**Disagreement about Evidence-Based Policy**

Nick Cowen[1]
School of Social and Political Sciences, University of Lincoln

Nancy Cartwright
Department of Philosophy, Durham University

Written for the *Routledge Handbook of The Philosophy of Disagreement* edited by Maria Baghramian, J. Adam Carter and Richard Rowland

Evidence based-policy (EBP) is a popular research paradigm in the applied social sciences and within government agencies (Davies, 2012). Informally, EBP represents an explicit commitment to applying scientific methods to public affairs, in contrast to ideologically-driven or merely intuitive "common-sense" approaches to public policy. More specifically, the EBP paradigm places great weight on the results of experimental research designs, especially randomised controlled trials (RCTs), and systematic literature reviews that place evidential weight on experimental results (Boaz et al., 2002; Young et al., 2002). One hope is that such research designs and approaches to analysing the scientific literature are sufficiently robust that they can settle what really 'works' in public policy.

EBP offers a tantalising answer to a key question in the philosophy of disagreement: what to do when experts disagree (Frances, 2014: 40). It appears as a promising solution to a challenging position that government policymakers often face when dealing with experts, especially when they are not perfectly trusted by the government authority (Cowen, 2019). We can characterise this as an example of what Goldman (2001: 86) terms the 'novice/expert problem' that arises when a novice struggles with a second-order problem of establishing who is really an expert to be relied on in a subject area without having first-hand knowledge of that subject area. Plausible epistemic strategies include evaluating the capacity of experts to answer and critique each other's arguments, weighting by the number of experts, appraising the credential of the experts, considering personal biases and interests, and evaluating the experts' past track-record (Goldman, 2001: 93).

Like everyone else, policymakers rely on a mix of all these strategies. The EBP solution is to require proponents of a policy (they could be experts but they do not need to be) to support their position using empirical evidence that is transparently generated and collected according to a relatively simple, standardised experimental method, with the results open to interpretation by

[1] This work was generously supported by the Institute for Humane Studies' Discourse Initiative grant.

non-experts. Experts must point to firm evidence accessible to generalists, at least to those trained in the basics of EBP. If the evidence does not point in their favor, their claims can be reasonably rejected.

Can EBP succeed in displacing reliance on domain-specific expertise? Can EBP generate evidence that a rational agent *should* accept as sufficient warrant to justify a policy or decision? On our account, this is seldom, if ever, the case. The key reason for this is that underlying this approach is generally an appeal to argument by induction, which always requires further assumptions to underwrite its validity, and if not induction, some other argument form that also requires assumptions that are very often not validated for the case at hand. As indispensable as informal inductive evidence is for everyday decision-making, such evidence cannot support further conclusions without additional assumptions that need to be warranted and thus has limited applicability in policy. How can people establish what conclusions a piece of experimental evidence helps to warrant where implementation of a policy is being considered? Only through appeal to broader knowledge, including established theories about the underlying properties and propensities of physical and social entities. This knowledge supplies the additional premises required to establish the value and relevance of a piece of experimental evidence for a case of interest. Yet, these further knowledge claims are often the source of the same disagreement that prompted the recourse to EBP in the first instance. Our conclusion is that reasoned disagreement is an inevitable outcome of serious reflection on the complex phenomena one finds in social policy.

We make our case as follows. First, we outline the case for evidence-based policy in broad but, we hope, fair terms, drawing particular attention to the imperative EBP proponents promote to avoid bias from unobserved factors. Second, we point out the weaknesses with EBP taken purely from the standpoint of avoiding this sort of bias that EBP proponents prioritise for attention. Third, we pan out to reveal that EBP neglects the sorts of research questions that can be highly consequential for policy efficacy.

**The case for evidence-based policy**

The label 'evidence-based policy' appeals to the widespread notion that our beliefs and actions should reflect and respond to what we can learn about the world. The specific practice of EBP as it has emerged in many policy settings consolidates this commitment in a way that attempts to tie policy actions to knowledge about what a policy action will achieve in a very direct fashion. Hence, a popular phrase that EBP proponents use is 'what works?' which characterises their

interest as being research about the effectiveness of interventions above other potential questions (What Works Team, 2018).

A key premise is that we cannot know with sufficient warrant what the result of a particular policy will be unless the policy and its outcomes have been observed empirically in the field and transparently recorded (Cochrane, 1972: 22). In normal situations it is hard to disentangle the role of other background factors as well as individual variations from the contribution of a specific policy intervention. The worry is that a positive correlation between the policy and the desired outcomes may not indicate a causal connexion between them since the correlation may instead be due to the operation of other factors not taken into account. These are commonly called *confounding factors*. If we knew exactly what the confounding features were, we could take them into account in our causal inferences. But what these are is often a highly contested matter, and it is sometimes agreed that we really have very little knowledge about what they are. Thus, it has been argued in EBP, the only way to be sure of the effect of a policy intervention is through an experiment that allows the researcher to compare phenomena with and without the intervention in otherwise similar environments, where the similarity of these environments is vouchsafed by study design alone and does not depend on contested assumptions about what is and is not a confounding factor. By far the preferred kind of evidence for proponents of EBP is then the *randomised controlled trial* (RCT): RCTs are characterised as the 'gold standard' of research designs against which other forms of research are measured (Goldacre, 2013; Sanders and Halpern, 2014; cf. Cartwright, 2007). In an RCT, a policy is randomly allocated to one set of cases as a 'treatment' while another set functions as a 'control'. Those in the control group are given a different intervention or observed as a 'business as usual' scenario. The advantage of this research design, presuming, of course, it is properly conducted, is that any significant average variation between the outcomes of the treatment and control groups is likely to be *caused* by the action of the treatment or intervention. In other words, it is a direct experimental test of the change that a particular policy brings about.

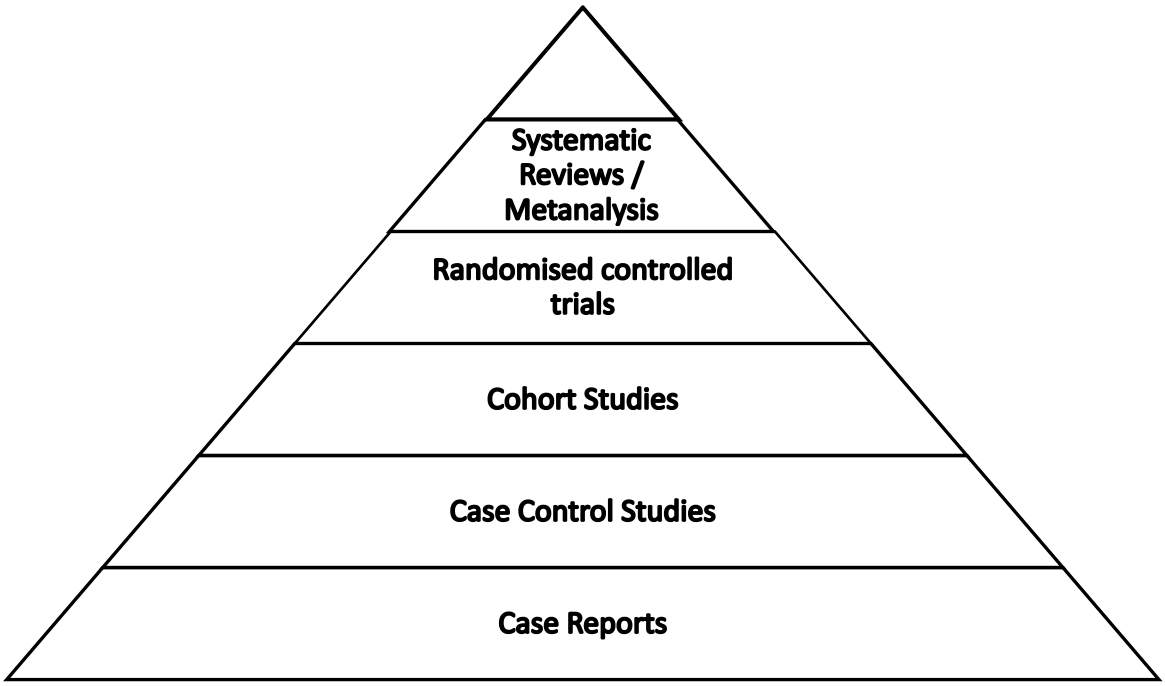
The classic claim among proponents of EBP is that well-conducted RCTs have the highest levels of internal validity (Siepmann et al., 2016). That is, they are supposed to measure more precisely the average effect of an intervention within a study than any alternative research design. The reason for this is that the process of randomising cases to treatment and control groups should by design alone remove all sources of systematic difference in the distribution of confounding factors in the treatment and control groups, at least at the time of randomization. This leaves an average effect size that can be estimated within reasonable margins of error so long as you have a

large enough number of cases (and so long as no systematic differences in the distribution of confounding factors between the treatment and control populations occur post randomization).[2]

In terms of value for the real-world, this approach faces the next hurdle: external validity (Bracht and Glass, 1968). This is how well the outcomes of the treatment in the observed cases transfer to unobserved cases. The prevailing answer to this challenge is to do repeated experimental studies of the intervention in a variety of different settings and contexts, and then synthesise the average effects of all the study results into an overall summary result. These 'studies of studies' are called *systematic reviews*. Critically, what studies are included or excluded from these reviews of the evidence can be very consequential for the overall average result. The concern is that the process of reviewing the literature can leave many opportunities for researchers to introduce biases (possibly to favour policy interventions they happen to think are better) through choices about how it is covered. To combat this, the major 'what works' clearing houses publish standard protocols for conducting systematic reviews that involve specifying precise search terms, the databases to be searched, and the inclusion criteria for studies, ideally before any significant search of the literature has been undertaken (Liberati et al., 2009).

The overall result is that many EBP publications and discourse offer a strikingly hierarchical conception of the quality of research methods. For example, variations on a 'pyramid' depicting levels of evidential rigour can be found dispersed both informally across evidence-based discourse and in textbooks and academic articles (Farrington et al., 2002; Murad et al., 2016). At the top of any pyramid will be systematic reviews with randomised controlled trials sitting just beneath, with subsequent studies ranked according to the closeness with which they approach the experimental RCT design. The central message is that the more capable a research design is at separating correlation from causation, as well as reducing opportunities for researchers to influence findings through the various choices they make while undertaking a study, the greater the weight you can give it.

[2] Note that random assignment does not guarantee that confounding factors are distributed between treatment and control in any single run of the experiment, even at the time of assignment. What it does guarantee is that (supposing no further systematic differences are introduced after assignment) the difference between outcome average in the treatment group and in the control group is an unbiased estimate of the true average treatment effect, where unbiased estimates have to do with what happens under indefinite repetitions of exactly the same experiment.




**Weaknesses with EBP**

We begin our critical assessment by considering the problems with EBP when taken on its own terms. As we have described, EBP's ranking of studies elevates research designs that eliminate (or at least reduce) one source of bias: uncontrolled or unobserved factors that explain a correlation between an independent variable (the intervention) and dependent variable (the outcome policymakers want to change). The greatest concern expressed by EBP proponents is to avoid positing causation from correlation, and RCTs are taken to be the most refined way of avoiding having to make that argumentative leap. The trouble is well known: no amount of certainty about what happens in one site can assure us about what happens in another. The population of interest (where a policy is supposed to be implemented) is seldom the same as the population that is in the various experimental studies that make up a systematic review.

On what grounds can policymakers conclude that the studied population is representative of the population of interest? While there are exceptional studies, generally the study population is likely to be selected in a very different way from the proposed implementation of a policy to a whole population of interest. Recruitment for individual-level studies is notoriously difficult; finding a large enough sample comes from trial, error, and luck. Potential study participants are typically accessed in ways that amount to convenience sampling. Researchers identify people or organisations that are already known to them or have a pre-existing relationship with their associated institutions. Even when studies are promoted more broadly, the eventual participants

will likely differ systematically from the targeted population. For example, they will be the sort people who are more responsive to study invitations and participant incentives.

Many proponents of EBP are alert to these sorts of concerns and have attempted to ameliorate them as best they can. For example, a now preferred standard in RCTs is to adopt an intention-to-treat principle whereby study results include people who fail to follow the initial protocol (for example, by dropping out of the study part way through its implementation) (Gupta, 2011). This reduces one source of bias: the population sampled being more likely to be consistently engaged in treatment compared to the target population. However, this cannot adjust for the people who never engage with researchers at all.

Moreover, in some projects, researchers deliberately exclude a range of people in the population of interest from participation. This might be for practical considerations, such as concern about certain types of participants being less likely to complete the study; or it might relate to ethical concerns (such as excluding young people, the elderly or the mentally less capable) when fully informed consent to participate in a study is hard to verify. Yet the population of interest (where the intervention will eventually be practiced) might well include these kinds of individuals.

When it comes to field trials that involve testing an intervention in a location or community, it is similarly rare that locations are selected randomly or in a way that plausibly achieves representativeness. The more common practice is to establish a set of potential locations based on accessibility and trust between researchers and relevant local communities and municipal governments. Only at that point, are treatment and controls randomly allocated.

Why do all these practical challenges for sampling participants and study locations matter? The issue is that these stages in research design and conduct introduce many potential sources of bias. Individuals or areas that turn out to be more conveniently selectable for studies are likely to differ systematically on a multitude of unobserved factors from those that are less selectable. As we have seen, the solution that EBP proponents offer is to replicate the experiment in many different settings, including different demographics, environments (e.g. urban and rural), countries, and regimes.

The theoretical flaw in this approach is that it relies on induction (Reiss, 2019: 3106). Any conclusions drawn from this piling up of studies based on sampling in different settings is as defeasible as claiming that "all swans are white" based on repeated observations of swans even if the observations happens in multiple locations (Cowen et al., 2017: 273). We do not mean to suggest though that induction from a set of instances to a general conclusion is never valid. It's just that piling up the numbers is far from sufficient. That's what the case of the swans in Sydney

Harbour is meant to illustrate. What must be always added if the inference is to be credible is the assumption that the targeted setting is the same as the tested settings with respect to everything necessary to afford the same connexion between intervention and outcome. No inductive conclusions can be more certain than the certainty with which this crucial additional assumption is established. In the nicest cases, where we have a great deal of very certain knowledge, the numbers don't really matter. For instance, suppose we develop a new experimental design to measure more precisely than so far possible the charge on an electron. Here because we have already established that all electrons have the same charge, we can in principle do an induction from a sample of 1—though we may decide to measure more because there is a chance the device is not working properly (i.e., because we worry about a failure of internal validity!).

Practically, when it comes to policy interventions, the additional assumption necessary to underwrite an inductive inference can be highly suspect. Study sites (even those selected with the aim of testing in a variety of contexts) will often possess some hidden, uncontrolled factors that correlate with their accessibility to researchers, making them unrepresentative of the population of interest. This inductive approach based on conducting an RCT for a promising intervention, then repeating and attempting to replicate the results in different contexts is exposed to two common types of error. First, interventions that appear to work within the tested contexts might turn out not to work when implemented more broadly. Second, negative or disappointing results in a few initial RCTs may lead researchers to reject, as lacking in evidence, interventions that might turn out be highly successful in contexts or configurations that have yet to be tested (Cowen and Cartwright, 2019).

None of these concerns subtract from the theoretical value, and sometimes great practical relevance, that the results of RCTs have on their own terms and especially when taken in combination with other evidence. What they do instead is establish that there exist important trade-offs. For example, a study based on analysis of administrative data that tracks outcomes following a policy being implemented might not include a randomisation element. There will not be a control group that is directly comparable to the treatment group. On the other hand, if the administrative data is collected unobtrusively as part of the ordinary practice of service provision within the field, then the sample might be much more likely to be representative of the target population than when collected as part of an RCT. Such a study could be more biased in the sense that it cannot control as effectively for unobserved factors within the sample, but might be less biased when applied to real world cases of interest. In this sense, the stepwise connection between so-called internal validity and external validity that EBP proponents presume in theory might not hold in practice. A research design might, in fact, be able sacrifice some degree of

internal validity yet offer results that have more real-world application (external validity in the parlance of the experimentalists).

Presuming for the moment that the main currency of research quality in comparative statistical studies of policy effectiveness is unbiasedness, as the standard justifications for the emphasis on RCTs imply, then opting to conduct RCTs, or weighting the results of RCTs heavily in a survey of the literature, may turn out to be costly in that same currency. It can involve accepting more biasedness overall to reduce bias in one part of the research process.

**The broader limits of 'what works'**

So far, we have pointed out some weaknesses with a narrowly inductive, experimentalist approach to EBP when taken purely on its own term. Now, we challenge the broader basis for conceptualising effective research for policy as simply 'what works'.

*Context matters*

In essence, 'what works?' focused research can easily glide over other important research questions such as 'what's going on?', 'what's going wrong?' and 'how does this work?' These address the background context in which some sort of intervention is being considered (what is sometimes characterised as 'business as usual' in the EBP framework), the diagnostic process of figuring out why there is a problem and why existing policies have disappointed as well as the details (or mechanism) of how a novel intervention (at least in this particular context) is supposed to work.

The EBP framework pushes these questions towards the periphery of an inquiry by characterising policies as discrete 'interventions'. These are essentially standardised packages of practices or activities that, in social policy, normally involve deploying people trained to carry out a programme. This reflects EBPs origins in clinical medicine where the classic 'intervention' is administrating a drug or performing a procedure. The purpose of aggregating the results from multiple contexts into an average is to figure out whether some intervention has a generally positive effect on a situation or ailment. The problem is that an average effect size of a particular action is often not meaningful or useful for action without further parameters and description. If we strike a match in a living room, it may light up. If we strike it in the same room where there is a gas leak, it may cause an explosion. If we strike the match outside in the wind, it may not light at all. What does striking a match do 'on average'? Such a question does not really make sense. The EBP framework is premised on their being something like a central tendency for every intervention that is instantiated, in some sense, whenever it is implemented (Cowen and Cartwright, 2019). However, such a tendency often does not exist. Rather, an action or

intervention uses a mechanism that achieves or changes something in some contexts but not at all in others (or does something completely different). Moreover, for most policies there are many intervening steps that must occur in the right way between the implementation of the policy and the end result if the final intended outcome is to be achieved. So what matters is not *the* mechanism by which the policy produce the intended outcome but rather the host of mechanisms that operate in between: the series of different mechanisms by which each step leads to the next.

Although rarely stated in such terms, modern clinical medicine is based on a strong background understanding of what we might call the 'business as usual' of the human body. Medical science draws on a range of scientific knowledge, including biology, anatomy and physiology. This knowledge is derived not from studies directly aimed at testing an intervention but rather research aimed at understanding the processes that make up human functioning. This includes laboratory-controlled studies that identify mechanisms of action in deliberately artificial environments. The overall result of this research that uses a variety of research designs and modes of inquiry is that the medical profession has developed a good idea of how human bodies work when healthy, as well as the expected healthy variations between them, such as how children, adults and elderly people systematically differ.[3] Moreover, doctors learn and study pathology, that is knowledge of the diseases and ailments that present in patients. This means medical practitioners can make diagnoses and prognoses: they can figure what is causing a set of symptoms and predict the course of a disease without some sort of clinical intervention or lifestyle change. Thus, medical science and effective clinical interventions rely on knowledge of 'what's going on' and of 'what's going wrong' and not just of 'what works'. In medicine, RCTs of specific interventions are not a way of avoiding disagreements within biological science. Rather their success is a result of having a great deal of background knowledge established through settling disagreements over centuries of research.

*Social impacts*

The imperative to measure the impact of an intervention against a pre-specified control group constrains the research questions and research designs that are tractable within the EBP paradigm. To produce interpretable results, research designs require a reasonably large number of cases that are plausibly separable from the control group and where the intervention can be

[3] Holdcroft (2007) identifies critical gaps in this knowledge of human variation, especially between ethnicity and gender as some traditional medical knowledge has implicitly treated the archetype subject of clinical intervention to be white, cis and male. For our argument, this serves to accentuate the importance of understanding the background context in which interventions are proposed.

randomly allocated. This typically limits the sort of units of analysis that are can be tested. Often these units will be individuals or, at best, relatively small groups such as family households or classes in schools. Moreover, the main treatment effect that is to be measured must impact the treatment group and not have substantial spill overs, especially on the control group. The result is that treatments are conceptualised mainly in individualistic or atomistic terms (Hope, 2005).

Yet, many policies aim to improve wellbeing not through directly impacting a set of individuals but through altering the shared environment to produce a positive impact on a local community and sometimes beyond. For example, the benefits of a new public transport amenity could improve individual welfare through several kinds of mechanisms, for example by improving journey times, reducing fuel costs, and thereby increasing attendance at school and expanding available employment opportunities reachable within a commuting distance. Moreover, many of the benefits do not relate to personal use of the amenity as such. More accessible public transport reduces congestion on roads for people using private transport and ameliorates environmental pollution improving the health of everyone in the community whether they use transport or not. Testing such an intervention against a control group, whether randomised or not, would be practically impossible to implement and would inevitably fail to assess the policy along the various dimensions along which it is intended to work.

Although this issue is clear when dealing with interventions where the main effects are supposed to be public, it continues to be a problem even for interventions where the key mechanism is through the individual. For example, interventions that are intended to help individual students achieve more learning or for convicted offenders to become rehabilitated might have a primary action on individuals. Yet, in these policy areas, the impact of peers (both positive and negative) is also well-established. Artificially segregating and allocating a treatment to one group of individuals, but a group that inevitably interacts with a wider community, can thus miss out important real-world outcomes.

As we have already seen, the origins of this focus on individual effects lies in the emergence of EBP from clinical medicine, where the main paradigm intervention is generally something that acts on an individual body (such as the administration of a drug) where the mechanism is presumed not to have spill-over effects on others. Yet, this emphasis on individual-level clinical interventions has problems in medicine as well. The Covid-19 pandemic highlighted the importance of the social impact of health interventions especially with regards to public health. Critical questions, such as the extent to which mask-wearing and vaccination reduces the transmission of infection within a community are effectively impossible to answer through an

RCT or through a research design that attempts to mimic an RCT as closely as possible. This lack of evidence that fits the paradigm of rigorous research led several proponents of EBP initially to underweight the value of non-pharmaceutical interventions (NPIs) like masks. For example, a single RCT, conducted in Denmark, that purported to show that interventions encouraging the wearing of masks were ineffective (Bundgaard et al., 2021) was used as grounds to dismiss NPIs in general. While there is still plenty of room for reasonable disagreement over mask-wearing, such a narrow appreciation of what high-quality research constitutes means that an enormous evidence-base more favourable to masks can be neglected on a poorly reasoned basis (Howard et al., 2021).

*Individualist interventions*

A problem with reifying a sub-set of research designs over others is that it channels research efforts towards developing and answering questions that those research designs are most capable of answering. Only coincidentally will these questions be the most consequential or urgent for policy. For example, behavioural interventions, commonly called 'nudges', are one substantial area of interest for researchers working within the field of EBP. Indeed, in the United Kingdom, the government-backed 'what works network' grew out of Downing Street's Behavioural Insights Teams (called 'the nudge unit'). The premise of such interventions is that public policy can correct for individual behavioural biases or cognitive inattention that produces bad individual and social outcomes (Chater and Loewenstein, 2022). Practically, nudges are often constituted by strategically-timed communications such as letters, e-mails, phone calls and text messages. The light-touch nature of these sorts of interventions means that it is more easy to justify including subjects within the study without explicit consent, thus making it more feasible to produce a large number of cases that is plausibly more representative of the population of interest. They can also be implemented at relatively low-cost without necessarily engaging in expensive fieldwork since the whole premise is the better use of pre-existing government communications channels. There are also a great many possible permutations of messaging, using different and more intense timings, and combinations of mediums. This means that it is possible to create several arms of a study with varied 'nudges' to test against a control group that is not subject to nudges. Researchers test interventions like these in areas such as fitness, immunisation, and tax compliance (Banerjee et al., 2021; Mascagni and Nell, 2022; Solomon et al., 2014).

This all makes sense from the practical perspective of producing a series of compelling and credible studies that fit an experimental conception of rigor, but it comes at the cost of

neglecting policy relevance. The mechanisms underlying nudge interventions are often under-theorised and hard to replicate from one case to another. This means that it is hard to generalise the findings when considering other contexts or closely related policy interventions, and sometimes even to predict whether the impact of an intervention will be sustained over time. More importantly, such studies depart from the assumption that changing individual behaviour is either the primary thing that matters for policy effectiveness, or that at least it is the primary lever available to policymakers. Systemic and structural factors that determine policy outcomes are neglected. In effect, EBP may often settle disagreement not by judging and assessing the effectiveness of controversial policies but by avoiding them altogether.

*Scaling-up problems*

Inattention to the broader community impacts and structural determinants of social outcomes does not merely rule out the measurement and assessment of potentially important policy interventions. In some cases, it can lead to problems with establishing the real-world limitations of the interventions that EBP can support. For example, one relatively generalisable outcome in education, supported both by experimental and non-experimental research designs, is that smaller class sizes in schools lead to better student achievement (Angrist and Lavy, 1999). One large-scale experiment in Tennessee (Krueger, 1999) that produced positive results from reducing class-sizes in early years education caught the attention of policymakers across the United States. As a result, in California there was a sustained attempt to roll out smaller class sizes to all schools. Subsequent analyses revealed that implementing this policy at a larger scale in different location produced disappointing results compared to the original study (Munro, 2014). One important reason for this was the lack of qualified and experienced teachers to enter the new posts. Moreover, the distribution of teaching expertise was not favorable to students from disadvantaged backgrounds. In this instance, the experimental evidence added some evidence in favor of an already warranted hypothesis: holding unobserved factors stable, smaller class sizes can improve student attainment. However, it did not contribute to the more challenging policy question of how to cultivate and allocate the necessary teaching expertise necessary to make class sizes smaller at scale.

**Conclusion**

EBP promises to make the results of empirical social science intelligible to policymakers and thus allow generalists to be less reliant on trust in expert testimony when making policy decisions. It does so through putting weight on the results of randomised experimental tests of policy interventions. We have shown that EBP falls short of this promise in two critical respects.

First, the results of experimental studies do not stand alone in justifying a policy approach. A great deal of additional background knowledge needs to be present to warrant that implementing a policy shown to be effective in a set of cases will work in a new case of interest. Second, the emphasis on experimental research as the most rigorous means that research efforts avoid answering questions about systemic and structural change in favor on policies that can be conceptualised in a more individualistic frame. Not only does this miss important policy questions, but inattention to structural elements of public policy means that the results of EBP can disappoint when they interact with structural determinants. EBP does not deliver us from disagreement about the fundamentals of what works in public policy.

**References**


Angrist JD and Lavy V (1999) Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement. *The Quarterly Journal of Economics* 114(2): 533–575. DOI: 10.1162/003355399556061.

Banerjee A, Chandrasekhar A, Dalpath S, et al. (2021) *Selecting the Most Effective Nudge: Evidence from a Large-Scale Experiment on Immunization*. w28726, April. Cambridge, MA: National Bureau of Economic Research. DOI: 10.3386/w28726.

Boaz A, Ashby D, Young K, et al. (2002) *Systematic Reviews: What Have They Got to Offer Evidence Based Policy and Practice?* ESRC UK Centre for Evidence Based Policy and Practice London. Available at: http://www.kcl.ac.uk/sspp/departments/politicaleconomy/research/cep/pubs/papers/assets/wp2.pdf (accessed 17 April 2014).

Bracht GH and Glass GV (1968) The External Validity of Experiments. *American Educational Research Journal* 5(4): 437–474. DOI: 10.3102/00028312005004437.

Bundgaard H, Bundgaard JS, Raaschou-Pedersen DET, et al. (2021) Effectiveness of Adding a Mask Recommendation to Other Public Health Measures to Prevent SARS-CoV-2 Infection in Danish Mask Wearers: A Randomized Controlled Trial. *Annals of Internal Medicine* 174(3): 335–343. DOI: 10.7326/M20-6817.

Cartwright N (2007) Are RCTs the Gold Standard? *BioSocieties* 2(1): 11–20. DOI: 10.1017/S1745855207005029.

Chater N and Loewenstein GF (2022) The i-Frame and the s-Frame: How Focusing on the Individual-Level Solutions Has Led Behavioral Public Policy Astray. *SSRN Electronic Journal*. DOI: 10.2139/ssrn.4046264.

Cochrane AL (1972) *Effectiveness and Efficiency: Random Reflections on Health Services*. Nuffield Trust.

Cowen N (2019) For whom does "what works" work? The political economy of evidence-based education. *Educational Research and Evaluation* 25(1–2): 81–98. DOI: 10.1080/13803611.2019.1617991.

Cowen N and Cartwright N (2019) Street-level Theories of Change: Adapting the Medical Model of Evidence-based Practice for Policing. In: Fielding N, Bullock K, and Holdaway S (eds)

*Critical Reflections on Evidence-Based Policing*. 1st ed. Routledge, pp. 52–71. DOI: 10.4324/9780429488153.

Cowen N, Virk B, Mascarenhas-Keyes S, et al. (2017) Randomized Controlled Trials: How Can We Know "What Works"? *Critical Review* 29(3): 265–292. DOI: 10.1080/08913811.2017.1395223.

Davies HTO (ed.) (2012) *What Works? Evidence-Based Policy and Practice in Public Services*. Reprinted. Bristol: Policy Press.

Farrington DP, Gottfredson DC, Welsh BC, et al. (2002) Maryland Scientific Methods Scale. In: Sherman LW (ed.) *Evidence-Based Crime Prevention*. London ; New York: Routledge, pp. 13–21.

Frances B (2014) *Disagreement*. Key concepts in philosophy. Cambridge ; Malden, MA: Polity.

Goldacre B (2013) Building evidence into education. Available at: http://dera.ioe.ac.uk/17530/1/ben%20goldacre%20paper.pdf (accessed 17 April 2014).

Goldman A (2001) Experts: Which Ones Should You Trust? *Philosophy and Phenomenological Research* 63(1): 85–110.

Gupta S (2011) Intention-to-treat concept: A review. *Perspectives in Clinical Research* 2(3): 109. DOI: 10.4103/2229-3485.83221.

Holdcroft A (2007) Gender bias in research: how does it affect evidence based medicine? *Journal of the Royal Society of Medicine* 100(1): 2–3. DOI: 10.1258/jrsm.100.1.2.

Hope T (2005) Pretend It Doesn'T Work: The 'Anti-Social' Bias In The Maryland Scientific Methods Scale. *European Journal on Criminal Policy and Research* 11(3): 275–296. DOI: 10.1007/s10610-005-9000-1.

Howard J, Huang A, Li Z, et al. (2021) An evidence review of face masks against COVID-19. *Proceedings of the National Academy of Sciences* 118(4): e2014564118. DOI: 10.1073/pnas.2014564118.

Krueger AB (1999) EXPERIMENTAL ESTIMATES OF EDUCATION PRODUCTION FUNCTIONS. *QUARTERLY JOURNAL OF ECONOMICS* 114(2): 497–532.

Liberati A, Altman DG, Tetzlaff J, 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. DOI: 10.1371/journal.pmed.1000100.

Mascagni G and Nell C (2022) Tax Compliance in Rwanda: Evidence from a Message Field Experiment. *Economic Development and Cultural Change* 70(2): 587–623. DOI: 10.1086/713929.

Munro E (2014) Evidence-based policy. In: Cartwright N and Montuschi E (eds) *Philosophy of Social Science: A New Introduction*. First edition. Oxford, United Kingdom: Oxford University Press, pp. 48–67.

Murad MH, Asi N, Alsawas M, et al. (2016) New evidence pyramid. *Evidence Based Medicine* 21(4): 125–127. DOI: 10.1136/ebmed-2016-110401.

Reiss J (2019) Against external validity. *Synthese* 196(8): 3103–3121. DOI: 10.1007/s11229-018-1796-6.

Sanders M and Halpern D (2014) Nudge unit: our quiet revolution is putting evidence at heart of government. *The Guardian*, 3 February. Available at: https://www.theguardian.com/public-leaders-network/small-business-blog/2014/feb/03/nudge-unit-quiet-revolution-evidence (accessed 29 November 2016).

Siepmann T, Spieth PM, Kubasch AS, et al. (2016) Randomized controlled trials - a matter of design. *Neuropsychiatric Disease and Treatment*: 1341. DOI: 10.2147/NDT.S101938.

Solomon E, Rees T, Ukoumunne OC, et al. (2014) The Devon Active Villages Evaluation (DAVE) trial of a community-level physical activity intervention in rural south-west England: a stepped wedge cluster randomised controlled trial. *International Journal of Behavioral Nutrition and Physical Activity* 11(1): 94. DOI: 10.1186/s12966-014-0094-z.

What Works Team (2018) *The What Works Network - Five Years On*. January. London: What Works Network.

Young K, Ashby D, Boaz A, et al. (2002) Social Science and the Evidence-based Policy Movement. *Social Policy and Society* 1(03): 215–224. DOI: 10.1017/S1474746402003068.