Critical Review

A Journal of Politics and Society

ISSN: 0891-3811 (Print) 1933-8007 (Online) Journal homepage: https://www.tandfonline.com/loi/rcri20

Evidence-Based Policy: The Tension Between the Epistemic and the Normative

Donal Khosrowi & Julian Reiss

To cite this article: Donal Khosrowi & Julian Reiss (2019): Evidence-Based Policy: The Tension Between the Epistemic and the Normative, Critical Review, DOI: 10.1080/08913811.2019.1688520

To link to this article: https://doi.org/10.1080/08913811.2019.1688520

Published online: 25 Nov 2019.

Donal Khosrowi and Julian Reiss

EVIDENCE-BASED POLICY: THE TENSION BETWEEN THE EPISTEMIC AND THE NORMATIVE

ABSTRACT: *Acceding to the demand that public policy should be based on "the best available evidence" can come at significant moral cost. Important policy questions cannot be addressed using "the best available evidence" as defined by the evidence-based policy paradigm; the paradigm can change the meaning of questions so that they can be addressed using the preferred kind of evidence; and important evidence that does not meet the standard defined by the paradigm can get ignored. We illustrate these problems in three contexts of evidence-based policy.*

Keywords: *evidence-based policy; evidence-based initiatives; randomized controlled trials; behavioral economics; development economics; nudge theory; Abhijit Banerjee; Esther Duflo*

The last twenty years have seen increasing academic and public debate about how scientific evidence should inform policy. This has led to countless "evidence-based initiatives" and "What Works Centers" in

Donal Khosrowi, donal.khosrowi@philos.uni-hannover.de, a postdoctoral researcher at the Institute for Philosophy, Hannover University, Im Moore 21, 30167, Hannover, Germany, would like to thank Magdalena Małecka and the audience at PSA2018 in Seattle for helpful discussions of some of the ideas developed here. This research was financially supported by an AHRC Northern Bridge Doctoral Studentship (grant AH/L503927/1), a Durham Doctoral Studentship, and a Royal Institute of Philosophy Jacobsen Studentship. Julian Reiss, julian.reiss@jku.at, Institute for Philosophy and Scientific Method at Johannes Kepler University Linz, Altenberger Str. 50, A-4040 Linz, Austria, is the author of *Causation, Evidence, and Inference* (Routledge, 2015), *Philosophy of Economics* (Routledge, 2013), and *Error in Economics* (Routledge, 2008). He acknowledges funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program, Grant Agreement No. 667526 K4U.

Critical Review 1–19 ISSN 0891-3811 print, 1933-8007 online
https://doi.org/10.1080/08913811.2019.1688520

the United States, the United Kingdom, and many other countries (Abraham et al. 2017; Banks 2009; Parsons 2002).

Proponents of "evidence-based" policy often defend their approach by means of commonplaces such as the following: "It is a fundamental principle of good public services that decisions are made on the basis of strong evidence and what we know works" (Cabinet Office 2013, i). However, once it is spelled out what is meant by "strong evidence" and "knowledge of what works," evidence-based policy becomes highly controversial.

Advocates of evidence-based policy propose that randomized controlled trials (RCTs) should serve as the "gold standard" of evidence (Haynes et al. 2012). Epistemic criticisms of this idea are well known and will not be repeated here (see Byrne et al. 2009, Deaton and Cartwright 2017, Heckman 1992, Pawson 2006, and Scriven 2008). We will henceforth ignore questions of reliability and grant to evidence-based policy proponents, for the sake of argument, that (if and when the conditions for application are satisfied) RCTs are the best method to address "what works" questions, i.e., questions about the effectiveness of policies in changing the mean value of specific target variables in a desired direction.

The concern we raise in this paper is a different one. Any scientific method is limited by the types of questions it can address; the types and availability of data it needs if it is itself going to "work well"; and the types and amount of background knowledge it requires. For example, the regression-based estimation of policy effects is capable of testing causal hypotheses, but only if sample sizes are large enough, there is enough variation in the data, there are no unmeasured confounders, and the joint probability distribution of the variables of interest is known. Importantly, while (under these assumptions) regression can help us estimate the causal effect of an independent policy variable X on a dependent outcome variable Y, it tests only a hypothesis concerning a "black-box" causal relationship; it sheds no light on why or how such a causal relationship holds. Another method, process tracing, can be used to learn the causal mechanism(s) through which X is causally relevant for Y, but it requires substantive background knowledge (for instance about possible alternative explanations of event sequences) to distinguish a genuine causal mechanism from a mere sequence of events, and it cannot address quantitative issues. Similar points can be made about virtually any research method. Regarding "evidence-based" policy, i.e., policy based on randomized controlled trials, the problem is that the methods for producing evidence constrain the types of questions the evidence can help us answer: i.e., only those questions for which the preferred ("gold-standard") methods can provide informative evidence.

In what follows, we dicuss this problem in three contexts: the use of randomized controlled trials in public policy, development economics, and behavioral nudge policy, with the intention of showing that the problem, if ignored, can lead to inferior policy outcomes. We will focus on illustrations that emphasize consequences of methodological choices related to the *justice* of outcomes, especially distributive justice.

Our discussion exemplifies a more general, well-known feature of social-scientific practice: that the methodological and moral aspects of research are entangled and cannot neatly be separated. It is not possible, for instance, to develop methodological standards for good research practice without deciding, if only implicitly, which research questions are important or significant (cf. Kitcher 2001, Kitcher 2011). Likewise, we cannot neatly distinguish between a "technocratic" stage of research, in which it is determined which policies work best, and a "political" stage, in which the most desirable policies are chosen among the feasible options.

Problems with the Evidence in Evidence-Based Policy

Our first illustration concerns a peculiarity of RCTs as the "gold standard" of evidence. RCTs identify a quantity called the "average treatment effect," or ATE. One way to motivate the use of RCTs is by invoking the "potential outcomes approach" to causation (Holland 1986). The core concept here is the "*individual* treatment effect" (ITE):

$$Y_t(u) - Y_c(u)$$

where $Y_i(u)$ is the outcome that would have obtained for unit/individual u had it received treatment i, and t and c refer to treatment and control conditions, respectively. For instance, Y might be the duration of an individual's unemployment, t a new policy that requires her to show up biweekly at the local unemployment office and answer specific questions, and c the standard policy, requiring her to show up only every three months and without answering any questions.

The problem with ITEs is that they are unobservable. An individual can be subject to only one of the alternative policies, not both (at the same time). This motivates the search for a next-best identifiable quantity, which is what ATEs provide. They measure the difference of the expected value of an outcome between two *groups* of individuals subject to the two treatments t and c:

$$E_u[Y_t(u)] - E_u[Y_c(u)]$$

RCTs help us identify this quantity when their main identification assumptions are satisfied: most importantly, that outcomes are statistically independent of treatment assignment.

An immediate problem with ATEs is that unless we can assume that all individuals are perfectly alike in all characteristics relevant to the causal effect of interest, the ATE will generally not be equal to the ITEs experienced by individuals. ITEs may in fact differ significantly from the ATE—for example, a positive ATE is consistent with some individuals experiencing large positive effects, while others experience negative effects. This is called *treatment-effect heterogeneity*: a systematic variation in the sign or magnitude of a causal effect among individuals. Such heterogeneity is common in many evidence-based policy areas, such as development economics, where some individuals may greatly benefit from access to, say, microcredit, while others will not (Banerjee et al. 2017); in education policy, where students may respond differentially to interventions as a function of initial ability; and in economic policy, where behavioral responses to interventions may differ significantly between industries, firms, and other units (Khosrowi 2019).

Heterogeneity, by itself, would be neither surprising nor much of a problem if policy makers only cared about improving *average* outcomes. But that is not generally the case. Many policy makers pursue distributive goals. They may want to maximize aggregate welfare, which is sensitive to who gets what. They may aim for equality, for which they need to know whether policy effects are equally distributed or help equalize some outcome. They may prioritize increases in welfare for the worst off. To pursue such goals, and, more generally, any goal that is concerned with the distribution of policy effects, policy makers will need more information than the ATE; they will need evidence concerning the *distribution* of treatment effects.

Such information cannot, at present, be recovered from ATEs, and there is no obvious remedy. While it is possible to conduct subgroup analyses to help produce information on heterogeneity (Varadhan and Seeger

2013), these analyses are generally not considered credible on epistemic grounds, so performing them is discouraged in widely circulated methodological guidelines (Khosrowi 2019). The main problem with subgroup analyses is that they are prone to produce spurious results and do not permit clear-cut causal interpretation, for two reasons. First, since partitioning a population into subgroups decreases sample sizes within groups, their statistical power becomes questionable. Smaller samples make it more likely that we detect significant results simply due to chance, and that truly significant results remain masked by the coarse-grained noise encountered in small samples. Second, even if we have more adequate sample sizes and find a significant difference in effects between, say, older and younger subgroups, this does not imply that age is responsible for bringing about this observed difference; other things besides age, but correlated with it, might induce the differences we see.

These are not idle concerns. However, it is important to recognize that they are worst-case scenarios. Any study, if poorly designed, can raise concerns about the credibility of its results. This applies as much to subgroup analysis as it does to RCTs with insufficiently large samples, spillovers, and extensive attrition rates. We do not wish to engage in extensive methodological commentary here, so we will merely point out that at least for *well-designed* subgroup analyses, it seems plausible to think that some remaining risk of bias might sometimes appear to be an acceptable price to pay in exchange for gaining access to potentially important information on effect distributions, information that would otherwise remain entirely unavailable.

Despite such considerations, standard methodological guidelines maintain that subgroup analyses generally have low credibility (JPAL 2017), should not be explicitly incorporated into general assessments of the effectiveness of interventions (What Works Clearinghouse 2014, 28), and may be ignored even if they are reported (What Works Clearinghouse 2017, 18). The upshot of such methodological stances is that even if subgroup evidence were more broadly available, it would not only be considered epistemically inferior to the ATE estimates reported in virtually all evaluations, but it would also likely remain ignored.

This raises a problem: at least at present, the type of average treatment-effect evidence favored and prevalently produced in evidence-based policy is *differentially useful* for the pursuit of different values and goals. It might be useful for those who are interested only in average effects, but it remains largely uninformative for those who care about distributive issues. This can have various undesirable consequences. First, policy makers who pursue distributive goals may find it more difficult to justify their calls for policy action by appeal to evidence, e.g., because evidence on distributive effects is not available; or, if it is available, because appealing to such evidence makes them more vulnerable to resistance on epistemic grounds. Second, they may also be more vulnerable to resistance that is motivated by political reasons, but is dressed up as being motivated by epistemic concerns (see Parkhurst and Abeysinghe 2016). Third, faced with these pressures, policy makers may experience incentives to shift the values that they pursue to those for which actionable evidence exists. In short, our concern is that, in several different ways, the current emphasis on RCTs, combined with the distaste for subgroup evidence expressed in methodological guidelines, can skew what kinds of policies are, and can be, justified by appeal to standard evidence-based policy evidence.

One might object that issues concerning "what works" and distributive issues can be separated at a later stage. According to this objection, policy researchers provide evidence about what works best to solve social ills such as poverty, unemployment, drug abuse, absentee fathers, or what have you, and inequality is one of the significant social ills on this list. Fixing broken families, let us say, is a goal independently of its interrelations, should they exist, with inequality. However, we believe that such issues are always entangled. Generally, policies have different effects on different individuals. Even when average effects are positive, some individuals may be made worse off or benefit less than others. Thus, depending on policy makers' distributive goals, policies can make outcomes worse even though they are effective on average—for example, because they make those who are already relatively well off significantly better off, while making those already badly off slightly worse off.

Ex-post redistribution will not usually work as a solution to this problem because we do not have information about the distributive effects of policies (because RCTs do not provide this information), and because policies tend to target outcomes that do not lend themselves to redistribution. For example, a health policy might aim to prioritize those who fare worst with respect to health. But health cannot be redistributed from the healthy to the sick after the health effects of a policy have already materialized. Distributive issues therefore have to be addressed already at the point of evidence production. The evidence produced should supply information that is helpful for decision-making, and, for those decision-makers who care about distributive effects, this requires supplying information about distributive effects.

The concerns outlined above give rise to a serious challenge for evidence-based policy, which is motivated by the idea that evidence can figure as a neutral arbiter among competing, politically and morally motivated convictions pertaining to what policies should be implemented. It seems that evidence-based policy is currently not neutral in this respect, as it can systematically bias what the evidence can be used for. Moreover, if one thinks that this should be remedied, then this requires changes to evidence-based policy methodology that encourage the production of evidence that is useful for pursuing a broader range of purposes and values. However, this could be understood to threaten value freedom in evidence-based policy, as it seems that if such changes were implemented, this would be motivated by non-epistemic, moral concerns (see Khosrowi 2019). This may be taken to suggest that non-epistemic, moral values, just like the epistemic values currently endorsed in evidence-based policy, have an important role to play in deciding what evidence is good enough for evidence-based policy.

Development Economics

Development economics examines means of promoting economic development, economic growth, and structural changes aimed at improving the well-being of populations in low-income countries. It is also concerned with the effectiveness of development aid directed to the same end. According to a World Bank paper, nearly half of the world's 6.5 billion inhabitants lived on the equivalent of less than $2 per day in 2005, and about a quarter of the world lived on the equivalent of less than $1.25 per day (Chen and Ravallion 2009). At the same time, the ratio between the top and bottom 10 percent of people (ranked by their countries' GDP per capita) was about 39:1 (Milanovic 2008). Even if extreme poverty has declined in recent years, global inequality remains a great concern (Milanovic 2016) because it is generally believed to conflict with widely held views about global justice (Boylan 2018), especially insofar as the wealth of those at the top of the distribution comes at the expense of those at the bottom.

Development aid went through a number of phases during the course of the past half century (Cohen and Easterly 2009). In the 1960s its main focus was on economic growth, so money was used to support large-scale infrastructure projects such as dams and airports. In the 1970s the emphasis was on how to assist the least well-off in developing countries directly. The 1980s was the era of structural adjustment policies following the so-called "Washington Consensus" (a set of policy prescriptions such as fiscal discipline, tax reform, and trade liberalization). In the 1990s the focus was on governance: democratization, fighting corruption, and so on. What all these phases had in common was a focus on macroeconomic factors that affect entire nations or at least regions.

In the 2010s, though, we entered the era of "randomization in the tropics" (Deaton 2011), emblematized by the award of the 2019 Nobel Memorial Prize in economics to Abhijit Banerjee, Esther Duflo, and Michael Kremer. Countless studies have followed their lead in using RCTs to investigate the effectiveness of development interventions, such as whether providing mosquito nets for free, at subsidized price, or at full cost is most effective for getting people to use them; and whether providing information, training community members, or training and organizing volunteers is the best strategy to improve community involvement in public schools. What these experiments have in common is that they address microeconomic, individual-level behavioral questions.

There is a simple reason: most macroeconomic questions cannot (realistically) be answered by RCTs, whether the policy at hand is the building of a dam or airport, the liberalization of trade, or the enforcement of anti-corruption laws. Thus, if we want an "evidence-based" answer to a policy question, we had better ask the right kind of question—which, following currently fashionable methodology, often means a micro question. Banerjee and Duflo (2011) open their *Poor Economics: A Radical Rethinking of the Way to Fight Global Poverty* by suggesting that "whereof one cannot speak" in an evidence-based fashion, "thereof one must be silent." Commenting on the debate between proponents of development aid, such as Jeffrey Sachs, and its critics, such as Dambisa Moyo, William Easterly, and others, Banerjee and Duflo ask:

> Whom should we believe? Those who tell us that aid can solve the problem? Or those who say that it makes things worse? The debate cannot be solved in the abstract: We need evidence. But unfortunately, the kind of data usually used to answer the big questions does not inspire confidence. . . . Fortunately, we don't need to be quite so defeatist.

> There are in fact answers. . . . It is just that they are not the kind of sweeping answers that Sachs and Easterly favor. This book will not tell you whether aid is good or bad, but it will say whether particular instances of aid did some good or not. (Banerjee and Duflo 2011, 4)

Ultimately, however, ignoring the "wrong kinds of questions" comes at a risk. An increasing number of development economists maintain that development aid does more harm than good. Moyo (2009, xix) contends that "aid has been, and continues to be, an unmitigated political, economic, and humanitarian disaster for most parts of the developing world." Of course, it is possible that, because we haven't had enough "evidence-based" research on development aid yet, most funded projects have been flying blind and producing negative effects. However, it is also possible that even if only demonstrably effective interventions have been funded, Moyo and other critics are right. Among other things, this is because positive effects in one place are consistent with negative effects elsewhere, positive local effects are consistent with negative effects when programs are scaled up, and positive short-run effects are consistent with negative long-run effects (e.g., because of effects on incentive structures). Ultimately, then, we cannot get around addressing the "big questions" if development is the goal.

Nudge Policy

The first two illustrations emphasized different aspects of evidence-based policy. An RCT identifies the causal effect of a policy variable X on an outcome variable Y. But without the aid of additional assumptions, which are treated with suspicion in the evidence-based policy paradigm, it can identify only the average causal effect and not the distribution of effects among the individuals in the population affected by the policy. What this shows is that certain questions cannot be addressed even when randomized experiments can be performed. The second illustration examined cases where the question of interest itself is such that it is not open to experimental investigation. Our third illustration concerns the use of psychological research to justify policy "nudging."

"Nudges" are micro-interventions that aim to affect individuals' behavior in ways that improve their well-being. More technically:

> A nudge … is any aspect of the choice architecture that alters people's behavior in a predictable way without forbidding any options or significantly changing their economic incentives. (Thaler and Sunstein 2008, 6)

By "choice architecture," the authors mean any aspect of the context in which a choice is made that may influence the decision maker's behavior. Among these architectural features are which option constitutes the default option, where alternative goods are physically located, how they are presented, what information is presented about them, and how it is presented. Libertarian paternalism is the advocacy of using nudges rather than outright coercion as policy instruments (Barton and Grüne-Yanoff 2015). Nudges are paternalistic in that they are interventions, against an individual's expressed will, aimed to benefit him or her; they are libertarian in that they do not constrain the individual's choice set.

Libertarian paternalism and nudge theory draw heavily on behavioral economics, which investigates the effects of cognitive limitations and biases on human decision making (see Mullainathan and Thaler 2001). Its results are experimentally established phenomena such as "loss aversion" (people weigh losses more heavily than gains), "non-linear probability weighting" (people overweight small probabilities and underweight large probabilities), and "hyperbolic discounting" (people discount future rewards by a factor that increases with the length of the delay). As an example, consider the "save more tomorrow" program (Thaler and Benartzi 2004), which aims to help people save money for retirement. Instead of asking employees to stash away a larger fraction of their income *now*, which purportedly leads to undersaving, the idea is to ask employees to commit to saving a share of *future* salary raises. The program exploits both loss aversion and hyperbolic discounting: loss aversion because present savings are perceived as a loss, hyperbolic discounting because people weigh future "losses" less heavily.

Randomization is not the core issue here; accordingly, libertarian paternalism does not use the "evidence-based" label. However, libertarian paternalism is frequently justified on the basis of its being firmly grounded in evidence, and the evidence in which it is grounded is experimental. Moreover, libertarian paternalism compounds the two difficulties highlighted above: it both uses misleading average results and constrains the range of questions that can be addressed.

For a policy to be paternalistic, it needs to improve people's well-being (or at least it must be justified by the claim that it does so). Libertarian paternalists understand well-being in terms of preference satisfaction, as do most mainstream economists (see Hausman and McPherson 2009). However, they depart from mainstream economics in that they do not think that people's actual choices reliably indicate what contributes to their well-being. If they did, any choice would trivially improve people's well-being and policy could not make a difference one way or another. Rather, libertarian paternalists take those preferences as relevant for the assessment of well-being that people *would* experience if they were not subject to cognitive limitations and biases such as loss aversion or hyperbolic discounting. Thus, survey evidence suggests that many people in fact prefer to save more; but they fail to act on these preferences because of cognitive limitations and biases. Nudges aim to correct for such limitations and biases while leaving other options available for those who prefer them.

But how do we know what people's underlying preferences are? Insofar as people *differ* in their saving preferences, only those who in fact want to save more should be nudged towards that outcome. However, the evidence that is used to justify nudge interventions is concerned with average or typical preferences, rather than individual preferences, and is thus insufficient to justify nudge interventions. The very survey evidence that indicates heterogeneity in people's desire to save is used to justify nudging *everyone* to save more, because such evidence cannot identify which members of a target population should and which should not be nudged. Of course, we could justify this policy by contending that individuals *should* save more, but this would use a different (objective) standard of well-being and the policy would therefore not be libertarian-paternalist in spirit.

Nudge theory also encourages victim blaming (Binns and Low 2017). People are portrayed as irrational, myopic, weak-willed, and insufficiently self-interested (Mullainathan and Thaler 2001), and thus as responsible for their adverse outcomes. This may not only blind policy analysts to alternative causes of the undesirable outcomes in question, but also foreclose policy options concerned with such alternative causes (see Saltelli and Giampietro 2017 for similar concerns about evidence-based policy more broadly). Historically sensitive economists and social scientists such as Tocqueville, Schumpeter, and Polanyi have provided strong reasons to reject the idea that economic rationality is a universal human capacity—an idea that has plagued mainstream economics since Adam Smith. Instead, they suggest, economic rationality is itself a product of the broad institutional setting within which individuals act. In focusing only on the interaction between particular choice architectures and agents' cognitive shortcomings, nudge theory diverts attention from various genuinely social factors that may importantly codetermine the putatively undesirable behaviors at issue. Among these factors may be inadequate education systems, poor incentive structures, misleading media, emaciated social relations, and dysfunctional health systems.

There are a number of reasons for thinking that such structural factors should not be ignored even when policies targeting individuals are effective. Suppose, for example, that the underlying cause of individuals' short-sighted behavior is "artificially" low interest rates, as some Austrian economists have argued (e.g. Stöferle 2015). To nudge people away from these short-sighted behaviors would then be like treating a symptom rather than the disease. If the disease is treatable (which we can find out only if we ask the right questions), treating a symptom will tend to be suboptimal because most diseases have more than one symptom (artificially low interest rates have many undesirable consequences that have nothing to do with myopic behavior), and treating all symptoms instead of the underlying disease will be neither possible nor cost-effective. Most treatments have side effects (e.g., nudging may make people less competent and less autonomous). Moreover, if the disease is not treated, the symptom will recur once the treatment has been stopped (as people will revert to short-sighted behavior when not nudged). Finally, there are moral costs. Thaler and Sunstein argue that libertarian paternalism is less problematic than traditional paternalism both because the former uses individuals' "laundered" preferences as the standard of welfare, and because the policies libertarian paternalists advocate do not foreclose, or make significantly more expensive, individuals' alternative options. However, "laundered" preferences cannot be observed, and there is a danger that when information about "what people really want" is unreliable or hard to get, policy makers will substitute their own preferences for the target population's preferences, making the proposal resemble traditional paternalism much more closely (see Rizzo and Whitman 2008; Rizzo and Whitman 2009). Finally, it simply seems unfair to "treat" individuals when the underlying cause of the social ill is a failed policy such as zero interest rates.

Methodological-Policy Debates

The preference for experimental evidence can thus constrain the types of questions we can ask and the purposes and values we can pursue on the basis of the evidence. This can skew what kinds of policies can be evaluated; it can create blind spots concerning significant distributive issues; and it can put at a disadvantage policy makers who pursue distributive goals, or who place special emphasis on prioritizing the welfare of the worst-off. The overarching question this raises is that of balancing epistemic concerns, which have to do with the production of reliable social policy knowledge, and non-epistemic concerns about the usefulness of such knowledge for the pursuit of particular moral and political values. Is it sometimes permissible to sacrifice important epistemic values (such as those underlying the methodological preference for RCTs) to produce evidence that is useful for a broader range of non-epistemic purposes? Should preferences for particular kinds of evidence be revised in light of concerns about the limited usefulness of that evidence? Or should evidence-based policy proponents bite the bullet and concede that their preferred kinds of evidence cannot address many important questions, and cannot help evidence users pursue many important purposes and values?

It is unclear that there are univocal answers to these questions. It seems more plausible that our answers will depend on many contextual details, including the types of question that users of evidence seek to address, the nature of the policy settings to which various research questions pertain, and the nature of the methods that are available for addressing these questions. All of these can vary importantly.

What is more, it does not seem that philosophers of science, such as ourselves, will be, or indeed should be, the ones deciding how these questions are eventually settled. We do not issue quality-of-evidence guidelines or methodological handbooks. Instead, such wide-ranging methodological decisions should be the product of joint deliberation involving social scientists, methodologists, policy makers and relevant stakeholders. One of our contributions here is to highlight the need of engaging in such deliberation.

There are, of course, recurring methodological debates among those who *do* issue guidelines for evidence production and use, on the one hand, and among social scientists and practitioners who are expected to follow these guidelines (see, e.g., Head 2010 and other articles in the same journal). But policy makers and stakeholders need to be involved in methodological debate, too. First, they are the intended users of the researchers' evidence. But if the purposes and goals they will pursue in practice are not adequately catered to by the kind of evidence prevalently produced, then this will tend to undermine the role of evidence in underwriting policy. Anticipating such problems, it seems reasonable to think that involving policy makers in deliberating about what constitute adequate relationships of fit between the practical and value-related aims of policy, the epistemic needs deriving from them, and the methods available to address these needs could help improve the usefulness of evidence for policy. It also seems helpful that those affected by policy interventions be involved in such deliberations since it is, at least in part, their welfare outcomes that raise the issues of distributive justice that we highlight here. One might respond that stakeholders are already involved "by proxy," as it were, since policy makers ideally care about stakeholders' welfare outcomes when deciding which policies to implement. However, beyond being the ultimate subjects of welfare analyses, stakeholders can also play an important epistemic role. Put simply, agents are sometimes in a privileged epistemic position to determine how a certain intervention has affected them.

Consider a population that was granted access to microfinance loans with overall positive average effects. One subgroup might report that nevertheless, the program enabled them to pursue unprofitable business opportunities and that they have since fallen into a "debt trap" as they had to take out high-interest loans to repay the initial loans. A collection of N = 1 narratives of this sort can be epistemically valuable, even though, or precisely because, the individual adverse effects are not easily detected in a standard RCT design. Conversely, without considering stakeholders' first-hand perspectives and focusing only on average effects, social scientists might be unable to detect individual-level adverse effects, and policy makers might be complacent in light of positive average effects. Thus, stakeholders' first-hand experiences can sometimes help elucidate, to methodologists and the producers and users of evidence alike, whether or not a certain method adequately detects the types of effects that a policy might have.

Beyond highlighting the need for a more open and inclusive methodological debate, we would like to emphasize that there is no neutral stance on balancing epistemic and non-epistemic concerns. To insist on the use of particular methods on purely epistemic grounds is to place a higher weight on epistemic than non-epistemic concerns. To resist such an orientation toward particular methods is to place a higher value on non-epistemic concerns over epistemic concerns (although one might also resist such an orientation on the grounds that the methods being treated as the "gold standard" are in fact deficient, as we initially pointed out). Since there is no neutral stance, the participants in methodological debates need to make transparent which value commitments are implied by the choices they advocate.

Such debates should also attend to the severity of the problems with different methods. It is clear that there will be simple cases where alternative methods are available that can help overcome some of the limitations that we have outlined above. Here, all it will take is some good old criticism to challenge existing methodological preferences. However, there are also going to be truly hard cases where basic features of the available methods inevitably constrain their informativeness about certain aspects of policy effects. In such cases, it might be wise for evidence to be disseminated with appropriate instructions alerting its users to important limitations and blind spots. The more interesting cases will sit in the middle. These are cases where alternative methods are available, but adopting them may come at the price of sacrificing deeply held methodological tenets. The cases discussed above suggest that, at least sometimes, it seems reasonable to sacrifice such tenets in the pursuit of producing evidence that is useful for a broader range of purposes and values. As suggested by the discussion above, it is clear that the quality of evidence about a policy cannot be assessed on purely epistemic grounds. Hence, quality-of-evidence guidelines should take into account the purposes for which evidence is generated, such that plausible evidentiary needs of policy makers, weighted by the moral and political significance of these needs, would codetermine judgements of evidentiary adequacy.

A starting point would be to replace traditional hierarchies of evidence with more complex *matrices* that couple epistemic and ethical dimensions of evidentiary adequacy. Such matrices would not assess the quality of evidence *simpliciter*, but only relative to a variety of purposes for which evidence may be used, and to the significance of the non-epistemic stakes involved. For instance, it is clear that there will be cases where, although there is some tension between epistemic and non-epistemic concerns, the non-epistemic concerns at stake are just not very significant. Not all public policies raise distributive-justice concerns, such as when policies produce windfall benefits in well-off populations without hurting anyone else. But there are surely other cases, such as those in development economics, where we start out with highly vulnerable populations in precarious circumstances, and at least some programs have the capacity to make these agents even worse off. Here, it seems plausible to think that non-epistemic concerns may justifiably override epistemic concerns.

The bottom line is: the right mix of epistemic and non-epistemic considerations depends on the details of the case.

Evidential Adequacy and Normative Considerations

In light of the normative considerations we have tried to bring to bear, what constitutes evidentiary adequacy?

We suggest that ideally, adequate evidence (a) should answer to the research question at hand (and not to some merely related question that is more tractable to a putatively surer method); (b) that it should be produced by a method that is applicable to the intended domain, and that (c) does not contradict the background knowledge we have; and, importantly, (d) that it should help evidence users realize their intended value schema. There are important questions that need to be addressed in fleshing out such an ideal, however. For instance, should this apply to individual studies or whole bodies of evidence? It is clear that evidence from particular methods can be highly limited in its scope of feasible use. But it also seems that there can be complementary relations between different tokens of individual cases of severely limited evidence. Here, it would seem appropriate that evidential adequacy applies to bodies of evidence, and takes into account such complementary relations. (For a proposal for a framework for aggregating evidence, see Reiss 2015.)

We also need to think about what to do when the values to be pursued on grounds of evidence are not fully determined at the time when the evidence is produced. How can the adequacy of evidence for the pursuit of yet to be determined values be assessed and promoted here? Should evidence be produced with a view towards being useful for the pursuit of the broadest range of value schemas? Other important questions include: what purposes and values should we take into account when assessing whether evidence is adequate? Who gets to decide which values are important? When different values imply conflicting demands for the type of evidence that should be produced, how can we adjudicate such conflicts?

One might also ask whether there are perhaps cases in which we should limit, instead of increasing, the range of purposes and values for which evidence should be adequate. Consider cases in which highly detailed and discriminate effectiveness evidence might allow policy makers to target interventions so they affect only those individuals who will benefit the most. This might be highly cost effective but, depending on who the individuals are who are likely to benefit the most, might also aggravate existing inequalities, which may threaten to undermine important values such as fairness. For instance, one might think that it is preferable to prioritize less privileged individuals, even if these individuals benefit less from a policy than others, and thus incur an efficiency cost. On such a view, it might seem problematic to produce highly discriminate effectiveness evidence and make it available to public servants who are under increasing pressure to maximize the cost-effectiveness of public services, and to do so in ways that gloss over important value-related issues. So are there sometimes reasons to discourage the production of certain kinds of evidence, such as when its existence might incentivize policies that are likely to have controversial properties in terms of distributive justice and when there is predictably little space for debate about whether these properties seem acceptable? Or should discouraging the production of highly discriminate effectiveness evidence be rejected as illegitimate censorship?

Highlighting these issues, as well as emphasizing that there is no neutral stance towards the balancing of epistemic and non-epistemic considerations, are important first steps in facilitating an explicit debate about evidence-based policy: not about whether using evidence to inform policy is a good idea, but about what kinds of evidence should be produced and used and what role values should play in making such decisions.

REFERENCES

Abraham, Katherine, Ron Haskings, Sherry Glied, Robert Groves, Robert Hahn, Hilary Hoynes, Jeffrey Liebman, Bruce Meyer, Paul Ohm, Nancy Potok, Kathleen Rice Mosier, Robert Shea, Latanya Sweeney, Kenneth Troske, and Kim Wallin. 2017. "The Promise of Evidence-Based Policymaking: Report of the Commission on Evidence-Based Policymaking." Washington, DC: Commission on Evidence-Based Policymaking.

Banerjee, Abhijit, Emily Breza, Esther Duflo, and Cynthia Kinnan. 2017. "Do Credit Constraints Limit Entrepreneurship?" Working Paper, Buffett Institute Global Poverty Research Lab.

Banerjee, Abhijit, and Esther Duflo. 2011. *Poor Economics: A Radical Rethinking of the Way to Fight Global Poverty*. New York: PublicAffairs.

Banks, Gary 2009. "Evidence-Based Policy Making: What Is It? How Do We Get It?" ANU Public Lecture Series, presented by ANZSOG, 4 February. Canberra: Australian Government Productivity Commission.

Barton, Adrien, and Till Grüne-Yanoff. 2015. "From Libertarian Paternalism to Nudging—and Beyond." *Review of Philosophy and Psychology* 6(3): 341–59.

Binns, Colin, and Wah Yun Low. 2017. "Nobel Prizes, Nudge Theory, and Public Health." *Asia Pacific Journal of Public Health* 29(8): 632–34.

Boylan, Michael, ed. 2018. *The Morality and Global Justice Reader*. London: Routledge.

Byrne, David, Wendy Olson, and Sandra Duggan. 2009. "Causality and Interpretation in Qualitative Policy Related Research." In *Handbook of Case-Study Research*, ed. David Byrne and Charles Ragin. Thousand Oaks, Calif.: SAGE.

Cabinet Office. 2013. "What Works: Evidence Centres for Social Policy." London: Cabinet Office.

Chen, Shaohua, and Martin Ravallion. 2009. *The Developing World Is Poorer than We Thought, But No Less Successful in the Fight against Poverty*. Washington, DC: Development Research Group, World Bank.

Cohen, Jessica, and William Easterly. 2009. "Introduction: Thinking Big versus Thinking Small." In *What Works in Development: Thinking Big and Thinking Small*. Washington, DC: Brookings Institution.

Deaton, Angus. 2011. "Instruments of Development: Randomization in the Tropics, and the Search for the Elusive Keys to Economic Development." Working Paper, National Bureau of Economic Research.

Deaton, Angus, and Nancy Cartwright. 2017. "Understanding and Misunderstanding Randomized Controlled Trials." Working Paper, National Bureau of Economic Research.

Hausman, Daniel M., and Michael S. McPherson. 2009. "Preference Satisfaction and Welfare Economics." *Economics and Philosophy* 25(1): 1–25.

Haynes, Laura, Owain Service, Ben Goldacre, and David Torgerson. 2012. "Test, Learn, Adapt: Developing Public Policy with Randomized Controlled Trials." Behavioral Insights Team. London: Cabinet Office.

Head, Brian. 2010. "Evidence-Based Policy." *Policy and Society* 29(2): 77–94.

Heckman, James. 1992. "Randomization and Social Policy Evaluation." In *Evaluating Welfare and Training Programs*, ed. Charles F. Manski and Irwin Garfinkel. Cambridge, Mass.: Harvard University Press.

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

JPAL. Abdul Latif Jameel Poverty Action Lab 2017. https://www.povertyactionlab.org/methodology/how/how-obtain-results.

Khosrowi, Donal. 2019. "Trade-Offs Between Epistemic and Moral Values in Evidence-Based policy." *Economics and Philosophy* 35(1): 49–71.

Kitcher, Philip 2001. *Science, Truth and Democracy*. Oxford: Oxford University Press.

Kitcher, Philip 2011. *Science in a Democratic Society*. Amherst, N.Y.: Prometheus Books.

Milanovic, Branko. 2008. "Even Higher Global Inequality than Previously Thought: A Note on Global Inequality Calculations Using the 2005 International Comparison Progam Results." *International Journal of Health Services* 38(3): 421–29.

Milanovic, Branko. 2016. *Global Inequality: A New Approach for the Age of Globalization*. Cambridge, Mass.: The Belknap Press of Harvard University Press.

Moyo, Dambisa 2009. *Dead Aid: Why Aid Is Not Working and How There Is a Better Way for Africa*. New York: Farrar, Straus and Giroux.

Mullainathan, Sendil, and Richard H. Thaler. 2001. “Behavioral Economics.” *International Encyclopedia of Social Sciences*. Oxford: Pergamon Press.

Parkhurst, J., and S. Abeysinghe. 2016. “What Constitutes ‘Good’ Evidence for Public Health and Social Policy-Making? From Hierarchies to Appropriateness.” *Social Epistemology* 30: 665-79.

Parsons, Wayne. 2002. “From Muddling Through to Muddling Up: Evidence Based Policy Making and the Modernisation of British Government.” *Public Policy and Administration* 17(3): 43-60.

Pawson, Ray. 2006. “Evidence-Based Policy: A Realist Perspective.” Thousand Oaks, Calif.: SAGE.

Reiss, Julian. 2015. “A Pragmatist Theory of Evidence”. *Philosophy of Science* 82(3): 341-62.

Rizzo, Mario, and Douglas Glen Whitman. 2008. “Little Brother is Watching You: New Paternalism on the Slippery Slopes.” New York University Law and Economics Working Papers.

Rizzo, Mario and Douglas Glen Whitman 2009. “The Knowledge Problem of New Paternalism.” *Brigham Young University Law Review* 2009(4): 905-968.

Saltelli, Andrea, and Mario Giampietro. 2017. “What Is Wrong with Evidence Based Policy, and How Can It Be Improved?” *Futures* 91: 62-71.

Scriven, Michael. 2008. “A Summative Evaluation of RCT Methodology and an Alternative Approach to Causal Research.” *Journal of MultiDisciplinary Evaluation* 5(9): 11-24.

Stöferle, Ronald-Peter. 2015. “The Unseen Consequences of Zero-Interest-Rate Policy.” URL = https://mises.org/library/unseen-consequences-zero-interest-rate-policy.

Thaler, Richard, and Shlomo Benartzi. 2004. “Save More Tomorrow: Using Behavioral Economics to Increase Employee Saving.” *Journal of Political Economy* 112(1): S164-187.

Thaler, Richard H., and Cass R. Sunstein. 2008. *Nudge: Improving Decisions About Health, Wealth, and Happiness*. New Haven: Yale University Press.

Varadhan, Ravi, and John Seeger 2013. “Estimation and Reporting of Heterogeneity of Treatment Effects.” In *Developing a Protocol for Observational Comparative Effectiveness Research: A User’s Guide*, ed. P. Velentgas, N. A. Dreyer, P. Nourjah, S. R. Smith, and M. M. Torchia. Rockville, Md.: Agency for Healthcare Research and Quality.

What Works Clearinghouse. 2014. “What Works Clearinghouse: Procedures and Standards Handbook v.3.0.” Institute of Education Sciences, U.S. Department of Education.

What Works Clearinghouse. 2017. “What Works Clearinghouse: Procedures Handbook v.4.0.” Institute of Education Sciences, U.S. Department of Education.