Author version. Published as/cite as: **Khosrowi, D. 2021.** "Evidence-Based Policy" In *Routledge Handbook of Philosophy of Economics*, edited by Julian Reiss and Conrad Heilmann, New York, Routledge. https://doi.org/10.4324/9781315739793

Evidence-Based Policy

Donal Khosrowi

Leibniz University Hannover

ABSTRACT

Evidence-Based Policy (EBP) maintains that public policy should be based on high-quality evidence of 'what works', in particular, experimental evidence from randomized controlled trials (RCTs) that measure the effects of policy interventions. While intuitively compelling, EBP has attracted significant criticisms from philosophers, methodologists, and political scientists. This chapter pursues three aims: 1) to provide an accessible overview of EBP; 2) to consider some of the most pressing challenges that EBP faces, including issues surrounding extrapolation and external validity as well as important value-related tensions; and 3) to explore a range of ameliorative proposals for improving the ways that evidence bears on the design and implementation of public policies.

Keywords: evidence-based policy; values; methodology; external validity; internal validity; extrapolation

1 Introduction

Public policy- and institutional decision-makers routinely face questions about whether interventions 'work': does universal basic income improve people's welfare and stimulate entrepreneurial activity? Does gating alleyways reduce burglaries or merely shift the crime burden to neighbouring communities? What is the most cost-effective way to improve students' reading abilities? These are empirical questions that seem best answered by looking at the world, rather than trusting speculations about what will be effective.

Evidence-Based Policy (EBP) is a movement that concretizes this intuition. It maintains that policy should be based on evidence of 'what works'. Not any evidence will do, however. Following on the heels of its intellectual progenitor, Evidence-Based Medicine (EBM; see Evidence Based Medicine Working Group 1992; Sackett et al. 1996), EBP insists on the use of high-quality evidence produced in accordance with rigorous methodological standards. Though intuitively compelling, EBP has attracted significant criticism from methodologists, political scientists, and philosophers (see, for instance, Cartwright et al. 2009; Cartwright 2013a; Reiss 2013; Strassheim and Kettunen 2014; Muller 2015; Parkhurst 2017; Deaton and Cartwright 2018a; Favereau and Nagatsu 2020).

This chapter provides a critical overview of EBP through a philosophical lens, reviewing and discussing some of the most pressing challenges that EBP faces, and outlining some proposals for improving it. *Section 2* provides a brief overview of EBP and distinguishes a broader and narrower understanding of it. *Section 3* reviews existing criticisms, two of which are considered in detail. The first elaborates how EBP struggles with *extrapolation*, i.e. using evidence from study populations to make

inferences about the effects of policies in novel target populations. The second maintains that EBP's methodological tenets are deeply entwined with moral and political values, which can threaten EBP's promise to promote objectivity in policy-making. Finally, *Section 4* considers some proposals for how EBP could be improved going forward and concludes the discussion.

2 What's EBP?

2.1 Broad and Narrow EBP

With EBP proliferating in various policy areas, there are many ways to spell out what exactly it amounts to. Not all of these can be discussed here, but to help organize our thinking it is useful to draw a provisional distinction between a narrow and a broad understanding of EBP.

Broad EBP simply emphasises that policy should be informed by evidence, but no *general* criteria are put in place to govern what kinds of evidence to seek out and how to use them. Importantly, there is no requirement that evidence should speak directly to whether policies are *effective* - evidence could simply be used to monitor certain policy variables, for instance. To use a fringe example, gathering data on whether plankton concentration in a marine ecosystem is within a certain desirable range could count as an evidence-based way of informing marine ecology management (see e.g. Addison et al. 2018). Here, evidence guides real-world decision-making, but it is not used to determine the impacts of specific interventions, nor are there strict guidelines for what kinds of methods and evidence are good enough to inform decision-making.

By contrast, *narrow EBP*, which is the focus of this chapter, concentrates primarily on learning which policies 'work' and applies concrete strictures on what evidence is good enough for this purpose. In doing so, narrow EBP focuses on evidence from high-quality *effectiveness studies*, in particular Randomized Controlled Trials (RCTs), which are considered best for determining the effectiveness of policy interventions. Results from these studies are often 1) amalgamated in *meta-analyses* that compute an overall best estimate of a policy effect (Haynes et al. 2012), and 2) collated in *systematic reviews* and other evidence syntheses that grade available evidence according to quality and summarise it to provide a broader picture. While narrow EBP is far from a unified paradigm (see Head 2010; 2016), several general features help characterize it in more detail.

Narrow EBP is advocated, governed, and conducted by a wide range of institutions, collaborations, and research networks (see Parkhurst 2017 ch.8) such as the Campbell and Cochrane Collaboration, the GRADE and CONSORT working groups, and others[1]. These are complemented by more specific governmental institutions focusing on EBP in particular areas, such as the US Department of Education's What Works Clearinghouse or the UK's eight What Works Centres (Cabinet Office 2013), which cover a wide range of policy areas, including health, education, policing, and local economic growth. Finally, there are also numerous NGOs, academic and private institutions, and research centres, such as 3ie, J-PAL, etc., who champion the EBP approach in areas such as international development (see Duflo and Kremer 2005; Banerjee and Duflo 2009).

[1] Although Cochrane, CONSORT, and GRADE are EBM institutions, their recommendations are frequently adopted in EBP contexts.

These institutions perform a variety of functions: they offer general guidelines and concrete assistance to evaluators in conducting effectiveness studies and meta-analyses; they produce systematic reviews that survey and summarize the existing evidence-base and grade it according to quality; and they disseminate information about the (cost-) effectiveness of different interventions and the strength of the evidence underwriting these assessments. Importantly, while many of these activities involve backward-looking policy evaluation, a substantial portion of the evidence produced and summarized is supposed to help decision-makers implement policies in new environments. Ideally, decision-makers and practitioners, often termed 'users' or 'consumers', can go 'shopping' for evidence collated in so-called 'warehouses', 'libraries of evidence', or 'toolkits', which provide off-the-shelf information to help address common policy issues.

In performing these functions, a distinctive feature of narrow EBP is its reliance on rigorous methodological guidelines and so-called *evidence-hierarchies*, which rank the quality of different kinds of evidence and methods to produce them (Nutley et al. 2013). While available hierarchies differ in their details, they follow the same general blueprint. RCTs (and meta-analyses thereof) rank highest; the fact that RCTs involve experimental control is thought to make them most reliable for determining what a policy's effects are. Climbing down the ladder, one finds quasi-experimental and observational approaches, such as matching methods and simpler multivariate regression-based studies. For lack of experimental control, these study-types are thought to face more severe concerns about *risk of bias*, i.e. whether they can properly distinguish the effect of a policy from other things that influence an outcome at the same time. Finally, evidence hierarchies bottom out with the least credible kinds of studies and evidence, e.g. cohort studies, qualitative case-control studies, expert judgment, etc., which are considered even more prone to bias, or are not believed to help us determine policy effectiveness at all.

When building systematic reviews of evidence pertaining to the effectiveness of particular interventions, these hierarchies and more specific evidence-grading guidelines are used as exclusion and ranking criteria. Studies exhibiting risk of bias are discounted, or even excluded when determined to be below a certain level of quality. This *manualization* is supposed to streamline evidence production and use; ensure that rigorous standards are applied in synthesizing evidence; and make the criteria underlying evidence synthesis more transparent in the pursuit of promoting accountability and objectivity in evidence-based decision-making. With these general features in place, let us consider in more detail why RCTs are at the top of evidence-hierarchies.

2.2 Gold Standards

In producing evidence that is informative for decision-making, it is important to obtain evidence of the *causal effects* of policies. Without knowledge of what causes what, it is unlikely that our policy interventions will be successful. However, the demand for evidence of causal effects presents a formidable epistemic challenge.

Consider an example: does gating alleyways reduce burglaries (Sidebottom et al. 2017)? To answer this question, we could compare the incidence of burglaries in 'gated' and 'ungated' neighbourhoods. Yet, even if we found that gated neighbourhoods experienced fewer burglaries than ungated ones, it is not obvious

whether this difference is an effect of alley-gates or rather of some *confounding factor* (also *confounder*) that induces a spurious correlation between alley-gates and burglary rates. Perhaps those neighbourhoods that tend to have gates installed are targeted less by burglars regardless of gates, e.g. because people are more concerned about burglaries and hence more watchful, there is more CCTV, etc. We might also compare one and the same sample of neighbourhoods before and after gates are installed. Yet, even if we observed that the incidence of burglaries decreased after gates have been installed, this could be due to a variety of other reasons, such as a change in police activity.

In each of these cases, there is a worry about obtaining a *biased* measurement of the effect we are interested in, either because alley-gates don't have any effect at all and something else is responsible for any observed differences between gated and ungated neighbourhoods, or because the effect of alley-gates is muddled together with other things that happen simultaneously. Clearly, if our estimates of policy effects should be informative for decision-making, we need to ensure that we obtain *unbiased* estimates, i.e. measurements that capture only the effects of our policy and nothing else.

The standard framework underlying attempts to accomplish this is the *potential outcomes framework* (Neyman 1923; Rubin 1974; Holland 1986). When identifying a causal effect, the aim is to compare two states of the world that are alike in all respects except for the cause, intervention, or treatment of interest. More specifically, an *individual treatment effect* (ITE) is defined as the difference between a unit u's (say, a neighbourhood or individual) outcome Y in two states: a factual state $Y_t(u)$ where the unit is 'treated' (say, where a gate is installed), and a *counterfactual* state $Y_c(u)$ where the unit is 'untreated', all else equal. The *fundamental problem of causal inference* (Holland 1986), however, is that we can never observe both states for the same unit at the same time. So, how can we measure causal effects at all?

RCTs offer a deceptively simple solution by randomly allocating units to (at least) two groups: the *treatment group*, which receives a treatment (e.g. alley-gates), and the *control group*, which does not (or receives some alternative treatment, think placebos in medical trials). Randomization promises to mitigate bias because it helps ensure that the net effects of all confounding factors are balanced between the two groups. For instance, by randomly allocating neighbourhoods from a given sample to the treatment group (receiving gates) and the control group (not receiving gates) we can ensure that the net effects of policing and watchfulness on burglary rates are the same for both groups. Importantly, while any two *specific* neighbourhoods might still differ in how confounders influence their outcomes, randomization helps make sure that these differences wash away on *average*. So, when measuring the difference in the means of the outcome between treatment and control groups, we get an unbiased estimate of the *average treatment effect* (ATE) of an intervention (at least in expectation – more on this shortly).

Quasi-experimental methods that do not involve experimentally controlled assignment of treatment status (e.g. instrumental variables, regression discontinuity, differences-in-differences, and matching methods) can sometimes achieve similarly credible estimates of treatment effects. But while there has been increasing enthusiasm for these methods in empirical microeconomics (see Angrist and Pischke 2010) and beyond, EBP's methodological guidelines insist that they are not as credible as RCTs because they require a host of assumptions that are more difficult to

support by reference to features of the study design than is the case for RCTs (see Deaton and Cartwright 2018a).

For instance, matching methods (Rosenbaum and Rubin 1983) purposely select units from a population so that they differ only in their treatment-status but are similar in all other respects, particularly in regards to potential confounders. In our alley-gate example, we could match gated and ungated neighbourhoods on the level of policing, watchfulness, CCTV prevalence etc., and then compare them with respect to burglary rates. Importantly, however, the features on which units are matched must exhaust all features that can relevantly influence the outcome, which is often difficult to support. RCTs, by contrast, are argued to be applicable without knowledge of potential confounders and to generally require fewer substantive assumptions. In virtue of these features, RCTs are believed to be more credible than other study designs and are often touted as the 'gold standard' for clarifying policy effectiveness.

To summarize, while broader versions of EBP may be open to using various kinds of evidence to inform policy, narrow EBP insists on more rigorous methodological standards for what evidence is sufficiently credible. This insistence is supposed to underwrite two important promises: first, that we can build evidence libraries collating credible and ready-to-use evidence that speaks to policy issues of interest to decision-makers. Second, that this evidence, in virtue of its credibility, can push back on the role that individuals' values and convictions play in deciding which policies should be implemented, thus promoting objectivity, transparency, and accountability in policymaking (Nutley 2003:3; Abraham et al. 2017:60). Let us turn to some of the challenges that have been levelled against narrow EBP and consider how they cast doubt on whether it can deliver on its promises.

3 Challenges for EBP

Narrow EBP faces a wide range of 1) methodological as well as 2) value-related and practical challenges. Not all of these can be discussed here, so I will offer brief overviews of both kinds, each followed by a more detailed look at what I consider their most pressing instances.

3.1 Methodological Challenges

Methodological challenges take issue with the central methodological tenets of EBP, in particular evidence-hierarchies proclaiming the superiority of RCTs (see Heckman 1992; Pawson 2006; Scriven 2008; Deaton 2009; 2010; Deaton and Cartwright 2018a and other articles in the same issue). One of the key concerns is that, on closer inspection, RCTs in fact require a whole array of substantive background assumptions, which are not always satisfied and can be difficult to validate.

First, the balance in confounders between treatment and control groups that randomization is supposed to achieve only obtains *in expectation* and not necessarily on any particular measurement (see Deaton and Cartwright 2018a:4-6 for a discussion; see also Leamer 1983; 2010; Cartwright 2007; see Worrall 2002; 2007 for similar concerns about RCTs in EBM). In other words, while RCTs can provide unbiased effect estimates when results are averaged over repeated measurements, on any single occasion there may still be substantial background differences between groups that can distort an effect measure. Although this can be remedied by

inspecting the balance of confounders and re-randomizing if necessary, doing so raises similar concerns as those faced by non-randomized alternatives: one needs to know which confounders (or combinations thereof; see Fuller 2019) need to be balanced. This has led some to conclude that the assumptions required by RCTs are no less problematic in practice than those required by other methods (see Muller 2015; Bruhn and McKenzie 2009).

Second, there is a host of additional concerns about how bias can creep into RCTs even if randomization is successful (see Deaton and Cartwright 2018a for an overview). These focus on how units are selected into trials; attrition during a trial (e.g. individuals dropping out); the blinding of participants, administrators, and evaluators; spillover and equilibrium effects; and many other issues. For instance, a central assumption supposedly facilitated by randomization is that units' potential outcomes are independent of treatment status, e.g. how much an individual will benefit from an intervention should not influence whether they are assigned to the treatment or control group. This assumption is undermined when units non-randomly leave a trial in a way that correlates with their potential outcomes, e.g. low-crime neighborhoods could ultimately resist having inconvenient gates installed and drop out of a trial, which could yield an upwards-biased effect estimate for the overall population. Moreover, in many cases it will be important that units are blinded to their treatment-status: for instance, individuals aware of participating in an alley-gate trial could become unintentionally more watchful once gates are installed, thus inadvertently affecting the outcome and biasing the effect estimate. Similarly, the effects experienced by treated units can sometimes spill over to untreated ones, for instance, when gated and ungated neighbourhoods are geographically close and burglars are deterred from attempting burglaries in a whole area rather than just where gates are installed. Relatedly, some interventions, when implemented at large scale, can change not only the values of particular variables, but also more fundamental structural features of the causal setup one is seeking to meddle with (see Lucas 1976) - think would-be burglars turned violent robbers because burglary becomes too cumbersome.

A third important concern about RCTs maintains that even if they are helpful in successfully identifying causal effects, they still leave unclear *how* the effect of interest came about, e.g. by which *mechanism* or *process* an intervention worked. Such knowledge can be crucial for understanding why and how a policy is effective or not, how affected individuals experience its effects, and how a policy might be improved. For instance, we could imagine a case where alley-gates are effective in reducing burglaries not because they physically prevent access to backdoors, but because increased demand for gates happens to create job opportunities for skilled would-be burglars - something that could also be achieved in other, socially more desirable and more sustainable ways. The point here is that without ancillary analyses tracing the mechanisms and building theories of how interventions work, RCTs may not be informative enough for designing, deploying, and maintaining policy interventions; the effects they estimate remain a 'black box' that tells us too little about how policies work (see Heckman 2010 for positive proposals).

The above concerns largely target the *internal validity* of RCTs (Guala 2010), i.e. their ability to successfully measure what they seek to measure in a particular context. We now take a more detailed look at challenges that focus on *external validity*, i.e. problems encountered in reasoning beyond a particular study context.

3.2 Extrapolation: Getting From A to B Without Walking All The Way

As outlined above, a major promise of EBP is that we can build libraries of evidence, i.e. collections of expertly curated and ready-to-use evidence on a variety of policy options targeting problems routinely faced by policymakers.

This promise comes under pressure when recognizing that the populations studied in trials and the eventual target populations of interest to policymakers can often differ in important ways (Vivalt 2020), so concluding that what works in a study population A will also work in a target population B is often simply implausible (Steel 2009; 2010; Cartwright 2013b; Fuller 2019; Reiss 2019). But if, on their own, the only thing that evidence libraries can do for us is tell us what happened in a number of study populations, then why should we bother building them? To make the evidence collated in libraries useful, we need a theory of how to let it speak to questions about our eventual policy targets, i.e. a theory of how to *extrapolate* from available effectiveness evidence.[2]

There are two ideas that can help us overcome problems of extrapolation: first, not all differences between populations matter. If we can support that populations are sufficiently similar in relevant respects, we might still be entitled to draw well-supported conclusions about a novel target. Second, even if populations differ relevantly, we might nevertheless be able to account for *how* these differences bear on the effects to be expected in a target.

There is now a menu of different approaches to extrapolation that draw on these ideas. Perhaps the most general is Cartwright's *Argument Theory of Evidence* (2013a), which maintains that inferences about the effects of policies in new environments should be cast in terms of valid and sound *effectiveness arguments*. 'It works here, therefore it works there' is not a valid argument, for instance, and any methodological rigour exercised in producing evidence is undermined if one relies on bad arguments when putting it to use (Cartwright and Stegenga 2011; Cartwright and Hardie 2012).

According to Cartwright, a useful way of thinking about building better arguments is in terms of *causal principles* and *causal support factors*. Causal principles represent the causal arrangements that connect an intervention to an outcome variable. These arrangements need to be similar between populations for an intervention that works in A to also work in B - think burglars in B who prefer front door access and are unlikely to be hindered by alley-gates. Support factors are factors that interact with an intervention and need to be suitably realized for an intervention to yield its envisioned effects - think alley-gates that only work when people are willing to lock them. In making an inference to a new target one needs to learn whether similar causal principles are at work in A and B, which support factors are important for an effect, and whether they are suitably realized in B. The following sketch illustrates how we could integrate these ideas into an effectiveness argument (adapted from Cartwright 2013a:14):

P1: X plays a causal role in the production of Y in A.

[2] Extrapolation is also discussed under the rubric of external validity (Guala 2010; Marcellesi 2015; Reiss 2019), generalizability (Vivalt 2020), transferability (Cartwright 2013b), and transportability (Bareinboim and Pearl 2012).

P2: *X* can play a causal role in the production of *Y* in *B if* it does so in *A and* the support factors necessary for *X* to produce *Y* are present in *B*.

P3: The support factors necessary for *X* to produce *Y* are present in *B*.

C: Therefore, *X* can play a causal role in the production of *Y* in *B*.

Cartwright stresses that the only thing we get from an RCT is P1, but that P1 alone is not enough to infer C. P2 is needed to ensure that the causal principles are similar (or indeed the same) and clarifies the importance of support factors, and P3 ensures that these support factors are indeed present in the target. Importantly, both premises must be true, and to offer support or warrant for their truth we must invoke additional resources, including strong background theory, extensive causal knowledge of the study and target populations, etc.

Cartwright's Argument Theory provides general constraints on evidence use by emphasising that not only the quality of evidence matters (i.e. how well-supported P1 is), but that plenty of additional resources are needed to make this evidence speak compellingly to questions of interest to us. At the same time, the Argument Theory mainly provides a high-level account of how warranting conclusions about new targets should proceed. It tells us that good arguments are needed, including which general kinds of premises they might involve, but not what these arguments would look like in more concrete and more involved cases (but see Cartwright and Hardie 2012 for more detailed proposals). Importantly, the exemplary arguments used to illustrate the Argument Theory (such as the above) also do not tell us what to do in cases where populations differ relevantly. While these are not principled shortcomings, we still need additional strategies that can help us spell out more concrete and sophisticated recipes for extrapolation. Let us look at some candidates that can help make progress on this front.

Reweighting strategies (Hotz et al. 2005; Crump et al. 2008; Bareinboim and Pearl 2012; 2016; see also Athey and Imbens 2017; van Eersel et al. 2019; see Duflo 2018 for related machine learning-based methods) aim to permit extrapolation even when populations differ in relevant ways. Say the effect of *X* on *Y* depends on individual's age *Z*, i.e. *Z* is a so-called *moderating variable* that can amplify or diminish the effect (Baron and Kenny 1986). Suppose further that two populations *A* and *B* differ in their *Z*-distribution, so the *X*-*Y*-effect is likely to differ between them. Then, despite this difference, if we can capture *how* the effect depends on *Z*, we can reweight the effect measured in *A* by the observed *Z*-distribution in *B* to correctly predict the effect of interest there.

Hotz et al. (2005) provide an approach that articulates this idea through a simple but general reweighting formula. Bareinboim and Pearl's (2012; 2016) *causal graph-based approach* is more involved, allowing a wider range of more sophisticated inferences. In doing so, it requires that we can write down a *structural causal model*, i.e. a system of equations describing how variables hang together causally, and a corresponding graphical causal model (called *directed acyclic graph*, DAG) that encodes these relationships (see Scheines 1997 and Pearl 2009 ch.1 for introductions). Together with a powerful calculus, this framework helps derive formulae that permit more involved extrapolations, including in cases where populations differ in several ways at once, and where it is important to accommodate which of these differences matter and how.

Both approaches involve extensive assumptions to licence the inferences they can enable. For instance, Hotz et al.'s (2005) approach requires that we have an extensive grasp of what the important moderating variables are (Muller 2013; 2014; 2015). Moreover, it requires that study and target populations exhibit a wide range of causally relevant similarities, including at the level of the structure of causal mechanisms (Khosrowi 2019a). For instance, adjusting an effect to accommodate age differences between populations only works if the *way in which* age meddles with an effect is similar between populations. Bareinboim and Pearl's approach involves even stronger assumptions (see Hyttinen et al. 2015; Deaton and Cartwright 2018b). So, while reweighting approaches enable a wide range of more sophisticated extrapolations, they also require more support to *justify* these inferences, i.e extensive background knowledge and supplementary empirical evidence that help clarify important similarities and differences between populations.

This requirement creates two problems: first, acquiring such support can be epistemically demanding. Especially in social sciences, we rarely find sufficiently developed causal knowledge to confidently assert, for instance, that the causal mechanisms governing an outcome in two populations are similar. A second, more pernicious problem is the *extrapolator's circle* (LaFolette and Shanks 1996; Steel 2009). In a nutshell, the supplementary knowledge about the target required for an extrapolation should not be so extensive that we could identify an effect in the target based on this knowledge *alone*. For instance, if we would need to implement a policy *in a target* to learn whether the mechanisms governing its effects are similar between populations, we could simply measure the policy effect of interest there, thus, disappointingly, rendering the evidence from the study population redundant to our conclusion. The extrapolator's circle is a serious challenge for any strategy for extrapolation: in mapping out ways to make a leap from A to B we need to avoid walking all the way to B first, as otherwise there remains no leap to be made.

Steel (2009; 2010) offers detailed proposals to overcome the extrapolator's circle in biomedical sciences. His *comparative process tracing* strategy outlines how, by focusing on clarifying certain downstream similarities between populations, we can avoid learning about the target in its full causal detail. However, others remain sceptical about whether the extrapolator's circle is indeed evaded (Reiss 2010), and whether Steel's approach can be helpful for extrapolation in EBP unless a much wider range of evidence than ordinarily used there is admitted to bear on issues of causally relevant similarities and differences (Khosrowi 2019a).

In sum, there are a number of promising approaches to extrapolation. General accounts, such as Cartwright's, stress the importance of making crucial assumptions explicit and adequately supporting them. More specific strategies help detail which inferences are feasible in principle and what particular assumptions we need to bet on. However, while the inference-templates licenced by these approaches are promising, there is still a persistent lack of concrete recipes for *supporting* these inferences. And while some authors have argued that existing strategies have solved the problem of extrapolation, at least in the abstract (Marcellesi 2015), it remains doubtful whether they are sufficient to overcome concrete real-world problems of extrapolation. These strategies tell us which assumptions are needed, but they don't provide a compelling story as to how these assumptions could be underwritten in practice without falling prey to the extrapolator's circle. Let us now turn to a second set of challenges, which put additional pressure on the principled promises of EBP.

3.3 Practical and Value-Related Challenges

The second set of challenges has two strands: practical and value-related. First, various authors have voiced concerns about the practical feasibility of EBP as ideally envisioned. They worry that a simplistic template of EBP, where policy issues are identified and evidence is sought to help resolve them, rests on a naïve understanding of policy processes (Weiss 1979; Cairney 2016; Head 2016; Cairney and Oliver 2017; Parkhurst 2017). Public policymaking is often a complex and incremental struggle over political and epistemic authority (Strassheim and Kettunen 2014). In these muddied waters, where competing convictions, dogmas, and values clash, and concessions are made on some issues in exchange for authority over others, evidence is unlikely to play the role sketched out by the simplistic template. Rather, critics worry that EBP routinely degenerates into its 'evil twin', *policy-based evidence* (ibid.), where policy-makers might cherry-pick or commission the production of evidence that speaks in favour of specific agendas.

Second, even if no such attempts to unduly instrumentalize evidence take place, there remain important worries about the entanglement of moral and political values in EBP. Specifically, some authors argue that central methodological tenets of EBP can introduce (rather than mitigate) bias concerning what policy questions are considered salient and what policy options are implemented (Parkhurst 2017; Khosrowi and Reiss 2020). Let us take a closer look at this worry.

3.4 Value-Entanglement: Precision Drills and Wooden Sculptures

In promoting the use of evidence for policy, one of the key promises of EBP is that evidence can figure as a neutral arbiter to adjudicate competing value-laden convictions pertaining to what should be done (Hertin et al. 2009; Teira and Reiss 2013; Reiss and Sprenger 2020). This picture has been challenged by authors arguing that central methodological tenets of EBP are in tension with a variety of moral and political values that policy-makers might be interested in pursuing (Khosrowi 2019b; Khosrowi and Reiss 2020). Specifically, subscribing to EBP's tenets can make pursuing some kinds of questions and promoting some kinds of values substantially more difficult for policy-makers.

An analogy can help us understand this concern: crafting policies on the basis of evidence is more like crafting an intricate sculpture than building a simple bench: there is no general recipe for how to do it right; it takes creativity and vision, a great deal of dedication, and lots of skill; and having some good tools will be helpful, too. EBP's tools of choice are RCTs; they are methodologically vindicated precision drills. Yet, while they are excellent tools for getting some really difficult jobs done (drilling with precision), they are not the right tool for every job (carving), nor can they help us do any complex job from start to finish. Two important limitations of RCTs stand out: they are limited in the range of *questions* they can be applied to and in the range of *answers* they can provide, both of which can hamper their ability to cater to the complex evidentiary needs that arise in policymaking.

First, RCTs are only usefully applicable to micro-questions. Consider large-scale interventions such as tax reforms, infrastructure projects, or trade policies. Of course, observational studies also often find it difficult to estimate the effects of such interventions, but RCTs are at a distinct disadvantage: individuals or communities can

often not (realistically) be randomly subjected to the effects of expensive infrastructure projects, trade policies, or novel institutional designs that need to be implemented at scale.

Second, RCTs can only measure an *average treatment effect* (ATE), which, on its own, does not permit inferences about the individual-level effects that constitute it or the distribution of these effects (Heckman and Smith 1995). This is because any ATE can be realized in various ways, and it is impossible to distinguish between these alternatives through mere inspection of the ATE. For example, a small positive effect might be the result of all treated individuals benefitting by roughly the same amount, or by a small group experiencing significant benefits while many others are made significantly worse off.

These two limitations give rise to two subsequent problems. First, insisting on the superiority of RCTs can constrain the range of policy questions that can be clarified by evidence, e.g. by privileging the pursuit of micro-questions. This is problematic because it can distort what kinds of policy issues are targeted as relevant and what sorts of interventions are considered (see Barnes and Parkhurst 2014; Parkhurst and Abeysinghe 2016; Parkhurst 2017 on *issue bias*). Second, since RCTs and meta-analyses only supply ATEs, policymakers who are interested in distributive issues (say, prioritizing the worst-off individuals in a population) are put at a disadvantage because the information they need is not supplied (e.g. whether a policy benefits important subgroups) (Khosrowi 2019b). To make progress in clarifying whether a policy has desirable distributive effects, the evidence that conforms to existing quality standards is not informative enough. At the very least, it would need to be complemented by ancillary subgroup-analyses that clarify the distribution of effects (Varadhan and Seeger 2013). Yet, even if such analyses were routinely available, which they are not, existing guidelines would not award them the same credibility as the primary RCT results that they are supposed to complement (Khosrowi 2019b).

This situation is undesirable for policymakers interested in distributive issues. First, it can lead them to pursue only those value-schemes for which highly-ranked evidence exists, e.g. focusing on average outcomes instead of politically and morally salient subgroups. Second, resisting this pull might force policymakers to call on putatively inferior evidence, which makes it easier for opponents to challenge them. Either way, it seems that policies pursuing distributive aims could be crowded out of policy debates.

Together, these concerns suggest that EBP has a problem with values. Ideally, the evidence produced would be equally useful for the pursuit of a broad range of values and purposes (Khosrowi and Reiss 2020). However, since existing methodological tenets in EBP seem to render this unlikely, it remains doubtful whether evidence can really play the role of a neutral arbiter that promotes objectivity in policy-making processes. With these challenges in place, let us briefly consider some ameliorative proposals and draw some broader conclusions.

4 Conclusions & Outlook: Towards Better EBP

Despite pressing challenges, many critics of EBP agree that there is something sensible about the idea that evidence can, at least under some circumstances, and in some ways, make valuable contributions towards improving the design and implementation of good public policy (see e.g. Cartwright 2012: 975). Yet, while a

minimalist commitment to using evidence, along the lines of broad EBP, seems sensible, the way in which narrow EBP has detailed how evidence is supposed to inform policy is doubly problematic: First, its strictures on evidence-*production* seem too heavy-handed, inviting important value-related tensions. Second, evidence-*use* (e.g. extrapolation) remains largely unregimented and little is said on how we can make evidence persuasively speak to questions about novel targets. There is hence a clear need for further refinements. Let us briefly consider some recent proposals addressing these problems.

Concerning value-related tensions, it seems clear that the limitations of gold-standard methods should not dictate which policy questions are considered relevant and which policy options appear salient. However, advocating for methods other than RCTs on the grounds of meeting important evidential needs must still remain accountable to genuine concerns about their credibility and departures from good scientific practice (Parkhurst 2017). Resolving these tensions is far from trivial. Recent proposals taking issue with value-entanglements emphasise that governing the production of evidence for policy must combine concerns about the quality and credibility of evidence with considerations of its *appropriateness* or *usefulness*; something which existing guidelines often remain silent on (see Parkhurst 2017; Khosrowi and Reiss 2020).

For instance, Parkhurst (2017 chs. 7&8) makes detailed proposals for building a general framework that helps govern how evidence figures in policy-making. Retaining commitments to important ideals concerning quality of evidence, he places particular emphasis on facilitating the *democratic legitimacy* of evidence advisory systems and on ensuring, through institutional design and procedural precautions, that stakeholder values play a central role in governing how evidence informs policy.

Khosrowi and Reiss (2020) call for open methodological debate among methodologists, stakeholders, and producers and users of evidence about how to weigh different epistemic, value-related, and pragmatic criteria in refining the use of evidence for policy (see e.g. Head 2010 and other articles in the same issue for an exemplary instance). To use the metaphor once more: if our aim is to craft appealing sculptures, we need to consider the full array of tools available (files, sanders, saws, hammers, chisels, drills - some fine, some coarse) and investigate how *combining* these tools can help us address the full range of sculptory needs. Without such attempts, we'll be stuck with the oddly-shaped sculptures that precision drills can give us.

In contrast to value-related challenges, concerns about extrapolation are now more widely recognized (Bates and Glennerster 2017; Cowen et al. 2017; Duflo 2018; see Favereau and Nagatsu 2020 for a discussion). Yet, while abstract, general strategies for extrapolation are available, the arguments discussed above suggest that more work is needed to devise concrete, practical recipes that work.

Promising proposals adressing this need have been made in the realist evaluation literature (see e.g. Pawson and Tilley 1997; 2001; White 2009; Astbury and Leeuw 2010; Pawson 2013; Davey et al. 2018) and were recently reinforced by philosophers (Cartwright 2020). Rather than focusing on black-box estimates of policy effects, the aim in realist evaluation is to develop explicit *programme theories* (also called 'logic frames' or 'theories of change'), which elucidate 1) *how*, i.e. by which mechanisms, interventions are effective, 2) what circumstances promote and hinder their success, 3) how interventions may work differently for different individuals, and 4) how their effects are experienced.

In addition to emphasising the important role of theory in facilitating extrapolation, there have also been calls to (re-)consider what kinds of *supplementary* evidence are needed for underwriting extrapolation and to provide recipes for how to produce and use such evidence (Khosrowi 2019a). RCTs, by themselves, cannot clarify whether there are important similarities and differences between populations. Additional evidence is hence required to support extrapolation, and this may include not only familiar kinds of evidence, such as quantitative observational data, but also evidence that is rarely considered in EBP, such as mechanistic evidence obtained from process tracing studies (Beach and Pedersen 2019), or evidence from qualitative studies, such as ethnographies.

Yet, while some of these proposals have recently been taken up by EBP institutions such as 3ie (Peters et al. 2019), they have yet to gain traction in other corners. The Campbell Collaboration's guidelines, for instance, mostly refer to the guidelines issued by its EBM relatives, the Cochrane Collaboration and Grade Working Group. While these guidelines alert authors and users of systematic reviews to some of the pitfalls involved in extrapolation (Guyatt et al. 2011) and recommend that evidence should be downgraded when differences between study and target populations seem likely (Schünemann et al. 2019), they do not suggest strategies for how to overcome the problems encountered and leave it to evidence-users to judge whether findings are applicable to their contexts. So, despite important theoretical progress, systematic proposals for how to manage extrapolation have yet to be accommodated in EBP's methodological guidelines.

In sum, the challenges outlined in this chapter and the ameliorative proposals addressing them suggest that continued attention by methodologists, practitioners, EBP researchers, and others is needed to build a more compelling model for how evidence can bear on policy. In advancing this project, it seems important to recognize that a more compelling version of EBP is unlikely to work in the same way across different policy-domains (cf. Head 2010). Instead, it seems that many different, domain-specific approaches are needed that each take into account 1) what types of policy issues arise in specific domains, 2) what values are pertinent to identifying and addressing them, 3) what kinds of questions require attention, 4) how (combinations of) existing and heretofore neglected methods can best cater to these questions, and 5) how evidence production and use can be best integrated into existing institutional and decision-making structures.

Some EBP areas, such as education, child safety, and policing have made good progress on this front, with domain-specific methodologies being developed that recognize the limits of narrow EBP (Cowen et al. 2017; Munro et al. 2017). There are also cases involving more principled obstacles to adapting a narrow EBP template to the concrete needs arising in a particular domain. For instance, in evidence-based environmental policy it is widely recognized that RCTs cannot be feasibly implemented to assess the effects of environmental policies (Hayes at al. 2019) and that environmental management often requires information at higher spatial and temporal resolution than typical effectiveness studies can provide (Ahmadia et al. 2015). Problems like these suggest that simply adapting existing EBP templates to particular domains is not always possible and that new, local models for how evidence can inform decision-making might often be needed.

In refining EBP, adapting it to specific domains, and devising new models for how evidence informs policy, important opportunities arise for philosophers to engage in

methodological debates with EBP advocates, practitioners, and methodologists. Key to making these debates productive will be to ensure an up-to-date grasp of developments in the various areas in which EBP is gaining traction, as well as a commitment to making positive proposals that speak to ongoing practices.

The present chapter has provided a foundational overview of the philosophical and methodological debates surrounding EBP, but, of course, makes no claim to be comprehensive. Nevertheless, it is hoped that by focusing on a selection of recent and largely unresolved issues, the overview provided here will help stimulate further critical and constructive contributions by philosophers towards improving EBP.

References

Abraham, K., R, Haskings, S. Glied, R. Groves, R. Hahn, H. Hoynes, J. Liebman, B. Meyer, P. Ohm, N. Potok, K. R. Mosier, R. Shea, L. Sweeney, K. Troske, and K. Wallin. 2017. *The Promise of Evidence-Based Policymaking: Report of the Commission on Evidence-Based Policymaking*. Washington (DC), Commission on Evidence-Based Policymaking.

Addison, P. F. E, D. G. Collins, R. Trebilco, S. Howe, N. Bax, P. Hedge, G. Jones, P. Miloslavich, C. Roelfsema, M. Sams, R. D. Stuart-Smith, P. Sanes, P. von Baumgarten, and A. McQuatters-Gollop. 2018. "A new wave of marine evidence-based management: emerging challenges and solutions to transform monitoring, evaluating, and reporting." *ICES Journal of Marine Science* 75(3): 941-52.

Ahmadia, G. N., L. Glew, M. Provost, D. Gill, N. I. Hidayat, S. Mangubhai, Purwanto, and H. E. Fox. 2015. "Integrating impact evaluation in the design and implementation of monitoring marine protected areas". *Philosophical Transactions of the Royal Society B* 370: 20140275.

Angrist, J., and J. Pischke. 2010. "The Credibility Revolution in Empirical Economics: How Better Research Design Is Taking the Con out of Econometrics." *Journal of Economic Perspectives* 24(2): 3–30.

Astbury, B., and F. Leeuw. 2010. "Unpacking Black Boxes: Mechanisms and Theory Building in Evaluation." *American Journal of Evaluation* 31(3): 363-81.

Athey, S., and G. W. Imbens. 2017. "The state of applied econometrics: causality and policy evaluation." *Journal of Economic Perspectives* 31(2): 3–32.

Banerjee, A., and E. Duflo. 2009. "The Experimental Approach to Development Economics." *Annual Review of Economics* 1: 151–78.

Bareinboim, E., and J. Pearl. 2012. "Transportability of causal effects: Completeness results." In *Proceedings of the Twenty-Sixth Conference on Artificial Intelligence (AAAI-12)*, Menlo Park, CA.

— **2016.** "Causal inference and the data-fusion problem." *Proceedings of the National Academy of Sciences* 113: 7345-52.

Barnes, A. and J. Parkhurst. 2014. "Can global health policy be depoliticised? A critique of global calls for evidence-based policy." In G. Yamey and G. Brown (eds.) *Handbook of Global Health Policy*, 157–73. Chichester: Wiley-Blackwell.

Baron, R. M., and Kenny, D. A. 1986. "The moderator-mediator variable distinction in social psychological research: Conceptual, strategic and statistical considerations." *Journal of Personality and Social Psychology* 51: 1173-82.

Bates, M. A., and R. Glennerster. 2017. "The generalizability puzzle." *Stanford Social Innovation Review*. Retrieved June 2020 from: https://ssir.org/articles/entry/the_generalizability_puzzle.

Beach, D., and R. B. Pedersen. 2019. *Process-Tracing Methods - Foundations and Guidelines*. 2nd edition. Ann Arbor: University of Michigan Press.

Bruhn, M., and D. McKenzie. 2009. "In Pursuit of Balance: Randomization in Practice in Development Field Experiments." *American Economic Journal: Applied Economics*, 1(4): 200-32.

Cabinet Office 2013. *What Works: evidence centres for social policy*. London: Cabinet Office.

Cairney P. 2016. *The politics of evidence-based policymaking*. London: Palgrave Pivot.

Cairney, P., and K. Oliver. 2017. "Evidence-based policymaking is not like evidence-based medicine, so how far should you go to bridge the divide between evidence and policy?" *Health research policy and systems* 15(1): 35.

Cartwright, N. D. 2007. "Are RCTs the Gold Standard?" *BioSocieties* 2(2): 11-20.

— **2012.** "Presidential Address: Will This Policy Work for You? Predicting Effectiveness Better: How Philosophy Helps." *Philosophy of Science* 79(5): 973-89.

— **2013a.** "Evidence, Argument and Prediction." In V. Karakostas, and D. Dieks (eds.), EPSA11 Perspectives and Foundational Problems in Philosophy of Science, The European Philosophy of Science Association Proceedings. Cham: Springer International Publishing Switzerland.

— **2013b.** "Knowing What We Are Talking About: Why Evidence Doesn't Always Travel." *Evidence and Policy: a Journal of Research, Debate and Practice* 9(1): 97-112.

— **2020.** "Lullius Lectures 2018: Mid-level theory: Without it what could anyone do?" In C. Martínez Vidal and C. Saborido (eds.) *Nancy Cartwright's Philosophy of Science*, Special Issue of *Theoria*.

Cartwright, N. D., A. Goldfinch, and J. Howick. 2009. "Evidence-based policy: where is our theory of evidence?" *Journal of Children's Services* 4(4):6-14.

Cartwright, N. D., and J. Hardie. 2012. *Evidence-Based Policy: A Practical Guide to Doing it Better*. Oxford: Oxford University Press.

Cartwright, N. D., and J. Stegenga. 2011. "A Theory of Evidence for Evidence-Based Policy" *Proceedings of the British Academy*, 171: 289–319.

Cowen, N., B. Virk, S. Mascarenhas-Keyes, and N. D. Cartwright. 2017. "Randomized Controlled Trials: How Can We Know 'What Works'?" *Critical Review* 29(3): 265-92.

Crump, R. K., Hotz, V. J., Imbens, G. W. and Mitnik, O. A. (2008). "Nonparametric tests for treatment effect heterogeneity", *The Review of Economics and Statistics*, 90 (3): 389–405.

Davey, C., S. Hassan, N. D. Cartwright, M. Humphreys, E. Masset, A. Prost, D. Gough, S. Oliver, C. Nonell, and J. Hargreaves. 2019. "Designing evaluations to provide evidence to inform action in new settings". CEDIL Inception Paper No 2: London.

Deaton, A. 2009. "Instruments of Development: Randomisation in the Tropics, and the Search for the Elusive Keys to Economic Development." *Proceedings of the British Academy* 162: 123–60.

— **2010.** "Instruments, randomization, and learning about development." *Journal of Economic Literature*, 48(2): 424-55.

Deaton, A., and N. D. Cartwright 2018a. "Understanding and misunderstanding randomized controlled trials." *Social Science & Medicine* 210: 2-21.

— **2018b.** "Reflections on Randomized Control Trials". *Social Science & Medicine,* 210: 86-90.

Duflo, E. 2018. "*Machinistas meet randomistas: Useful ML tools for empirical researchers.*" Summer Institute Master Lectures. National Bureau of Economic Research.

Duflo, E., and M. Kremer. 2005. "Use of Randomization in the Evaluation of Development Effectiveness." In G. Pitman, O. Feinstein, and G. Ingram (eds.) *Evaluating Development Effectiveness*. New Brunswick, NJ: Transaction.

Evidence-Based Medicine Working Group. 1992. "Evidence-based medicine. A new approach to teaching the practice of medicine." *JAMA* 268: 2420–25.

Favereau, J., and M. Nagatsu. 2020. "Holding back from theory: limits and methodological alternatives of randomized field experiments in development economics." *Journal of Economic Methodology.* 40(1): 1-21.

Fuller, J. 2019. "The Confounding Question of Confounding Causes in Randomized Trials." *The British Journal for the Philosophy of Science* 70(3): 901–26.

Guala, F. 2010. "Extrapolation, analogy, and comparative process tracing." *Philosophy of Science* 77(5):1070–82.

Guyatt, G. H., A. D. Oxman, R. Kunz, J. Woodcock, J. Brozek, M. Helfand, P. Alonso-Ciello, Y. Falck-Yter, R. Jaeschke, G. Vist, E. A. Akl, P. N. Post, S. Norris, J. Meerpohl, V. K. Shukla, M. Nasser, and H. J. Schünemann. 2011. "Grade Guidelines: 8. Rating the Quality of Evidence - Indirectness." *Journal of Clinical Epidemiology* 64(12):1301-10.

Hayes, K. R., G. R. Hosack, E. Lawrence, P. Hedge, N. S. Barrett, R. Przeslawski, J. M. Caley, and S. D. Foster. 2019. "Designing Monitoring Programs for Marine Protected Areas Within an Evidence Based Decision Making Paradigm." *Frontiers in Marine Science* 6: 746.

Haynes, L., Service, O., Goldacre, B., and Torgerson, D. 2012. *Test, Learn, Adapt: Developing Public Policy with Randomised Controlled Trials*. London: Cabinet Office Behavioural Insights Team.

Head, B. W. 2010. "Reconsidering evidence-based policy: Key issues and challenges." *Policy and Society* 29(2): 77-94.

— **2016.** "Toward More 'Evidence-Informed' Policy Making?" *Public Administration Review* 76: 472-84.

Heckman, J. J. 1992. "Randomization and Social Program Evaluation". In C. Manski and I. Garfinkel (eds.) *Evaluating Welfare and Training Programs* 201-30. Cambridge, MA: Harvard University Press.

— **2010.** "Building Bridges between Structural and Program Evaluation Approaches to Evaluating Policy." *Journal of Economic Literature* 48(2): 356-98.

Heckman, J. J., and J. A. Smith. 1995. "Assessing the Case for Social Experiments." *Journal of Economic Perspectives* 9(2): 85-110.

Hertin, J., K. Jacob, U. Pesch, and C. Pacchi. 2009. "The Production and Use of Knowledge in Regulatory Impact Assessment – An Empirical Analysis." *Forest Policy & Economics* 11(5-6):413–21.

Holland, P. 1986. "Statistics and Causal Inference." *Journal of the American Statistical Association* 81(396):945-60.

Hotz, V. J., G. W. Imbens, and J. H. Mortimer. 2005. "Predicting the efficacy of future training programs using past experiences at other locations." *Journal of Econometrics* 125: 241–70.

Hyttinen, A., F. Eberhardt, and M. Järvisalo. 2015. "Do-calculus when the true graph is unknown." UAI'15 Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence, 395-404.

Khosrowi, D. 2019a. "Extrapolation of Causal Effects – Hopes, Assumptions, and the Extrapolator's Circle." *Journal of Economic Methodology* 26(1):45-58.

— **2019b.** "Trade-Offs Between Epistemic and Moral Values in Evidence-Based Policy." *Economics and Philosophy* 35(1):49-71.

Khosrowi, D., and J. Reiss. 2020. "Evidence-Based Policy: The Tension Between the Epistemic and the Normative." *Critical Review* 31(2):179-97.

LaFollette, H., and N. Shanks. (1996). *Brute Science: Dilemmas of Animal Experimentation*. New York: Routledge.

Leamer, E. 1983. "Let's Take the Con Out of Econometrics." *The American Economic Review* 73(1):31-43.

— **2010.** "Tantalus on the Road to Asymptopia." *Journal of Economic Perspectives* 24(2): 31-46.

Lucas, R. 1976. "Econometric Policy Evaluation: A Critique." In K. Brunner and A. Meltzer (eds.) *The Phillips Curve and Labor Markets*, Amsterdam: North-Holland.

Marcellesi, A. 2015. "External Validity: Is There Still a Problem?" *Philosophy of Science* 82(5): 1308-17.

Muller, S. M. 2013. „External validity, causal interaction and randomised trials: the case of economics", Unpublished manuscript.

— **2014**. "Randomised trials for policy: a review of the external validity of treatment effects." Southern Africa Labour and Development Research Unit Working Paper 127, University of Cape Town.

— **2015**. "Interaction and external validity: obstacles to the policy relevance of randomized evaluations." *World Bank Economic Review* 29(1): 217-25.

Munro, E., N.D. Cartwright, J. Hardie, and E. Montuschi. 2017. *Improving Child Safety: deliberation, judgement and empirical research*. Durham: Centre for Humanities Engaging Science and Society (CHESS).

Neyman, J. 1923. "On the application of probability theory to agricultural experiments: essay on principles." (Section 9). Translated in *Statistical Science* 5:465–80 (1990).

Nutley, S. 2003. "Briding the polcy/research divide: reflections and lessons from the UK". Keynote paper: Facing the future: engaging stakeholders and citizens in developing public policy. National Institute of Governance Conference, Canberrra, Australia.

Nutley, S., A. Powell, and H. Davies. 2013. "What Counts as Good Evidence?" Discussion paper. London: Alliance for Useful Evidence.

Parkhurst, J. 2017. *The politics of evidence: from evidence-based policy to the good governance of evidence.* Abingdon, Oxon, UK: Routledge.

Parkhurst, J. and S. Abeysinghe. 2016. "What constitutes "good" evidence for public health and social policy-making? From hierarchies to appropriateness." *Social Epistemology* 5: 665-79.

Pawson, R. 2006. *Evidence-Based Policy: A Realist Perspective*. London and Thousand Oaks (CA): SAGE.

— **2013.** *The science of evaluation: a realist manifesto*. London: SAGE Publications.

Pawson, R., and N. Tilley. 1997. *Realistic Evaluation*. London: SAGE Publications.

— **2001.** "Realistic Evaluation Bloodlines." *American Journal of Evaluation* 22:317-24.

Peters, J., M. Jain, and M. Gaarder. 2019. "External validity: policy demand is there but research needs to boost supply". 3ie blog post. Retrieved June 2020 from: https://www.3ieimpact.org/blogs/external-validity-policy-demand-there-research-needs-boost-supply

Pearl, J. 2009. *Causality: Models, Reasoning, and Inference*. 2nd edition. New York: Cambridge University Press.

Reiss, J. 2010. "Review: Across the boundaries: Extrapolation in biology and social science." *Economics and Philosophy* 26:382-390

— **2013.** *The Philosophy of Economics: A Contemporary Introduction*. New York: Routledge.

— **2019.** "Against external validity." *Synthese* 196(8): 3103-21.

Reiss, J. and J. Sprenger. 2020. "Scientific Objectivity." The Stanford Encyclopedia of Philosophy (Winter 2020 Edition), Edward N. Zalta (ed.), https://plato.stanford.edu/archives/win2020/entries/scientific-objectivity.

Rosenbaum, P. R., and D. B. Rubin. 1983. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika* 70(1): 41–55.

Rubin, D. B. 1974. "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies." *Journal of Educational Psychology* 66(5): 688–701.

Sackett D., W. Rosenberg, M. Gray, B. Haynes, and S. Richardson. 1996. "Evidence based medicine: what it is and what it isn't." *BMJ* 312: 71.

Scheines, R. 1997. „An introduction to causal inference" In McKim and Turner (eds.) *Causality in Crisis? Statistical Methods in the Search for Causal Knowledge in the Social Sciences*, 185-99. Notre Dame, IN: University of Notre Dame Press.

Schünemann, H. J., J. P. T. Higgins, G. E. Vist, P. Glasziou, E. A. Akl., N. Skoetz, and G. H. Guyatt. 2019. "Chapter 14: Completing 'summary of findings' tables and grading the certainty of the evidence." In: J. P. T. Higgins, J. Thomas, J. Chandler, M. Cumpston, T. Li, M. J. Page, et al. (eds.) *Cochrande handbook for systematic reviews of interventions version 6.0*. Retrieved June 2020 from https://training.cochrane.org/handbook.

Scriven, M. 2008. "A Summative Evaluation of RCT Methodology & An Alternative Approach to Causal Research." *Journal of Multi-Disciplinary Evaluation* 5(9): 11-24.

Sidebottom, A., L. Tompson, A. Thornton, K. Bullock, N. Tilley, K. Bowers, and S. D. Johnson. 2018. "Gating Alleys to Reduce Crime: A Meta-Analysis and Realist Synthesis." *Justice Quarterly* 35(1): 55-86.

Steel, D. 2009. *Across the boundaries: Extrapolation in biology and social science*. Oxford: Oxford University Press.

— **2010.** "A New Approach to Argument by Analogy: Extrapolation and Chain Graphs", *Philosophy of Science* 77(5): 1058-69.

Strassheim, H., and P. Kettunen. 2014. " When does evidence-based policy turn into policy-based evidence? Configurations, contexts and mechanisms." *Evidence & Policy* 10(2): 259-77.

Teira D., and J. Reiss. 2013. "Causality, Impartiality and Evidence-Based Policy". In: Chao H. K., Chen S. T., Millstein R. (eds) *Mechanism and Causality in Biology and Economics*. History, Philosophy and Theory of the Life Sciences, vol 3. Dordrecht: Springer.

van Eersel, G. G., G. V. Koppenol-Gonzalez, and J. Reiss. 2019. "Extrapolation of Experimental Results through Analogical Reasoning from Latent Classes." *Philosophy of Science* 86(2): 219-35.

Varadhan, R and J. D. Seeger. 2013. "Estimation and reporting of heterogeneity of treatment effects." In P. Velentgas, N. A. Dreyer, P. Nourjah, S. R. Smith and M. M. Torchia (eds) Developing a Protocol for Observational Comparative Effectiveness Research: A User's Guide, pp. 35–44. Rockville, MD: Agency for Healthcare Research and Quality.

Vivalt, E. 2020. "How Much Can We Generalize from Impact Evaluations?" *Journal of the European Economics Association* (online first).

Weiss, C. H. 1979. "The Many Meanings of Research Utilization." *Public Administration Review* 39(5): 426-31.

White, H. 2009. "Theory–Based Impact Evaluation: Principles and Practice." *The Journal of Development Effectiveness* 1(3): 271-84

Worrall, J. 2002. "What Evidence in Evidence-Based Medicine?" *Philosophy of Science* 69: 316–30.

— **2007**. "Why There's no Cause to Randomize." *The British Journal for the Philosophy of Science* 58(3): 451-88.