attainment (Borman & Dowling, 2008).

Second, stepping away from experimental methods would allow EBE to be more sensitive to critical factors outside classroom practice or the control of school. By external factors, I refer to concrete aspects of the local environment, including access to housing, transport and amenities that constitute the quality of life of students, their families and teachers. I also mean local resources and institutional features such as access to qualified teachers, the degree and kinds of autonomy that schools have, rules governing inclusion and exclusion of students, and the curriculum they are required to pursue. These are the sorts of topics that frequently concern teachers and school leaders, which is at least indicative that this could be where major gains in educational effectiveness could be found.

The importance of understanding external factors is illustrated by one of the largest recent reversals of fortune in the education sector: the sustained improved performance of disadvantaged students in London state schools (Blanden, Greaves, Gregg, Macmillan, & Sibieta, 2015). This remarkable success is a puzzle because the improvement was not predicted and resists explanation from commonly understood factors. Demographic changes cannot explain the improvement (Blanden et al., 2015). It appears that more resources, a successful teacher recruitment campaign and new buildings have played a supportive, if not decisive, role and that new institutions focused on school management helped (Baars et al., 2014).

Narrative accounts suggest that an important factor was long-term cross-party commitment from Westminster (Baars et al., 2014). If this is so, London schools may have gained from geographic proximity to where final decisions are taken. Political leaders and policymakers may struggle to discover what is going on throughout the whole school system but perhaps they are more likely to have shared knowledge of what schools need in their more immediate locality, as well as a commitment to their improvement. This is valuable to know as it suggests that sufficient political will and sustained coordination can generate impressive outcomes. But it is also indicative of a possible pitfall of relying on a centralised education system. Repeating this success in less connected regions may be impossible. Critically, this successful drive to school improvement did not involve reforming only on the basis of what EBE proponent take to be strong evidence. This inability to explain the trends in London suggests that there is open frontier for a more methodologically diverse approach to EBE to discover what has worked there and how it might work elsewhere.

It could be objected, *pace* these important, yet sometimes mysterious, system-level and external influences on educational attainment, that there are still good reasons to study experimentally what happens in schools and classrooms so as to ensure they are making the best contribution they can. This is true for some purposes, but what this objection misses is that these structural factors cannot be taken as fixed. The world does not stand still while schools adapt their practice according to the results of RCTs. These factors are subject to frequent change in ways that overwhelm the impacts of interventions or even reverse their impact. They are very often the moderating factors that impact whether a school-level intervention will be successful or not. Or, a change in environment may solve a problem more comprehensively than a classroom intervention.

As Christodolou (2016) illustrates, it is very hard to implement successfully something even as well-supported as feedback unless systems of accountability permit it and there is shared

understanding of how feedback is meant to fit into the process of learning. Thus even if EBE research avoids making claims about structural reforms or changes, it cannot treat structural and environmental features as irrelevant. To be useful, EBE needs to suggest school-level and classroom strategies that are robust to social and policy environments which are subject to change. Because many of these changes will happen outside any previously observed circumstances, the reasons for taking these strategies to hold in new conditions will have to be theoretically inferred. It is precisely these issues that an EBE unencumbered by particular methodological commitments is well-placed to tackle.

Conclusion

In this article, I have explained the limits of an experimental approach to EBE. I have discussed how, nevertheless, the experimental paradigm exerts disproportionate influence on policymakers because of the particular role it can play in guiding the policy of a centralized bureaucracy. Finally, I have described what a less distorted version of EBE might look like.

There is a risk when making these overview critiques of a policy process, especially one that draws on tensions between field professionals, administrators and policymakers, of appearing to diminish the sincere efforts of reformers to make the education sector better performing. There is an important caveat to my concerns. Separate from its bureaucratic function, the EBE movement has made some significant advancements on previous approaches to reform in education. Chiefly, it has brought the problem of causal inference to professional attention. This contrasts with some 'best practice' approaches that naively presume that whatever policies prevail in successful schools must be ones that are worth imitating in struggling school settings. This is critical for shifting practitioner-focused research away from duels between theories without rigorous empirical testing, as well as saving teacher training from roving charismatic consultants who in reality offer very little substantive guidance (Coe, 2013).

My concern is that having identified causal inference as a major challenge in education research, EBE proponents might leap to a mistaken conclusion: that causality can best be shown through repeated experimental studies or studies that imitate as closely as possible experimental approaches. At the same time, acceptable solutions to these complex problems have become too easily filtered by bureaucratic convenience. Ultimately, EBE will perform better if it draws on the full range of available research techniques so that it can select those best suited to dealing with each issue in public education.

References

Alexander, D., & Letwin, O. (2013). What Works: evidence centres for social policy. Retrieved from http://dera.ioe.ac.uk/17396/1/What_Works_publication.pdf

Ashman, G. (2018). *The truth about teaching: an evidence-informed guide for new teachers* (1st edition). Thousand Oaks, CA: SAGE Publications.

Baars, S., Bernardes, E., Elwick, A., Malortie, A., McAleavy, T., McInerney, L., … Riggall, A. (2014). *Lessons from London Schools: investigating the success* (p. 136). Reading: CfBT Education Trust.

Banerjee, A. V., & Duflo, E. (2012). *Poor economics: a radical rethinking of the way to fight global poverty* (Paperback first published). New York: PublicAffairs.

Behavioural Insights Team. (2011). Behavioural Insights Team annual update 2010–11. *Cabinet Office: London, UK*. Retrieved from http://casaa.org/wp-content/uploads/Behaviour-Change-Insight-Team-Annual-Update_acc.pdf

Bergman, M., & Lane, J.-E. (1990). Public Policy in a Principal-Agent Framework. *Journal of Theoretical Politics*, *2*(3), 339–352.

Blanden, J., Greaves, E., Gregg, P., Macmillan, L., & Sibieta, L. (2015). *Understanding the improved performance of disadvantaged pupils in London* (Working Paper No. 21) (p. 47). London: London School of Economics.

Boaz, A., Ashby, D., Young, K., & ESRC, U. (2002). *Systematic reviews: what have they got to offer evidence based policy and practice?* ESRC UK Centre for Evidence Based Policy and Practice London. Retrieved from http://www.kcl.ac.uk/sspp/departments/politicaleconomy/research/cep/pubs/papers/assets/wp2.pdf

Borman, G. D., & Dowling, N. M. (2008). Teacher Attrition and Retention: A Meta-Analytic and Narrative Review of the Research. *Review of Educational Research*, *78*(3), 367–409.

Buchanan, J. M. (1999). Politics without romance. In *The logical foundations of constitutional liberty* (pp. 45–59). Indianapolis: Liberty Fund.

Campbell, D. T. (1979). Assessing the impact of planned social change. *Evaluation and Program Planning*, *2*(1), 67–90.

Cartwright, N. (2007). Are RCTs the Gold Standard? *BioSocieties*, *2*(1), 11–20.

Cartwright, N. (2013). Knowing what we are talking about: why evidence doesn't always travel. *Evidence & Policy: A Journal of Research, Debate and Practice*, *9*(1), 97–112. 1

Cartwright, N., & Hardie, J. (2012). *Evidence-based policy: a practical guide to doing it better*. Oxford ; New York: Oxford University Press.

Cartwright, N., & Munro, E. (2010). The limitations of randomized controlled trials in predicting effectiveness: Limitations of RCTs for predicting effectiveness. *Journal of Evaluation in Clinical Practice*, *16*(2), 260–266.

Christodoulou, D. (2016). *Making good progress? the future of assessment for learning* (NEW edition). Oxford, U.K: Oxford University Press.

Cochrane, A. L. (1972). *Effectiveness and efficiency: Random reflections on health services*. Nuffield Trust.

Cochran-Smith, M., & Lytle, S. (2006). Troubling Images of Teaching in No Child Left Behind. *Harvard Educational Review*, *76*(4), 668–697.

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

Coe, R. (2013). *Improving education: a triump of hope over experience* (Speech). Durham University: Centre for Evaluation and Monitoring.

Cowen, N. (2017). Why be robust? The contribution of market process theory to the Robust Political Economy research program. In P. J. Boettke, C. J. Coyne, & V. Storr (Eds.),

Interdisciplinary Studies of the Market Order: New Applications of Market Process Theory (pp. 63–85). London: Rowman and Littlefield International Ltd.

Cowen, N., Cartwright, N., Virk, B., & Mascarenhas-Keyes, S. (2015). Making the Most of the Evidence: Evidence-based policy in the classroom, 49.

Cowen, N., Virk, B., Mascarenhas-Keyes, S., & Cartwright, N. (2017). Randomized Controlled Trials: How Can We Know "What Works"? *Critical Review*, *29*(3), 265–292.

Crafts, N. (1996). "Post-neoclassical endogenous growth theory": what are its policy implications? *Oxford Review of Economic Policy*, *12*(2), 30–47.

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

Deans for Impact. (2015). *The Science of Learning*. Austin, TX: Deans for Impact.

Deaton, A., & Cartwright, N. (2018). Understanding and misunderstanding randomized controlled trials. *Social Science & Medicine*, *210*, 2–21.

Donmoyer, R., & Galloway, F. (2010). Reconsidering the Utility of Case Study Designs for Researching School Reform in a Neo-Scientific Era: Insights From a Multiyear, Mixed-Methods Study. *Educational Administration Quarterly*, *46*(1), 3–30.

Duflo, E., Glennerster, R., & Kremer, M. (2007). Chapter 61 Using Randomization in Development Economics Research: A Toolkit. In *Handbook of Development Economics* (Vol. 4, pp. 3895–3962). Elsevier.

Dunleavy, P. (2002). *Democracy, bureaucracy and public choice: economic explanations in political science*. London: Prentice Hall.

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

Eisenhardt, K. M. (1989). Agency Theory: An Assessment and Review. *The Academy of Management Review, 14*(1), 19.

Eisenhart, M., & Towne, L. (2003). Contestation and Change in National Policy on "Scientifically Based" Education Research. *Educational Researcher*, *32*(7), 31–38.

Every-Palmer, S., & Howick, J. (2014). How evidence-based medicine is failing due to biased trials and selective publication: EBM fails due to biased trials and selective publication. *Journal of Evaluation in Clinical Practice*, *20*(6), 908–914.

Flinders, M. (2000). The enduring centrality of individual ministerial responsibility within the British constitution. *The Journal of Legislative Studies*, *6*(3), 73–92.

Fortin, M. (2006). Randomized Controlled Trials: Do They Have External Validity for Patients With Multiple Comorbidities? *The Annals of Family Medicine*, *4*(2), 104–108. https://doi.org/10.1370/afm.516

Gibb, N. (2017, February). *The importance of an evidence-informed profession*. Speech, University of Buckingham.

Goldacre, B. (2013). Building evidence into education. Retrieved from http://dera.ioe.ac.uk/17530/1/ben%20goldacre%20paper.pdf

Goodhart, C. (1981). Problems of monetary management. In A. S. Courakis (Ed.), *Inflation, depression, and economic policy in the West* (pp. 111–146). Totowa, N.J: Barnes & Noble Books.

Gorard, S., & Siddiqui, N. (2018). Grammar schools in England: a new analysis of social segregation and academic outcomes. *British Journal of Sociology of Education*, 1–16. https://doi.org/10.1080/01425692.2018.1443432

Gough, D. A., Oliver, S., & Thomas, J. (2013). *Learning from research: systematic reviews for informing policy decisions: a quick guide*. Nesta London, UK. Retrieved from http://ktdrr.org/cgi-bin/ktstrategies_search.cgi?location=sr&sel_1=174

Halpern, D. (2016). *Inside the nudge unit: how small changes can make a big difference*.

Hattie, J. (2009). *Visible learning: a synthesis of over 800 meta-analyses relating to achievement*. London ; New York: Routledge.

Higgins, S., Katsipataki, M., Villanueva-Aguilera, A. B., Coleman, R., Henderson, P., Major, L. E., … Mason, D. (2016). *The Sutton Trust-Education Endowment Foundation Teaching and Learning Toolkit.* Education Endowment Foundation. Retrieved from http://dro.dur.ac.uk/20987/

Hofmann, S. G., Asnaani, A., Vonk, I. J. J., Sawyer, A. T., & Fang, A. (2012). The Efficacy of Cognitive Behavioral Therapy: A Review of Meta-analyses. *Cognitive Therapy and Research*, *36*(5), 427–440.

Holmlund, H., McNally, S., & Viarengo, M. (2010). Does money matter for schools? *Economics of Education Review*, *29*(6), 1154–1164.

Inglis, M., & Foster, C. (2018). Five Decades of Mathematics Education Research. *Journal for Research in Mathematics Education*, *49*(4), 462.

Ioannidis, J. P. (2017). Meta-analyses Can Be Credible and Useful: A New Standard. *Jama Psychiatry*, *74*(4), 311–312.

Joyce, K. E., & Cartwright, N. (2018). Meeting our standards for educational justice: Doing our best with the evidence. *Theory and Research in Education*, *16*(1), 3–22.

Krauss, A. (2018). Why all randomised controlled trials produce biased results. *Annals of Medicine*, *50*(4), 312–322.

Lawn, M. (2013). A Systemless System: Designing the Disarticulation of English State Education. *European Educational Research Journal*, *12*(2), 231–241.

Lucas, D. S. (2017). Evidence-based policy as public entrepreneurship. *Public Management Review*, 1–21. https://doi.org/10.1080/14719037.2017.1412115

Major, L. E., & Higgins, S. (2015). The toolkit four years on: lessons for spending the pupil premium. In *The pupil premium: next steps* (pp. 16–18). Sutton Trust / Education Endowment Foundation.

Marshall, P., Rabindrakumar, S., & Wilkins, L. (2007). *Tackling educational inequality*. London: CentreForum.

Marzano, R. J., Pickering, D. J., & Pollock, J. E. (2005). *Classroom instruction that works: reserach-based strategies for increasing student achievement*. Upper Saddle River, NJ: Pearson/Prentice Hall.

McCaughey, D., & Bruning, N. S. (2010). Rationality versus reality: the challenges of evidence-based decision making for health policy makers. *Implementation Science*, *5*(1), 39.

Meadowcroft, J. (2003). The British National Health Service: Lessons from the "Socialist Calculation Debate." *The Journal of Medicine and Philosophy*, *28*(3), 307–326.

Meadowcroft, J. (2005). Health Care Markets, Prices, and Coordination: The Epistemic explanation of Government Failure and the UK National Health Service. *HEC Forum*, *17*(3), 159–177.

Mueller, D. C. (1976). Public choice: A survey. *Journal of Economic Literature*, *14*(2), 395–433.

Newton, P. M., & Miah, M. (2017). Evidence-Based Higher Education – Is the Learning Styles 'Myth' Important? *Frontiers in Psychology*, *8*. https://doi.org/10.3389/fpsyg.2017.00444

Niskanen, W. A. (1968). The peculiar economics of bureaucracy. *The American Economic Review*, *58*(2), 293–305.

Niskanen, W. A. (1994). *Bureaucracy and public economics*. Aldershot, Hants, England ; Brookfield, Vt., USA: E. Elgar.

Ostrom, V. (1993). Epistemic choice and public choice. *Public Choice*, *77*(1), 163–176.

Ostrom, V., & Ostrom, E. (1971). Public Choice: A Different Approach to the Study of Public Administration. *Public Administration Review*, *31*(2), 203.

Palmer, M. S. (1995). Toward an economics of comparative political organization: examining ministerial responsibility. *Journal of Law, Economics, & Organization*, 164–188.

Pashler, H., McDaniel, M., Rohrer, D., & Bjork, R. (2008). Learning Styles: Concepts and Evidence. *Psychological Science in the Public Interest*, *9*(3), 105–119.

Pennington, M. (2011). *Robust political economy: classical liberalism and the future of public policy*. Cheltenham, UK ; Northampton, MA, USA: Edward Elgar.

Pritchett, L. (2002). It pays to be ignorant: A simple political economy of rigorous program evaluation. *The Journal of Policy Reform*, *5*(4), 251–269.

Reddy, S. G. (2012). Randomise This! On Poor Economics. *Review of Agrarian Studies*, *2*(2), 60–73.

Sanders, M., & Halpern, D. (2014, February 3). Nudge unit: our quiet revolution is putting evidence at heart of government. *The Guardian*. Retrieved from https://www.theguardian.com/public-leaders-network/small-business-blog/2014/feb/03/nudge-unit-quiet-revolution-evidence

Simpson, A. (2017). The misdirection of public policy: comparing and combining standardised effect sizes. *Journal of Education Policy*, *32*(4), 450–466.

Slavin, R. E. (2002). Evidence-Based Education Policies: Transforming Educational Practice and Research. *Educational Researcher*, *31*(7), 15–21.

Thaler, R. H., & Sunstein, C. R. (2009). *Nudge: improving decisions about health, wealth and happiness*. London: Penguin Books.

Thomas, G. (2012). Changing Our Landscape of Inquiry for a New Science of Education. *Harvard Educational Review*, *82*(1), 26–51.

Tullock, G. (2005). *Bureaucracy*. (C. K. Rowley, Ed.). Indianapolis: Liberty Fund.

Whitty, G. (2008). Twenty Years of Progress?: English Education Policy 1988 to the Present. *Educational Management Administration & Leadership*, *36*(2), 165–184.

Young, K., Ashby, D., Boaz, A., & Grayson, L. (2002). Social Science and the Evidence-based Policy Movement. *Social Policy and Society*, *1*(03), 215–224.