Nick Cowen, Baljinder Virk,

Stella Mascarenhan-Keyes, and Nancy Cartwright

RANDOMIZED CONTROLLED TRIALS: HOW CAN WE KNOW "WHAT WORKS"?

ABSTRACT: We attempt to map the limits of evidence-based policy through an interactive theoretical critique and empirical case-study. We outline the emergence of an experimental turn in EBP among British policymakers and the limited, broadly inductive, epistemic approach that underlies it. We see whether and how field professionals identify and react to these limitations through a case study of teaching professionals subject to a push to integrate research evidence into their practice. Results suggest that many of the challenges of establishing evidential warrant that EBP is supposed to streamline re-appear at the level of choice of locally effective policies and implementation.

Keywords: *behavioral economics; evidence-based policy; meta-analysis; nudge; nudging; systematic review; education policy.*

Nancy Cartwright, nancy.cartwright@durham.ac.uk, Department of Philosophy, Durham University, 50 Old Elvet, DH1 3HN, United Kingdom, thanks the Center for Ethics and Education, the National Science Foundation (grant no. 1632471), and the European Research Council (under the European Union's Horizon 2020 research and innovation program, grant agreement no. 667526 K4U) for funding this project. Nick Cowen, nick.cowen@nyu.edu, New

York University School of Law, 110 West Third Street, Room 223, New York, NY, 10012,
thanks Carmen Pavel, John Meadowcroft, Anastasia de Waal, John Owens, Hayley Davies, and the Centre for Public Policy Research seminar group, King's College London, for useful comments on versions of this paper, as well as the financial support of the Economic and Social Research Council (ES/J500057/1) and the Cabinet Office. Stella Mascarenhas-Keyes, stellamaskey123@hotmail.co.uk, and Baljinder Virk, drvirkassociates@gmail.com, thank the research participants and fellow authors. It is acknowledged that the content of this work reflects only the authors' views and that the ERC and the Cabinet Office are not responsible for any use that may be made of the information it contains.

Evidence-based policy (EBP) is an approach to public decision-making that is informed by the results of scientific research. Proponents see EBP as a straightforward and enlightened approach to achieving shared social goals. “What works,” a common slogan amongst proponents of EBP, connotes an interest in which policy approaches are most effective, i.e., which interventions or approaches are known to produce a given desired outcome. By contrast, skeptics see EBP as leading to inappropriate one-size-fits-all policies that avoid interrogating both the complex structures of the variegated social contexts in which policies are imposed, and the values underlying the policies (Biesta 2007; Holmes et al. 2006).

Contemporary EBP has come to be associated with the use of research with two distinctive characteristics. The first is an emphasis on systematic review and formal meta-analysis (Boaz et al. 2002; Young et al. 2002). Systematic reviews scan the research literature according to a pre-established protocol and attempt to summarize all the empirical evidence fitting pre-specified eligibility criteria that can be found about a given hypothesis. Meta-analyses use statistical methods to produce a summary result of the findings of studies that meet pre-specified criteria—often an estimated average effect-size of a set of interventions or policy approaches. The second is the use of randomized-controlled trials (RCTs) as the “gold standard” of evidence (Cartwright 2007; Coe 2004; Goldacre 2013; Sanders and Halpern 2014). RCTs are often considered uniquely capable of determining a causal relationship between an intervention and outcome. Despite their increasing popularity in public policy, however, some researchers argue that RCTs have significant limitations (Cartwright and Hardie 2012; Deaton and Cartwright 2016; Every-Palmer and Howick 2014; Greenhalgh et al. 2014; Hammersley 2013; Morrison 2001; Slavin 2008; Slavin and Smith 2009).

Given the plausible theoretical limitations with evidence-based policy, how then do field professionals engage with and implement it? Our hypotheses are that:

1. Field professionals will identify evidence-based policy as general in nature, rarely having direct applicability to their local context.

2. As a result, field professionals will draw on a much wider range of sources than formal experimental evidence in order to make sense of their practice and when considering alternative policies.

The structure of this paper is as follows. First we outline the emergence of EBP in the United Kingdom[1] and summarize some of its limitations. Then we explore whether and how field professionals experience these limitations by means of a case study of the use of EBP in the education system in England, a public service that has come under increasingly detailed government direction in recent decades.

Our hypotheses emerged through an ongoing theoretical analysis to identify the limitations in EBP methodology that can affect how useful practitioners will find it, in tandem and interactively with analysis of our primary data. Our primary data are interviews with teaching professionals. Using this data, we discuss how the field practitioners in our study engage with research evidence disseminated with government support.

[1] The experimental turn in EBP also reflects international trends in public policy, especially in the United States. We focus on the United Kingdom partly because of the location of our case study but also because it is a relatively centralized state with a unitary jurisdiction and few checks on the Government of the day. As a result, it seems that relatively rapid shifts in policymaking approaches led by the executive are more practical to implement and can be more easily observed.

I. THE RISE OF EXPERIMENTAL METHODS IN EVIDENCE-BASED POLICY

EBP may not always have had its current name or initials, but it is a perennial feature of public administration. It ebbs and flows in popularity in public discourse, sometimes being at the forefront, other times disappearing from sight as other trends prevail. In the United Kingdom, the current cycle arguably started when the New Labour Government (1997-2010) announced its commitment to EBP through a lecture given by David Blunkett (2000) to the Economic and Social Research Council (the United Kingdom's primary public funder of research in the social sciences). Blunkett wanted academic research to inform better policymaking. His speech represented a promise that policy makers would listen to academic researchers, but it also suggested the obligation for academics to pursue research aimed at achieving the common good.

Over the course of the New Labour years, the commitment to EBP encountered increasing skepticism. Evidence did not feed directly or straightforwardly into policy, with apparently well-established evidence being ignored when it was politically inconvenient (Marmot 2004). A wry inversion of EPB, "policy-based evidence-making," started to become a common slogan among skeptics (Hunter 2009), with the understanding that evidence itself could not be assumed to be value neutral, especially when it was commissioned by government and corporate interests. As the polish started to come off, it was possible to imagine EBP losing prominence and becoming just one of many approaches of making (and publicly justifying) political decisions.

Under the coalition Government (2010-2015), however, the sun rose again on EBP, this time with a specific emphasis on experimental methods. The resurgence was initially associated with Prime Minister David Cameron's interest in "nudge" theory (Bovens 2009; Goodwin 2012; Loewenstein et al. 2012; McAnulla 2010). "Nudge" theory uses research evidence from the field

of behavioral economics to suggest policies aimed at promoting welfare gains by reframing the choices that individuals make, especially regarding personal health and financial management, where short-term decisions can fail to align with long-term personal interests (Thaler and Sunstein 2009). The new Government established the Behavioural Insights Team, led by David Halpern, within the Cabinet Office, to investigate how nudges could be implemented to improve policy outcomes (Halpern 2016). The Behavioural Insights Team conducted several of its own policy experiments (Behavioural Insights Team 2011, 2012; Harper 2013). The team has since been transformed into an independent social enterprise.

The project of bringing research evidence to bear on policy problems soon expanded beyond nudging to include several bodies of research evidence that relate to public-sector provision in general. "What Works" was transformed from slogan to institution with the establishment of the What Works Network (Alexander and Letwin 2013), with David Halpern as its national adviser, and its secretariat in Her Majesty's Cabinet Office. It is currently comprised of seven centers and two affiliate members, the most prominent being the National Institute for Health and Care Excellence, which sets guidance for medical practice as well as funding guidelines for the National Health Service. There are plans underway for an eighth center, the "What Works Centre for Children's Social Care."

II. LIMITATIONS OF RANDOMIZED CONTROLLED TRIALS (RCTs) AND OTHER ASPECTS OF EVIDENCE-BASED POLICY

Reflecting their inspiration by behavioral economics and clinical medicine, both the Behavioural Insights Team and the What Works Network place a premium on experimental research designs. As Sanders and Halpern (2014) explain:

> RCTs are the so-called gold standard of evidence-based policy. In an RCT, participants are randomly assigned either to receive an intervention . . . or not (a control group). Since these two groups tend to be the same in all other respects, they can be compared to analyse the effects of a policy.

Academic scholarship offers a more mixed appraisal of RCTs and their capacity to inform policy. Cartwright and Munro introduce some concerns about how to interpret RCTs, arguing that they are by themselves typically "insufficient to meet the needs of policy or practice decision makers" (Cartwright and Munro 2010, 265). The primary result of an RCT is a judgment about the effectiveness of a treatment in producing an outcome in the study population, although the contribution of other factors to the outcome is unknown. The purpose of RCTs in the context of EBP is to help estimate the effectiveness of a treatment in a target population or populations different from the one on which the experiment was conducted, which raises questions about when such inferences are valid.

As Cartwright and Munro (2010, 261) note, "it is widely acknowledged that we generally don't know all the important causes for a factor, let alone knowing [their] distribution . . . in the study and the target populations." There are two categories of causal factors, besides the intervention itself, that might affect the outcome. First, there are factors that operate independently of the intervention. These will affect the overall size of the outcome but have no effect on what the intervention itself contributes to the outcome. Second, there are factors that moderate how much effect the intervention can produce, factors that must be in place lest the intervention fail to produce its expected contribution. These are called *support factors* (sometimes also "interactive factors" or "moderator factors"). Suppose, for example, that the

intervention is the creation of an afterschool homework club. This might work, but only if the children can get home from the club after doing their homework. In some environments, that might not be a significant challenge, but in others, transport to and from school might be a critical support factor in the success of the treatment. Support factors will play a significant role in our discussion because they bear on the central question of EBP: whether the intervention or policy will "work" in a targeted setting, i.e., whether it will produce some positive contribution to the desired outcome there.

Since the same policy or intervention will have different effects in different populations that have different support factors or different distributions of support factors, it matters in new settings which support factors are present and in what proportions. Extrapolating from positive outcomes in an RCT in one study population (or even a set of RCTs in a few different study populations) to the claim that some intervention will work in a new setting, or that it works in general, assumes precisely the level of knowledge that is absent when RCTs are called for. If that level of knowledge were present, an RCT would hardly be necessary.

Cartwright and Munro (2010, 262) trace three kinds of causal claim:

1. It-works-somewhere claims: treatment T causes outcome O somewhere, under some conditions (e.g., in study population X, administered by method M).

2. Capacity claims: T has a (relatively) stable capacity to promote O, so that it can be expected to do so widely.

3. It-will-work-for-us claims: T would cause O in population Q if administered as directed by policy P (i.e., effectiveness claims).

RCTs are immediately relevant for estimating the first type of causal claim, since they tell us whether (or the degree to which) the intervention influenced the targeted outcome in the

population enrolled in the experiment. RCTs' relevance for the other two types of causal claim is indirect and incomplete. Whether a treatment can be assumed to be relevant to a given untreated population depends on a fabric of other knowledge.[2] The second type of claim requires enough theoretical, empirical, and conceptual knowledge to support the claim that what happens in one or a handful of study settings will happen widely. Imagine, for example, what it takes to support the claim that the charge on one electron is the same as the charge on all. For the third type of claim, there must be good reason, both theoretical and empirical, to warrant the assumption that the RCT population and the target are alike in just the right ways to support the same causal pathways from intervention to outcome. Moving from the first to the second or third types of claim, then, requires a great deal of knowledge that cannot be warranted by the RCT itself. The problem is that "this kind of complicated causal reasoning is hard, even if we are prepared to be rough in our approximations and figure out ways to tolerate uncertainties" (Cartwright and Munro 2010, 263).

This is not to suggest that making such inferences is impossible. Sometimes, large observed effects can be relied upon to overwhelm less observable features and thus can allow practitioners and policymakers to act confidently even without complete knowledge of other factors. Cartwright and Hardie (2012, 39) acknowledge that an "ideal" RCT may be able to "clinch" a case that an intervention works in some particular study population. However, they argue that, depending on what background knowledge is available, the same can equally be said

[2] One reason to make such an inference would be that the experimental population is a sufficiently large random sample of the target population. However, an RCT seldom involves randomly sampling the entire population of interest. Instead, the sample population is drawn from those who are accessible to researchers and suitable for participation in a trial, usually in a limited range of places and at a specific point in time. It is only after that selection that individuals are randomly allocated to treatment and control groups.

of other research designs, including causal Bayes nets models, econometric analysis, and process tracing.

Limitations of Systematic Reviews

As with RCTs, the proponents of systematic research reviews draw their inspiration primarily from evidence-based medicine. As one official statement of support for EBP puts it:

> A systematic review attempts to collate all empirical evidence that fits pre-specified eligibility criteria in order to answer a specific research question. It uses explicit, systematic methods that are selected with a view to minimizing bias, thus providing more reliable findings from which conclusions can be drawn and decisions made. (Liberati et al. 2009, 2)

A systematic review may be designed to include qualitative studies, while still typically preferring studies with quantitative measures if they are available (Coe 2004, 3). What EBP proponents take to be the core of systematic reviews is clear definitions of interventions and outcomes, formally set criteria, and subsequent rigor in analysis (Puttick and Mulgan 2013).

In contrast to the use of rigid formulas and stringent ranking of evidence, Cartwright and Hardie (2012) note an important role for researcher judgment in evaluating a full range of research evidence. This includes the acknowledgement of tradeoffs between different ways of including and excluding studies on the basis of their presumed capacity to allow for causal

inference. Many systematic reviews begin by identifying a large number of studies but often end up including only a handful that meet the precise criteria of the protocol. Is this to be preferred? A small number of "high-quality" experimental studies may indicate that a treatment works, but there may be common factors or circumstances that allow such high-integrity studies to be conducted. By contrast, a much larger number of lower-quality studies may indicate something different. The reason for considering only the higher-quality studies is that they should be better at identifying causality. But it could be that the wider range of studies indicates limits to the circumstances where the treatment works.

More fundamentally, it is difficult to see what a systematic review is supposed to provide evidence for. Each separate study can at best provide direct evidence about whether, or the extent to which, the intervention works in the population measured in that study. We could then be very modest in our claims, using the systematic review to support the claim, "The intervention works somewhere." The motivation for this modesty is that we know that no matter how well done a study is, it will have flaws; we also know that the frequency of the outcome observed in a study is only an *estimate* of the probability for the outcome in the study population, and will sometimes be far off the probability for that population. So we should never trust any single study to establish that the intervention works in that very population (i.e., somewhere). One way to support the claim that it does work somewhere would be to do a variety of different studies on the same population.

Systematic reviews, though, tend to deal with studies on different populations. Presumably, the thinking is that it is unlikely that a large number of different studies (each on its own population) should all be far off the mark each for the population it investigates. So if they all give the same result, that result is highly likely to be correct for at least some of those

populations, though we don't know which ones. This, though, is an odd way to proceed to such a conclusion. If one wanted to see whether the intervention works in some specific somewhere, then the best evidence would be a variety of strong studies of different kinds all showing either that it did or did not work in that particular place. If the question is whether it works somewhere or other (where a "yes" answer can at least show that it *can* work), the choice of study sites is equally important as the quality of the studies, especially if the risks from mistakenly inferring that it does not work somewhere are high. As is emphasized in case-study methodology: if the sites are ones where it is unlikely to work in the first place, failing to find it working is little evidence against the hypothesis that it can work.

Of course, the growing investment in systematic reviews, and much of the language and practice surrounding them, suggest far bolder ambitions than showing that interventions work "somewhere" (Liberati et al. 2009; Noonan and Bjørndal 2010). The aim often seems to be to estimate whether the treatment promotes the targeted outcome across a wide range of cases, or perhaps that its effectiveness (alternatively, ineffectiveness) should be the default assumption, barring reasons to the contrary. If so, however, what is generally billed as a strong evidence base is nothing of the kind (Fleischhacker 2017). Imagine that a review found many good studies in different places that indicated a positive (alternatively negative or zero) effect. What can be concluded? Very little, even supposing, what is often not the case, that the number of study sites is large. To conclude on this basis that the treatment has positive effects (alternatively negative or zero effects) generally is induction by simple enumeration, which is widely acknowledged to be a flawed form of inference: *Swan 1 is white, swan 2 is white, . . . swan 6754 is white; therefore all swans are white.*

Even if the sites are varied, this adds little to support the inference to generalizability without good reasons to suppose that they are varied *in the right way*, reasons that will necessarily be based on a combination of other empirical research and theoretical and conceptual development. As noted above, sometimes such reasons are available. Ideally, systematic reviews will have located such research before advancing any conclusions about general or widespread effectiveness, but it seems that not many do so.

Context and Abstraction

The overall lesson is that *context matters* (McCormack et al. 2002; Seckinelgin 2016; Waters et al. 2006, 288; White n.d.). Some interventions will work only because of very special circumstances; they can work in some places but don't have a widespread potential to succeed. Even those that have widespread potential do not operate on their own; they will work only when the requisite support factors are in place, or some suitable substitute for them.

That context matters is fast becoming accepted across the EBP literature, but its substantive implications are not. The problem for practitioners is that EBP clearing houses and what-works centers are far less good at providing information and advice about what it is about local contexts that matters to a policy success than they are at vetting and summarizing evidence about how well the policy has succeeded in study sites. This is sometimes because the relevant information is knowable (at least in principle) but is not an area of interest for researchers. In other cases, the relevant information is essentially inaccessible to researchers.

Supposing that a particular intervention has relatively wide potential to improve targeted outcomes, we have identified two specific kinds of information that need to be acquired if we are

to predict how successful the intervention would be in a new context: what the support factors are and how they are distributed in the new context and how to "de-abstract" or "contextualze" generalizations. The first we have discussed at some length; support factors are those factors required for the intervention to work in the context in question. We need to know that hey (or an appropriate substitute) will be in place at the right time and to the right degree. The second needs some explanation.

General truths -- claims that apply consistently in new and different contexts -- tend to use fairly abstract concepts. Consider a pastiche of a military example. You might know that when firing cannonballs with a normal charge at a 35-degree angle, they reach the enemy's line 400 yards away. That's very helpful for setting the angle of your cannon so long as you are just 400 yards from the enemy line. For more general purposes, you need far more abstract concepts: e.g., that the *trajectory* of a cannonball is a *parabola*. More useful still is the abstract formula that connects horizontal distance travelled with the initial angle and velocity of firing. Or consider feedback for students to improve learning. There is by now a large typology of types of feedback—*direct, formative, grading, praise, verification of response accuracy, explanation of the correct answer*, *hints, worked examples,* and more, all fairly abstract descriptions. And there are theories about why they are or are not effective. You may not be able to create precisely the same type of feedback that produced positive outcomes in a given study, but knowing that generally "feedback of type x works" can be helpful in designing your own intervention. But it can only be helpful if you can figure out what these more general concepts they amount to in your setting. Just what constitutes a "hint" in teaching your students long division, or what will serve as an "explanation", or what will your students feel as "praise"? Knowledge formulated in

abstract concepts is only of use in practice if we know, for the situation at hand, how to "de-abstract" or "contextualize" it.

Moving to a higher level of abstraction also has benefits when it comes to warranting general claims. The more studies that report the same outcome in the study population, and the more varied the settings, the stronger is the warrant for concluding that the intervention generally promotes an improvement. Thus, it is helpful to be able to lump together interventions that differ in a variety of ways but all satisfy the same abstract description, so long as that description is relevant. Measurements of the great variety of distances that cannonballs across the ages have traversed when the cannons are pitched at very different angles and fired with very different velocities, when grouped together, provide strong warrant that cannonballs travel along parabolas. The amount and variety of evidence for the parabola hypothesis is overwhelming. Moreover, there has been a great deal of conceptual and theoretical development to back it up. Neither of these are usually true with respect to social policies. Indeed, unlike cannonballs, human beings are affected by heterogeneous motivations, which is one commonly cited reason for the difficulty the social sciences have in coming up with general theories that hold reliably across individuals. Yet, lumping under more abstract descriptions does improve strength of warrant, so long as the descriptions are relevant (cf. Simpson 2017) so it is not surprising that researchers and EBP sites try to do so..

Return now to practitioner problems. The move to more abstract descriptions of interventions is a mixed blessing for practitioners. It provides them with a set of interventions with reasonably strong warrant. But it does not tell them what these interventions look like in their own setting. Abstract knowledge needs to be "de-abstracted.", to be of practical use.

Thus far, we have identified these two problems that practitioners face – the problem of *support factors* and the problem of *de-abstracting*—from a theoretical point of view, based on conventional theories of causation (Mackie 1965) and of evidential warrant (Ayer 1972; Hume 2008; Quine and Ullian 1978). Despite these imperfections, RCTs and systematic reviews have come to enjoy a privileged, though not exclusive, status amongst proponents of EBP (Puttick and Mulgan 2013). This raises the question: Are the problems we have identified felt as problems by the practitioners themselves?

In an attempt to begin answering this question, we have conducted a case study of educators in Britain subject to a government-backed push for EBP in education, with the aim of discovering how field professionals interpret and react to evidence from EBP resources. In the next section, then, we discuss education policy in England, the extension of experimental EBP into the sector by central government, the major institution supporting EBP in England—the Education Endowment Foundation—and the hypotheses we have extracted from our study, our methods and approach to data collection, and our results.

III. PUBLIC EDUCATION IN ENGLAND: A CASE STUDY

For decades after education became both free and compulsory, English schools enjoyed significant autonomy from the central government. Local authorities were responsible for establishing and maintaining schools so that all children had free access to a school. Teacher training was at first handled informally, then organized around relatively autonomous vocational teaching colleges associated with universities (Crook 2002). School inspections were conducted by Her Majesty's Inspectorate of Schools. The education sector maintained significant independence from Parliament and the executive governments of the day. National exams were

set by independent boards with links to higher education. Up until the 1970s, the public debate focused on structural factors such as selection: whether the state system should continue to provide an upper tier of Grammar Schools for academically gifted children or move to a fully comprehensive system where children of all abilities are taught in the same schools (Gorard 2015, 258–59).

This started to change with Labour Prime Minister James Callaghan's controversial "secret garden" speech in 1976. Callaghan announced that the content and pedagogy of classroom instruction was not a matter to be decided between professionals alone, walled off from scrutiny, but concerned the public; and that, as a result, the details of educational provision were a responsibility of government. From then on education became more commonly a subject of policy intervention and wider public debate.

Education reforms continued under subsequent Conservative governments. These included the introduction of a National Curriculum that defined expected course content for all schools (Whetton 2009). The inspection system was reformulated around a new national agency, Ofsted (de Waal 2006, 2008), which was intended to enforce standards and ensure that government policy was carried out in schools. These attempts to strengthen the hand of the central government were combined with a number of market-inspired reforms, including the introduction of more national examinations and school "league tables," which were intended to make objective measures of school quality available to parents (Chitty and Dunford 1999; Mansell 2007; Wyse and Torrance 2009).

Subsequently, spurred on by theories of human capital and the possibilities of "neo-endogenous growth" (Crafts 1996), the New Labour government came to see education as the key to future economic prosperity and as, therefore, intimately linked to all other aspects of

government objectives. The government introduced more detailed guidance for classrooms, including compulsory daily literacy (Machin and McNally 2008) and numeracy hours in primary schools. New Labour's flagship policy was a new school structure, Academies, which are independent of local authorities. Under New Labour, these were established, often with a private sponsor, in urban areas that were judged to be poorly served by existing local-authority maintained schools. This was sometimes intended to expand choice for parents and competition among existing schools (Woods, Woods, and Gunter 2007). In this sense, they are comparable to charter schools found in some U.S. jurisdictions. Under recent Conservative-led governments, Academies have expanded from being an exceptional structure to a norm, with the result that local authorities now have a limited role in education provision and almost no power to direct policy goals. This trend towards centralization has continued, although fitfully, through progressively turning local-authority-controlled schools into notionally "independent" Academies that have a direct funding arrangement with the central government's Department for Education. In addition to Academies, "free schools," a somewhat more autonomous school-type, often managed by associations of local teachers and parents, have also been established. They remain subject to most of the same regulatory structures as Academies.

To sum up, in terms of formal institutions, the power to influence educational provision has shifted from teaching professionals and relatively local forms of governance to central government and its related agencies. Into this situation entered the Education Endowment Foundation. The EEF was first established in 2011 in conjunction with the Sutton Trust (an independent charity). It was designated a "What Works Centre" for primary and secondary education in 2013 (since extended to early-years and post-secondary education). Although its

remit continues to expand, the EEF engages in two core activities that align with the configuration of EBP:

1. Analyzing, summarizing, and disseminating academic research evidence with the intention of improving school-level practice and with the overall policy goal of reducing inequalities of educational attainment amongst students.

2. Funding and supporting experimental trials of new interventions and approaches in schools in order to produce new evidence of what works.

One of EEF's key resources is a regularly updated "teaching and learning toolkit" (Higgins et al. 2016), which summarizes current approaches in terms of effectiveness at improving education progress, cost of intervention, and certainty of the available evidence. How do teaching professionals engage with research evidence of this kind?

The Two Practitioner Problems

As we indicated in raising the problem of de-abstracting, finding a large number of studies in different "contexts" that evaluate the same intervention executed to the same protocol can be extremely difficult. The task is easier if the intervention is described at a general, abstract level that can cover a great variety of different concrete strategies.

Consider "early literacy approaches." The EEF reports that "extensive evidence," including "a number of meta-analyses and high quality individual studies" (Education Endowment Foundation 2016a) supports these approaches, which are said to "include: storytelling and group reading, activities that aim to develop letter knowledge, knowledge of sounds and early phonics, or introductions to different kinds of writing" (Education Endowment

Foundation 2016a). This might bewilder practitioners who have to choose a specific set of strategies for their specific setting.

The EEF also takes note of the central fact we quoted from Munro and Cartwright (2010): that the same intervention can have widely different effects in different populations, depending on what other causal factors are present as support factors in the population. EEF includes in a discussion of *Implementation* the advice that a teacher can either adopt a highly structured program, presumably "as is," or can create "your own interventions by adopting the principles from a Toolkit strand (for example, combining the recommendations from the evidence on Phonics, Small Group Tuition and Teaching Assistants)" (Education Endowment Foundation 2016b). EEF advises that in so doing, one must consider "if the necessary support factors are in place to make the intervention successful" (ibid.), as illustrated with a pie chart based on Cartwright and Cowen (2014), showing a set of factors that support a successful homework strategy: good teacher/pupil relationship, parent buy-in or homework club, timely focused feedback, and the right training support for teachers. This, though, provides no guidance about how to figure out which support factors are needed in a particular case, how to judge whether a contemplated intervention fits the abstract description well enough to produce the expected results, nor how to de-abstract the principles to see what they amount to in a particular setting.

These considerations make it plausible that practitioners will find the EEF guidance too vague or abstract to be of direct help in deciding what might work for them, which is consistent with the following hypotheses:

Hypothesis 1: As a result of the process of aggregating and filtering evidence from multiple studies according to transparency criteria, field professionals will perceive the content

of the evidence disseminated as being of a general character with limited in applicability to local contexts.

Supposing hypothesis 1, hypothesis 2 might follow:

Hypothesis 2: Because of the perceived limited applicability of the disseminated evidence to local contexts, field professionals will draw on a much wider array of evidence, including theoretical scholarship and their own experience, to support or challenge EBP when developing their own practices.

Research Methods

We conducted open-ended interviews followed by thematic analysis of the resulting data. This approach reflects our aim to help understand how practitioners in complex fields who are prepared to engage with EBP react to and use the information available from EBP clearing houses and what-works centers. Because we aim to understand the process as experienced and understood by the practitioners, it is important to hear what they say for themselves, in their own voices. Interviews thus provide the best opportunity to gain a range of perspectives in a number of different school settings across England. The interviews were conducted during the summer of 2014.

Our procedure was to approach potential interviewees with a brief introduction about the study. If the practitioner responded positively, then a convenient time was scheduled for the interview. A more detailed explanation of the study and a consent form were sent in advance of the interview date. Notes were taken during interviews. Specific ideas and direct quotations were checked with the interviewees a few weeks after they were interviewed to ensure that the transcribed data accurately relayed their perspectives. This also gave interviewees an opportunity

to clarify positions they had taken the interview. Some interviewees agreed to be recorded in order to improve the accuracy of the notes taken.

We planned each interview to be a semi-structured conversation about an hour in length. Interviewees were informed of the topics to be covered and were given a series of questions to think about in advance if they requested more details, but the interview itself was open-ended and the discussion guided, insofar as it was possible, by the interviewees. Interviewees were encouraged to discuss their views on the use of research evidence in their current school setting, and to compare the research findings to their previous experiences, if appropriate.

Interviewees sometimes struggled to schedule a whole hour for an interview, especially during a school day. As a result, while interviews were typically an hour long, a few were significantly shorter, occasionally as short as 20 minutes. In one site visit, conducting a group interview with three participants proved to be the best way of getting the widest possible range of perspectives.

The interview notes were analyzed for patterns and contrasting perspectives to establish if there were any commonly perceived challenges with using research evidence to make judgments about local contexts; and to determine how practitioners dealt with problems in understanding or implementing research evidence.

Participants

We used several approaches to find interviewees. We realized that speaking to scholars was not a significant priority for many of our target participants, and that a flexible approach to accessing interviewees was required. Social media offered one useful way of reaching out directly to people in the education sector who were interested in research evidence. We found several

potential interviewers commenting on education policy on Twitter and writing articles on blogs. We approached them via email, and several were willing to speak about the role of evidence in their practice and in the education sector as they saw it. Using snowball sampling following these initial contacts, we found other suitable candidates to interview. We also attempted to contact some schools that had published material on the use of evidence on their websites. This did not produce many responses, with only one teacher willing to be interviewed.

Social media revealed that some teachers had used connections forged over the Internet to establish their own workshops outside of the formalized structures of "school leadership" conferences and continuing professional development events. They were meeting separately from conferences involving government departments and NGOs. We attended one teacher-led event to gain some additional context about how practitioners use evidence when interacting with each other. This also allowed us to contact several more interviewees.

The resulting sample of 22 individuals connected to 12 separate school sites is unbalanced, as it is biased towards those engaged in practices and debates about research evidence in the classroom. In addition, given the open-ended, qualitative nature of our evidence base, our results should be considered speculative—an initial attempt to refine hypotheses. The participants included new teachers, teachers with a few years of experience, more experienced teachers with some management role, deputy heads, and head teachers. They were employed by a range of schools, including community primary and secondary schools, Academies, and one Free School. Among secondary-school teachers, subject specialties included science, history, English, design and technology, and information and communications technology.

Results and Discussion

Hypothesis 1: Generality and Local Contexts

A consistent theme expressed by the participants was the difficulty of generalizing from research evidence and of applying supposedly generally effective practices in local school contexts.

One teacher pointed out that summary research evidence tends to be vague, with the result that it is not clear how its implications deviate, if at all, from current common practice among teachers. This vagueness also means there is always a way for a skeptic of a particular approach to argue that it does not apply in context:

> The argument is how do you even use evidence in schools? How do you make evidence generalizable, because it's so messy? How do you take an intervention done in an entirely different context and apply it in your school? No one argues against evidence in principle but people contest evidence when it goes against their own experience and will contest it on context grounds. By the time you try to isolate those variables, you end up with very broad principles like feedback. Who is arguing we shouldn't give kids feedback?

Some teachers echoed this point, saying that classroom approaches with the strongest results in the toolkit were already quite widely accepted by teachers:

> We very much talk about metacognitive, we've talked about collaborative learning. Homework is a thing we are building on at the moment. There is nothing terribly new in the toolkit, it's all stuff that is known. Homework

has been argued about a lot in the past. It is well-regarded in the toolkit.

Some teachers were skeptical of drawing parallels between knowledge in the natural sciences and the social sciences, suggesting limits to experimental research evidence in education even if it were applicable in a clinical domain:

> One of the problems is that education research is ultimately social science, not like physics where we can definitely say we found something, or medicine where we've got a new compound we think does something and we do a double-blind trial. I don't think education results work in quite the same way.

> My personal view is its very hard to predict 100 per cent what the outcome is going to be. You can quite confidently predict the impact of some things, but there is always an aspect of uncertainty. Can you increase the likelihood? You have to be adaptive. If you are using ideas from universities (I get stuff from Twitter, huge resource base but that's not evidence-based really). How can I make it more reliable? I don't really know. It's one of the treasures of the education system that you are working with a group of people who you can't predict.

Another teacher, taking a more radically skeptical position, argued that the huge range and heterogeneity in contextual and supporting factors rendered experimental research evidence ultimately inapplicable to school practice:

> I love how trendy [randomized controlled trials are] but what does it tell you? You can't negate the impact of one teacher's charisma, or one teacher's bad day because they are getting a divorce. And how are you measuring progress? I don't think randomizing one method will ever give you generalizable results. I get to the stage where I start to think that kids aren't guinea pigs. There are things you can test in the human body and fix it. In education, there isn't one thing in different settings that you can reliably fix to get the same outcomes. It's not consistent in the way the human body can be. In schools, if you do different methods—is it the method? Or is it the teaching assistant? Is it about class, race, or gender in the school? Not sure you can close down the variables enough for it to be that [useful].

One teacher gave an example of how his practice and experience deviated (in his eyes) from research evidence. The reasons for the deviation were the sheer number of children on free school meals at their school (free school meals are a measure of social disadvantage that is now used to channel additional resources to schools in deprived areas). This meant that one-on-one tuition (ranked as comparatively expensive, given its impact in the EEF's toolkit) was both necessary and affordable in her school's context, compared with alternative approaches:

> The EEF toolkit suggested one-to-one support is expensive, [with only] moderate impact. But we have found it's high cost, high impact. [We can afford it] because we have a huge amount from the pupil premium. We narrow the gap between FSM and non-FSM. We've basically bridged the gap, five per cent either way in terms of stats. The EEF evidence does not equate to what we do in our school. [When it comes to] homework: the quality and the differentiated homework and students self-selecting homework can have a significant impact on progress. The problem here is that children [in this school] do not have safe home environments. . . . So, we are very cautious; yes, it's great there is a model—but you cannot use that to tarnish everyone with the same brush.

Teachers cited a wide variety of factors that will affect the outcome of an intervention or approach:

> It's OK learning about what works in different places. We then have to think how it would work in our own setting.

> The ethos of a school, how results driven, very large schools [would] have to be run differently from this school. Teaching style, subject—things that work in English won't work elsewhere. Policies already in place at the school. We use a whole set of recurring processes that students respond to, so if you put in a policy that doesn't link

> with those existing processes, they contradict and students quickly notice that. That won't be as successful. . . . An example might be how a behavior management system and praise and reward system . . . did not co-exist. If a praise policy didn't link into that, it didn't fit together nicely. Instead of being sanctioned for bad behavior, you get rewarded for good behavior, that could confuse well-behaved students who don't see it working for them.

One emphasized the importance of parent reaction to successful implementation:

> [There is] pressure from parents in certain schools. You have to be more flexible when you have a wide spectrum of abilities. You have to follow local and school ethos. E.g., does the school use setting and streaming? Some parents prefer one approach to another. If you ask students to do independent work at home, do they have the time, space and ability to do that at home? There will be a lot of things helping to decide whether to proceed with something or not. [With] middle class parents, you can try just about anything, they don't kick up a fuss. Here we have Asian parents and Muslim parents who would question the way we are changing things. They keep an eye on things.

A specialist school teacher described how dealing with some background factors, completely unnoticed by those without specialist knowledge or experience, could be crucial for allowing

some children with special educational needs to learn: “An uncomfortable child cannot learn. You could have an autistic savant in this room but who wouldn’t learn because this fan would be driving them mental.”

EBP’s focus on transparent methods that resist researcher bias and manipulation may be attractive and useful for policymakers, but the intended beneficiaries of research evidence find it difficult to apply effectively to their own practice, at least in the field of education in England c. 2014. Sometimes it is not always clear what the relevance of a general approach is to practice. At other times, EBP seems to be reiterating what teachers already know from their experience in the field, suggesting that widely touted evidence-based results may be driven by researcher priorities and debates rather than by their usefulness to intended beneficiaries.

Hypothesis 2: Wider Range of Resources

As a result of the challenge of applying generalized evidence to specific contexts, teachers discussed and drew on many alternative forms of research evidence and instruction for their own practice.

One example is “action research” (Reason and Bradbury 2013), often conducted as part of a Master’s degree in education. This research design typically involves a researcher-practitioner introducing a new approach and evaluating it using feedback such as a survey. It tends not to include an experimental element or systematic quantitative outcome measures, although it may include a before/after comparison. While some teachers found this approach to be particularly valuable, others were skeptical: “I would not take seriously studies with one teacher, one classroom. You have to take those with a pinch of salt.”

Even when not citing action research specifically, some teachers relied on informal small-scale trials inside their school to see if a particular approach was working: “You can come up with a thousand and one ways of improving an outcome but I always wonder what the evidence is. So, when a member of my team comes with an idea, I always think trial it first.” Perhaps reflecting our social-media search strategy for finding study participants, the Internet was cited repeatedly as a source of evidence, especially for accessing knowledgeable peers:

> We don’t engage with [research evidence]. [But] This stuff is changing, [I’m a] big fan of using Twitter—Twitter is a wonderful driver for self-sought CPD [Continuing Professional Development] —it’s the world’s biggest staffroom. One in four schools don’t have a qualified [subject specific] teacher but on Twitter you can find them.

Another common theme was the value of teachers engaging and discussing evidence for themselves rather than being passive recipients of research:

> Because I found out research for myself, there wasn’t the sense of shame or humiliation that happens when someone tells you are wrong. That’s why teachers need to be supported to engage with research evidence for themselves.

One teacher, involved in organizing teacher-led conferences, explained that she put a lot of emphasis on dialogue rather than top-down instruction and on the inclusion of teacher-led

seminars: “We want to see more collaboration between teachers. [It’s] not about academics coming in and saying, ‘This works, do this.’”

Another teacher, who was also involved in online discussion and informal teacher-led conferences, suggested:

> I would like to have communities that look at research critically, not to trash it because a lot is really interesting, but also think about the things that make it not so relevant to your setting. There is a real gap between theory and practice. We argue about what we are doing in the classroom all day long but we should debate the theory too.

Teachers that made use of EBP-configured research evidence tended to interpret it alongside other sources of evidence. Experts mentioned by name included Carol S. Dweck (2006), John Hattie (2009), Debra Kidd (2014), Daniel T. Willingham (2009), and Dylan Wiliam and Paul Black (1998). One teacher emphasized the importance of going to the source academic material, suggesting in particular a need to understand the mechanism through which an intervention is supposed to work:

> [The EEF provides a] very helpful introduction but not enough information to design a feedback policy. Really, I need to look at the studies that have formed the meta-analysis and remember the flaws in meta-analysis. Take John Hattie, [who] aggregated primary and secondary home-work policy into one effect size. [The toolkit] doesn’t provide all the answers, which is

> fine, it's not its role. . . . You have to tailor it to your classroom. This is where reading the research is important. You have to understand that underlying rationale.

Other teachers argued that it is necessary to have some grounding in theoretical frameworks, which they conceptually distinguished from general evidence of efficacy:

> Research done beforehand would lend some weight to it. E.g., the Hattie research, meta-analysis that he carried out. [You give that] somewhat more credence than something mentioned in a [CPD] course. The question then is, does someone really understand what they are doing. Or if it is x [intervention], they have some idea of x, but by the time you have applied it in context, it's y. It's important that you understand what x was and why you are doing y now.

> We are pushed to be teachers as researchers, when really, we should be teachers as scholars. There is a theoretical aspect to teaching practice, not just classroom practice. As rounded practitioners, we should be thinking about all aspects of our work. A lot of people haven't read a lot of education theory, and we need to have both research and theory to read and understand. We need that in order to navigate around the various trials and say, "Well this one is no good because they did this and gives me a load of figures that mean nothing in my context or whatever." Just because

> someone has done an education degree from 18-21 [does not mean they have] a lot of experience of decoding research.

Some teachers favored particular academic pedagogies and had developed affinities for particular ways of understanding classroom practice and particular authors:

> I use Dylan Wiliam's book, *Assessment for Learning*, as a bible . . . what he gives us, modified by Daniel Kahneman's work, is essentially a theory of learning [which], unlike Vygotsky and Piaget, is about how teachers can train the mind in practice, the importance of retention and focus. This helps you design a lesson to aid retention. Hattie fits into a pattern of ideas that is confirmed by other people. For example, formative assessment fits with Paul Black and Dylan Wiliam.

Others were more eclectic, drawing on a variety approaches for their own practice: "For any Willingham or Hirsch, you have some counter-evidence. Even this [useful approach] is dangerous. [It's] nice to look at different aspects of things."

For some teachers, interpretation of evidence was linked to debates within the teaching profession, sometimes conducted over social media. A division between "progressive" and "traditional" education was mentioned and discussed. Teachers rarely identified themselves as being in one or the other ideological camp, suggesting instead that they drew on both approaches in their own practice. However, the debate itself seemed to be an important source of framing for evidence-based research and the underlying rationale of different approaches. The debate also

represented a source of motivation for further engagement with research evidence, and it provided a means for filtering research. When "allies" (or indeed opponents) cited evidence in a debate, it might be sought out and analyzed for corroborating or rebutting claims. Some teachers were encouraged to engage, in particular, with recent research in cognitive psychology as a result.

In sum, teachers were interested in understanding the underlying mechanism through which an intervention is supposed to work. In developing a mental picture of such mechanisms, they drew on wider frameworks that fit their experience, intuitions, and values. This included some underlying ideological presuppositions, although not in a sense that necessarily correlates to an explicit political ideology. This diversity of values and underlying frameworks means that the validity and reliability of research evidence remained open to contestation among teachers. In practice, teachers re-introduced aspects of research evidence that formal EBP criteria excluded on the grounds that they lacked rigor because without them, the evidence was impossible to interpret or apply.

* * *

The nature of our empirical evidence, based on a non-random sample of participants and thematic analysis, means that our conclusions are necessarily tentative. But for both theoretical and empirical reasons, we believe that it is reasonable to expect a tension between the perspectives of researchers, policymakers, and field professionals. We suggest that this tension indicates the difficulty of developing generally applicable policy approaches even on the basis of many high-quality experimental studies. Theoretically, such studies inevitably raise questions

about how relevant "evidence" comes to be defined, interpreted, and applied. Perhaps surprisingly, practitioners in the field recognize these questions and reach different conclusions about evidence-based research as a result.

For policymakers, EBP seems to present a way of deciding the effectiveness of a particular policy or intervention. Yet there may not be a determinate answer to the question of "what works," or at least not one that is generalizable to the scale at which the policymakers in central government are required to operate. As a result, "what works?" cannot perfectly replace the more overtly normative question of "who decides?"

REFERENCES

Alexander, Danny, and Oliver Letwin. 2013. "What Works: Evidence Centres for Social Policy." http://dera.ioe.ac.uk/17396/1/What_Works_publication.pdf (September 4, 2014).

Ayer, Alfred Jules. 1972. *Probability and Evidence*. New York: Columbia University Press.

Behavioural Insights Team. 2011. "Behavioural Insights Team Annual Update 2010–11." London: Cabinet Office. http://casaa.org/wp-content/uploads/Behaviour-Change-Insight-Team-Annual-Update_acc.pdf (May 8, 2017).

Behavioural Insights Team . 2012. *Applying Behavioural Insights to Reduce Fraud, Error and Debt*. London: Cabinet Office. www.gov.uk/government/uploads/system/uploads/attachment_data/file/%20%2060539/BIT_FraudErrorDebt_accessible.pdf (October 2, 2014).

Biesta, Gert. 2007. "Why 'What Works' Won't Work: Evidence-Based Practice and the Democratic Deficit in Educational Research." *Educational Theory* 57(1): 1–22.

Black, Paul, and Dylan Wiliam. 1998. "Assessment and Classroom Learning." *Assessment in Education: Principles, Policy & Practice* 5(1): 7–74.

Blunkett, David. 2000. “Influence or Irrelevance: Can Social Science Improve Government?” *Research Intelligence* (71): 12–21.

Boaz, Annette, Deborah Ashby, and Ken Young. 2002. "Systematic Reviews: What Have They Got to Offer Evidence Based Policy and Practice?" ESRC UK Centre for Evidence Based Policy and Practice, London.

Bovens, Luc. 2009. “The Ethics of Nudge.” In *Preference Change: Approaches from Philosophy, Economics and Psychology*, ed. Till Grüne-Yanoff and S. O. Hansson. Berlin: Springer.

Callaghan, James. 1976. “A Rational Debate Based on the Facts.” Paper presented at Ruskin College Oxford.

Cartwright, Nancy. 2007. “Are RCTs the Gold Standard?” *BioSocieties* 2(1): 11–20.

Cartwright, Nancy, and Nick Cowen. 2014. "Making the Most of the Evidence in Education: A Guide for Working Out What Works . . . Here and Now." Working paper, Durham University.

Cartwright, Nancy, and Jeremy Hardie. 2012. *Evidence-Based Policy: A Practical Guide to Doing It Better*. Oxford: Oxford University Press.

Cartwright, Nancy, and Eileen Munro. 2010. “The Limitations of Randomized Controlled Trials in Predicting Effectiveness.” *Journal of Evaluation in Clinical Practice* 16(2): 260–66.

Chitty, Clyde, and J. E. Dunford, eds. 1999. *State Schools: New Labour and the Conservative Legacy*. London: Woburn Press.

Coe, Robert. 2004. “What Kind of Evidence Does Government Need?” *Evaluation & Research in Education* 18(1–2): 1–11.

Crafts, N. 1996. “'Post-Neoclassical Endogenous Growth Theory': What Are Its Policy Implications?” *Oxford Review of Economic Policy* 12(2): 30–47.

Crook, David. 2002. “Educational Studies and Teacher Education.” *British Journal of Educational Studies* 50(1): 57–75.

Deaton, Angus, and Nancy Cartwright. 2016. "Understanding and Misunderstanding Randomized

Controlled Trials." NBER Working Paper No. 22595. Cambridge, Mass.: National Bureau of Economic Research.

Dweck, Carol S. 2006. *Mindset: The New Psychology of Success*. New York: Random House.

Education Endowment Foundation. 2016a. "Early Literacy Approaches/EY Toolkit Strand." https://educationendowmentfoundation.org.uk/resources/early-years-toolkit/early-literacy-approaches.

Education Endowment Foundation. 2016b. "Implementation." https://educationendowmentfoundation.org.uk/our-work/implementation.

Every-Palmer, Susanna, and Jeremy Howick. 2014. "How Evidence-Based Medicine Is Failing Due to Biased Trials and Selective Publication." *Journal of Evaluation in Clinical Practice* 20(6): 908–14.

Fleischhacker, W. Wolfgang. 2017. "A Meta View on Meta-Analyses." *JAMA Psychiatry* 74(7): 684.

Goldacre, Ben. 2013. "Building Evidence into Education." London: Department for Education http://media.education.gov.uk/assets/files/pdf/b/ben%20goldacre%20paper.pdf (April 17, 2014).

Goodwin, Tom. 2012. "Why We Should Reject 'Nudge.'" *Politics* 32(2): 85–92.

Gorard, Stephen. 2015. "The Uncertain Future of Comprehensive Schooling in England." *European Educational Research Journal* 14(3–4): 257–68.

Greenhalgh, Trisha, Jeremy Howick, and Neal Maskrey. 2014. "Evidence Based Medicine: A Movement in Crisis?" *BMJ* 2014; 348:g3725

Halpern, David. 2013. "Preface." In David Gough, Sandy Oliver and James Thomas, *Learning from Research: Systematic Reviews for Informing Policy Decisions: A Quick Guide*. EPPI-Centre, Social Science Research Unit, Institute of Education, University of London.

Halpern, David. 2016. *Inside the Nudge Unit: How Small Changes Can Make a Big Difference*. London:

Ebury Press.

Hammersley, Martyn. 2013. *The Myth of Research-Based Policy & Practice*. Los Angeles: SAGE.

Harper, Hugo. 2013. *Applying Behavioural Insights to Organ Donation: Preliminary Results from a Randomised Controlled Trial*. London: Behavioural Insights Team.

Hattie, John. 2009. *Visible Learning: A Synthesis of over 800 Meta-Analyses Relating to Achievement.* London: Routledge.

Higgins, Steve, et al. 2016. *The Sutton Trust-Education Endowment Foundation Teaching and Learning Toolkit.* London: Education Endowment Foundation.

Holmes, Dave, Stuart J. Murray, Amelie Perron, and Genevieve Rail. 2006. "Deconstructing the Evidence-Based Discourse in Health Sciences: Truth, Power and Fascism." *International Journal of Evidence-Based Healthcare* 4(3): 180–86.

Hume, David. 2008. *An Enquiry Concerning Human Understanding*, ed. Peter J. R. Millican. Oxford: Oxford University Press.

Hunter, D. J. 2009. "Relationship between Evidence and Policy: A Case of Evidence-Based Policy or Policy-Based Evidence?" *Public Health* 123(9): 583–86.

Kahneman, Daniel. 2011. *Thinking, Fast and Slow*. London: Penguin Books.

Kidd, Debra. 2014. *Teaching: Notes from the Front Line*.

Liberati, Alessandro et al. 2009. "The PRISMA Statement for Reporting Systematic Reviews and Meta-Analyses of Studies That Evaluate Health Care Interventions: Explanation and Elaboration." *PLoS Medicine* 6(7): e1000100.

Loewenstein, George, et al. 2012. "Can Behavioural Economics Make Us Healthier?" *BMJ* 2012;344:e3482

Machin, Stephen, and Sandra McNally. 2008. "The Literacy Hour." *Journal of Public Economics* 92(5–6): 1441–62.

Mackie, John L. 1965. "Causes and Conditions." *American Philosophical Quarterly* 2(4): 245–64.

Mansell, Warwick. 2007. *Education by Numbers: The Tyranny of Testing*. London: Politico's.

Marmot, M. G. 2004. “Evidence Based Policy or Policy Based Evidence?” *BMJ* 328(7445): 906–7.

McAnulla, Stuart. 2010. “Heirs to Blair's Third Way? David Cameron's Triangulating Conservatism.” *British Politics* 5(3): 286–314.

McCormack, Brendan, et al. 2002. “Getting Evidence into Practice: The Meaning of `Context.'” *Journal of Advanced Nursing* 38(1): 94–104.

Morrison, Keith. 2001. “Randomised Controlled Trials for Evidence-Based Education: Some Problems in Judging 'What Works.'” *Evaluation & Research in Education* 15(2): 69–83.

Noonan, Eamonn, and Arild Bjørndal. 2010. “The Campbell Collaboration.” In *Cochrane Database of Systematic Reviews*, ed. The Cochrane Collaboration. Chichester, UK: John Wiley & Sons.

Puttick, Ruth, and Geoff Mulgan. 2013. “What Should the 'What Works Network' Do?” http://www.nesta.org.uk/sites/default/files/what_should_the_what_works_network_do_0.pdf

Quine, W. V. O., and J. S. Ullian. 1978. *The Web of Belief*, 2nd ed. New York: McGraw-Hill.

Reason, Peter, and Hilary Bradbury, eds. 2013. *The SAGE Handbook of Action Research: Participative Inquiry and Practice*, 2nd ed. London: SAGE.

Sanders, Michael, and David Halpern. 2014. “Nudge Unit: Our Quiet Revolution Is Putting Evidence at Heart of Government.” *The Guardian*, 3 February.

Seckinelgin, Hakan. 2016. *The Politics of Global AIDS: Institutionalization of Solidarity, Exclusion of Context*. Berlin: Springer.

Simpson, Adrian. 2017. “The Misdirection of Public Policy: Comparing and Combining Standardised Effect Sizes.” *Journal of Education Policy* 32(4): 450–66.

Slavin, R. 2008. “Perspectives on Evidence-Based Research in Education: What Works? Issues in Synthesizing Educational Program Evaluations.” *Educational Researcher* 37(1): 5–14.

Slavin, R., and D. Smith. 2009. “The Relationship Between Sample Sizes and Effect Sizes in Systematic Reviews in Education.” *Educational Evaluation and Policy Analysis* 31(4): 500–506.

Thaler, Richard H., and Cass R. Sunstein. 2009. *Nudge: Improving Decisions about Health, Wealth and Happiness*. London: Penguin Books.

de Waal, Anastasia. 2006. *Inspection, Inspection, Inspection!: How OFSTED Crushes Independent Schools and Independent Teachers*. London: Civitas.

de Waal, Anastasia, ed. 2008. *Inspecting the Inspectorate: Ofsted under Scrutiny*. London: Civitas.

Waters, Elizabeth, et al. 2006. "Evaluating the Effectiveness of Public Health Interventions: The Role and Activities of the Cochrane Collaboration." *Journal of Epidemiology & Community Health* 60(4): 285–89.

Whetton, Chris. 2009. "A Brief History of a Testing Time: National Curriculum Assessment in England 1989–2008." *Educational Research* 51(2): 137–59.

White, Howard. "Just Keep Testing: Five Principles for Evidence-Based Policy and Practice." *Campbell Collaboration*. https://www.campbellcollaboration.org/blog/just-keep-testing-four-principles-for-evidence-based-policy-and-practice.html

Willingham, Daniel T. 2009. *Why Don't Students like School? A Cognitive Scientist Answers Questions about How the Mind Works and What It Means for the Classroom*. San Francisco: Jossey-Bass.

Woods, Philip A., Glenys J. Woods, and Helen Gunter. 2007. "Academy Schools and Entrepreneurialism in Education." *Journal of Education Policy* 22(2): 237–59.

Wyse, Dominic, and Harry Torrance. 2009. "The Development and Consequences of National Curriculum Assessment for Primary Education in England." *Educational Research* 51(2): 213–28.

Young, Ken, Deborah Ashby, Annette Boaz, and Lesley Grayson. 2002. "Social Science and the Evidence-Based Policy Movement." *Social Policy and Society* 1(03): 215–24.