ANNUAL REVIEWS

Better Government, Better Science: The Promise of and Challenges Facing the Evidence-Informed Policy Movement

Jake Bowers[1] and Paul F. Testa[2]

[1]Department of Political Science, University of Illinois at Urbana-Champaign, Urbana, Illinois 61801, USA; email: jwbowers@illinois.edu

[2]Department of Political Science, Brown University, Providence, Rhode Island 02912, USA; email: paul_testa@brown.edu

ANNUAL REVIEWS CONNECT

www.annualreviews.org

- Download figures
- Navigate cited references
- Keyword search
- Explore related articles
- Share via email or social media

Keywords

Abstract

Collaborations between the academy and governments promise to improve the lives of people, the operations of government, and our understanding of human behavior and public policy. This review shows that the evidence-informed policy movement consists of two main threads: (*a*) an effort to invent new policies using insights from the social and behavioral science consensus about human behavior and institutions and (*b*) an effort to evaluate the success of governmental policies using transparent and high-integrity research designs such as randomized controlled trials. We argue that the problems of each approach may be solved or at least well addressed by teams that combine the two. We also suggest that governmental actors ought to want to learn about *why* a new policy works as much as they want to know *that* the policy works. We envision a future evidence-informed public policy practice that (*a*) involves cross-sector collaborations using the latest theory plus deep contextual knowledge to design new policies, (*b*) applies the latest insights in research design and statistical inference for causal questions, and (*c*) is focused on assessing explanations as much as on discovering what works.

The evidence-informed public policy movement is a way that new data, new questions, and new collaborators can help political scientists improve our theoretical understanding of politics and also help our policy partners to improve the practice of government itself.

EVIDENCE AS OPPOSED TO WHAT?

The appeal of an evidence-informed public policy is obvious, as on its face, the alternative is unclear. Evidence as opposed to what? Political science has long studied the "what" in this query, describing and explaining the dynamics of the policy-making process and the politics that surround it (Kingdon 1984, Baumgartner & Jones 1991, Cairney 2016). Increasingly, however, policy makers are inviting political scientists and other social and behavioral scientists to participate in policy making directly.

Governments invite social scientists to collaborate in the hopes that these academics will provide new insights to inform the design of policy and new methods to help governments learn what works and what does not. Academics join such collaborations to pursue a public service mission, to interrogate existing theory with the large body of data produced by governments, and to discover new questions arising from cross-sector collaboration to challenge existing theories that were developed mostly within the academy. A wide and diverse network of governments and academics working together promises to harness the insights and methods of the social and behavioral sciences to improve the practice of government, the lives of the public, and our understanding of human behavior and institutions.

This article introduces this movement to social scientists in general and political scientists in particular, many of whom are familiar with the concepts and principles of evidence-informed policy making but may be less aware of the growing opportunities to participate in this process. To do so effectively requires understanding two distinct roles played by evidence in this process—evidence as evaluation and evidence as insight. Highlighting this distinction helps us understand and address some of the most important criticisms of evidence-informed policy making. Many of the most common objections arise from a too narrow conception of the role of evidence in efforts to learn what works.

We focus much of our discussion on applications of insights from the behavioral sciences to public policy. We do so for two reasons. First, the growing recognition by policy makers that an understanding of human behavior can improve policy outcomes has opened the door for social scientists from a diverse range of disciplines to play an active role in the design and analysis of policy. Second, the way these collaborations have applied behavioral insights to public policy not only illustrates the dual use of evidence for the design and evaluation of policies, but also suggests ways in which this process can strengthen the link between insights and evaluations. Doing so is crucial to realizing the full promise of evidence-informed policy making for government, science, and citizens.

We begin by discussing the dual functions of evidence in policy making. First, we review the evolution of evidence as evaluation, highlighting the central role randomized field experiments have played. Next, we turn to more recent applications of behavioral insights to policy. Our review is by no means exhaustive [see Shafir (2013) for an extensive review of recent applications within the United States and OECD (2017) for a summary of applications around the world]. We use this discussion to reframe some common critiques of evidence-informed policy making in terms of the relationship between evaluation and insights. We show how ongoing efforts to apply behavioral insights to public policy have benefited from adopting and adapting best practices of

"good science"—e.g., credible research designs, transparent and open evaluation—and argue that such efforts create opportunities for collaboration between academics and policy makers that are often well suited to address larger questions of mechanism, context, and generalization. We hope to convince scholars of the tremendous potential such collaborations hold for addressing not only questions within our discipline but also the problems that face our society.

THE MEANING OF "EVIDENCE" IN PUBLIC POLICY

What does "evidence" mean? What does it mean to "base" a policy on evidence or to create a policy "informed by" evidence? When academics collaborate with government experts to solve specific policy problems, the term evidence can refer either to the past peer-reviewed studies that warrant belief in some theory or explanation (e.g., "Evidence from lab experiments suggests that social comparison can change behavior") or to future studies that will assess the success of the new policy intervention (e.g., "This evaluation of the new policy provides evidence that social comparisons can reduce opioid prescribing among doctors").[1]

The epistemic authority of both the evidence-as-evaluation approach and the evidence-as-scientific-consensus or what we are calling the evidence-as-insight approach arises from the same sources that give science its power to compel belief and change behavior. Insights from social, cognitive, and behavioral science enhance the generation of public policy because of the processes by which scientific consensuses are formed. Ideally, they arise from a collective effort to evaluate arguments and observations through the rigors of peer review. This evidence base, in principle, reflects a system aimed at objectivity and designed to avoid any personal or systemic bias. The evidence-as-evaluation approach hews to the same ideals: A given policy idea should be judged in a way that should share the epistemic authority of science in being impersonal, transparent, and unbiased.[2]

This idea that policy should be created by using knowledge that is collective as opposed to individualistic, and objective as opposed to subjective, is not new. The closest ancestor of the evidence-informed policy movement is the evidence-based medicine movement. To reduce medical costs and errors, a group of doctors and researchers turned to the idea that "the evidence base" or "the scientific consensus" should guide medical decisions rather than the expert judgments of individual doctors (Sackett et al. 1996, Sackett 1997, Giacomini 2009, Bluhm & Borgerson 2011, Djulbegovic & Guyatt 2017). They envisioned better health outcomes resulting from doctors following guidelines derived from dispassionate syntheses of the results of preregistered randomized controlled trials (RCTs) that had gone through blind peer review than from doctors following the evidence of their own idiosyncratic experience to guide clinical decisions. Evidence-based medicine has provided a template for evidence-based policy more broadly; "good evidence" would arise from the same social and technical processes that have yielded scientific evidence, a social process that famously uses theory and careful research design to overturn arguments based on the authority of conventional wisdom, religion, or individual expertise.

---

[1]Systematic measurement and observation also provide forms of evidence to governments, and since we know that observation is theory laden, often a simple description can catalyze policy change. For example, upon learning that one-fifth of all families receiving food aid lose their benefits each year even when their income does not change (Prell 2013), many policy makers would ask both why this happens and how it might be prevented.

[2]Of course, real scientists are also real humans, and so their own scientific objectivity is more of an ideal than a fact. Yet, by binding itself to certain institutions, the academic community has managed, in sometimes circuitous manners, to cumulate more or less impersonal understanding in multiple areas of investigation. See Reiss & Sprenger (2014) on the epistemic authority of science and the idea of scientific objectivity from the point of view of the philosophy of science.

In this article, we refer to the "evidence-informed" public policy movement rather than using the more popular term "evidence-based" because no scientific consensus alone has been enough to dictate a public policy. Instead, the scientific consensus and academics themselves tend to play a role in collaboratively creating new public policies; the evidence base and its interpreters, the academics, inform policy rather than dictate it.

Proponents of an evidence-informed policy-making process, a decade or two behind evidence-based medicine in its growth, tend to emphasize two distinct ideas: not only that policy makers and legislators should justify new policies using the scientific consensus, but also that governments should learn about the effectiveness of policies (new and status quo) using scientific methods. For example, the US Commission on Evidence-Based Policymaking (2017, p. 1), established by an Act of Congress in 2016, emphasizes the idea of evidence creation; the Commission "envisions a future in which rigorous evidence is created efficiently, as a routine part of government operations, and used to construct effective public policy." In principle, then, evidence serves two roles in policy creation, and those two roles can be combined—for example, the Office of Evaluation Sciences (OES), the behavioral insights unit of the US federal government, emphasizes the idea of building new policy interventions using the scientific consensus as well as randomized field experiments and reproducible and transparent research practices to assess the effectiveness of these new interventions (see **https://oes.gsa.gov/about/**). In practice, however, many debates around evidence-informed policy making sometimes obscure, conflate, or ignore these two roles, and so we next consider the roles of evidence for evaluation and insight separately to see how they can be productively linked.

EVIDENCE AS EVALUATION: USING RANDOMIZED FIELD EXPERIMENTS TO CRAFT INTERPRETABLE COMPARISONS

The evidence-as-evaluation approach to academic–practitioner collaborations changed the public debate about welfare and healthcare policies in the 1970s and 1980s, when firms such as Abt and RAND worked with the US federal government to field large-scale RCTs (Manning et al. 1987, Newhouse et al. 1993, Gueron & Rolston 2013). Typical of policy debates, the discussion at the time combined disagreements about values (e.g., "Providing free healthcare is wrong") with disagreements about effects (e.g., "Free healthcare will cause needless visits to the doctor"). Randomized trials promised to settle the second kind of debate: If an objective process could answer the empirical questions, then debate about the values and politics questions could be more fruitful.

The idea of randomization as a tool to address theoretical questions about political behavior and political psychology took off in political science in the late 1980s and early 1990s with survey experiments (Gaines et al. 2007) and lab experiments (Morton & Williams 2010, Iyengar 2011). Field experiments (Gerber & Green 2012) soon followed in the late 1990s.[3] Randomized field experiments in political science involved collaborations between nongovernmental organizations (NGOs) and academics from the beginning; it was, and is still, too difficult and costly if not unethical for academics to directly intervene in the political process without a nonacademic partner. Governments, too, began to collaborate with political scientists on such projects. For example, Bhatti et al. (2015, 2017) present voter turnout experiments done in direct collaboration

---

[3]In fact, field experiments in political science began as early as the 1930s, but they were rare thereafter. See Druckman et al. (2006) and Gerber & Green (2017) for their history. Early field experiments focusing on voter turnout were performed in collaboration with civic groups (Gerber & Green 2000, Morton & Williams 2008). The website of Evidence in Governance and Politics (EGAP; **http://egap.org**) provides many more examples of randomized field experiments, mostly focusing on topics in developing countries, designed and fielded in collaboration with NGOs.

with the Danish government. Groups within government, such as the Behavioral Insights Team in the United Kingdom and the OES and The Lab @ DC in the United States, soon joined the big research consulting firms (e.g., Abt, RAND, MDRC, Mathematica Policy Research), academic–practitioner collaboration-oriented research NGOs (e.g., J-PAL, EGAP, and ideas42), and private firms (e.g., DeLoitte and McKinsey) in designing, fielding, analyzing, and interpreting the results from field experiments meant to answer the question, "Did it work?"

Organizations use such experiments to learn whether a given policy or tactic worked well in a given context, at a given moment in time, compared to some other policy or tactic such as the status quo. If the question is whether policy X works better than policy Y, then a randomized research design, in principle, provides clear and easy interpretations of comparisons of the effects of policy X versus policy Y. We have known about the power of randomization at least since Fisher (1925, 1935) and Neyman [1990 (1923)] each built a version of statistical inference on the basis of random assignment. Fisher (1935, ch. 2) famously showed that randomization could be a "reasoned basis" for statistical inference about causal claims, although the use of randomization to make fair comparisons goes back further, perhaps to the psycho-physical experiments of Peirce & Jastrow (1885). The idea that an RCT provides clarity of comparison is what Kinder & Palfrey (1993) meant when they referred to experiments as creating "interpretable comparisons." A report that said that policy X worked better than policy Y could not be attacked on the grounds that the comparison was unfair—that those exposed to policy X were wealthier or healthier, for instance, than those exposed to policy Y—because randomization creates fair comparisons that can be easily interpreted as caused by the randomization alone. Random assignment, after all, would ensure no systematic differences in the kinds of people exposed to the two policies. Further, randomization allows researchers in the middle of policy debates to side-step certain thorny, yet secondary and distracting, questions of statistical method: When asked to justify analytic choices of standard errors, estimator, statistical test, or confidence intervals, researchers can refer to the design of the study itself rather than rules of thumb or other arguments from authority. The most famous example of this simplicity in statistical analysis comes from Fisher (1935, ch. 2), who introduces a statistical hypothesis test using eight cups of tea in which the only assumption to be justified is that the cups of tea were presented in a random order.[4] This clarity of comparison and method has enabled discussion about "what works" to focus on the substance. If a large RCT has shown that policy X is better than policy Y, then policy makers in NGOs and governments are able to argue in favor of policy X, and perhaps replicate and extend the study to learn more. If the study did not show evidence in favor of policy X, then the organization could use the lack of evidence to generate new ideas and to motivate replication and extension as well.

The task of learning what works is clearly aided by randomization and other designs that can demonstrate the effect of some policy or change while maintaining focus on the substance. However, if evidence is generated without some theory of change, some insight into the why and how of the intervention, the process of learning what works is likely to be slow, circuitous, and costly, as what works in one time, place, and context is not evidence of what works in general, nor a guarantee of what will work elsewhere (Cartwright & Hardie 2012).

EVIDENCE AS INSIGHT: USING BEHAVIORAL SCIENCE TO CREATE NEW PUBLIC POLICY

Even as RCTs began to show their power for policy evaluation, another, sometimes overlapping group of scholars began to focus on the translation of the scientific consensus into policy ideas.

---

[4]Contrast this with the arguments about data modeling assumptions common in academia.

Students of human decision making (mostly from psychology and economics) began to influence policy on the creation side while practitioners of randomized experiments and causal inference worked on the evaluation side. The early work on decision making within psychology (e.g., Kahneman & Tversky 1979) helped give rise to the field of behavioral economics (see Thaler & Ganser 2015, Thaler 2016, and CASBS 2018 on the history of behavioral economics). Together, this research helped launch a movement to use insights from social and behavioral sciences to improve policy.[5]

The popular book *Nudge* by Thaler & Sunstein (2008) further inspired this effort in the policy world. The pioneering Behavioral Insights Team, also known as "The Nudge Unit," founded in the UK Prime Minister's Cabinet Office in 2010, showed that such an approach could be put into practice. In 2015, President Obama signed Executive Order No. 13707 (3 C.F.R. 56365-67), instructing the federal agencies to attend to behavioral science as a part of the policy-making process. Organizations such as OES as well as the White House Social and Behavioral Sciences Team formally established by the executive order helped agencies implement this directive with a combined approach that tested nearly every one of their policy creations with an RCT (Congdon & Shankar 2015, 2018; Benartzi et al. 2017).

The behavioral-insight approach to evidence-based policy making is only one way that the scientific consensus can play a role in suggesting new avenues for policy creation, but we discuss it because it is growing in popularity and impact (e.g., Shafir 2013, OECD 2017). It is an example of evidence as insight or evidence as explanation in addition to evidence as evaluation. Like the evaluation-based efforts, the insight approach shares a general belief that findings generated from rigorous research studies (including policy evaluations) should help justify public policy where the behavior of individual humans is a focus. If humans do not react as expected to tax credits, for example, then the introduction of tax credits will not achieve its goals. What distinguishes this movement from more evaluation-focused efforts is its particular emphasis on the relevance of insights from behavioral science to the design of public policy. To illustrate this approach in practice, we discuss the default effect—specifically, the role of defaults in retirement savings.

The Default Example

One of the clearest examples of how evidence-as-insight has shaped public policy comes from the domain of retirement savings. Most Americans do not save enough for retirement (Morrissey 2016). One possible solution to this problem is to try to incentivize saving for retirement by means of the tax code. Yet, even with tax incentives and matching contributions from employers, many individuals eligible for programs like 401(k)s and Individual Retirement Accounts do not save enough or do not save at all (Munnell et al. 2012). Even those who do use such programs do not save at the rate that a rational actor would save and do not save enough for a comfortable retirement without the need for extra assistance (Benartzi & Thaler 2013).

Human beings often do not behave the way that rational actors would. Bettinger et al. (2012) found that, although the benefit of saving thousands of dollars on college tuition makes the cost of a four-hour effort to fill out a form worthwhile for any rational actor, an easier form-completion process caused more young people to take advantage of federal college loans (Bettinger et al. 2012). Thaler & Ganser (2015) explain how economics turned to psychology as the rational actor–based psychological microfoundations of earlier economics failed to explain a growing number of economically relevant outcomes, including retirement saving. Today, the list of cognitive biases

---

[5]"Behavioral science" is a catchall term for research from psychology, cognitive science, behavioral economics, and other fields in which human action is the focus of explanation.

by which actual human behavior diverges from the predictions of standard rational actor models is quite long.

The idea of a default option arose from work in psychology and economics that sought to develop theoretical understandings of seemingly anomalous behavior and provide practical advice to guide policy design in a world in which rational cost–benefit models often fail to predict people's actual behavior. A default option is the option that a chooser would receive if the chooser made no active choice. To improve retirement savings, for example, a policy maker could set automatic paycheck deductions for retirement savings at 5% in the hopes that rational actors would switch away from the default if they thought it was not optimal for them and that regular humans would find lack of action easier and thus achieve their own long-term goal of saving more for retirement. Attempts to harness the default effect have produced some successful public policies (e.g., Gale et al. 2005, Beshears et al. 2008). For example, Madrian & Shea (2001) find that moving from a regime in which individuals had to actively choose a savings plan to one in which they were automatically enrolled and permitted to opt out produced a 50-percentage-point increase in participation. Of course, getting people to enroll in retirement plans does not guarantee that people will save adequately for retirement. Automatic enrollment can increase participation, but individuals in such programs often contribute at low default rates of 2–4% (Choi et al. 2004, Madrian 2014). Thaler & Benartzi (2004) describe one behaviorally informed solution to this problem in which employees at one firm were offered the opportunity to meet with a financial consultant. Almost all were advised that they needed to be saving more for retirement, and about 25% chose to increase their contributions to the recommended 5% after meeting with the consultant. Individuals who said they could not afford to increase their contribution were offered the chance to enroll in the Save More Tomorrow plan, which tied increased savings rates to future pay raises. Three and a half years later, participants in the plan had an average contribution rate of about 13.6%, which was 4.6 percentage points higher than those who had increased their savings rate after the initial consultation without the Save More Tomorrow plan. In addition to automatic enrollment, the principle of automatic escalation of contributions (a form of process default) is an increasingly common feature of savings plans offered by US employers (Benartzi & Thaler 2013).

Why are defaults the default example of the way behavioral insights can work in public policy? One reason is that defaults work in a wide array of settings. Comparing rates of organ donation, Johnson & Goldstein (2003) find that the lowest effective consent rate among countries with opt-out systems is 85.9%, nearly 60 percentage points higher than the highest consent rate among countries requiring explicit consent (27.5% in the Netherlands). Similarly, evidence from both the lab and field suggests individuals are more likely to choose "green" energy options when these options are the default (Pichert & Katsikopoulos 2008, Sunstein & Reisch 2014).

Second, compared to other policy tools such as incentives, sanctions, and mandates, defaults are a relatively "low-touch" intervention—what Thaler & Sunstein (2008, p. 6) call a nudge: "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." A nudge is consistent with the principles of what they call libertarian paternalism when it preserves the choice set (i.e., does not change the possibilities for a person's action), is cheap and easy to avoid or opt out of, and leads to outcomes that individuals themselves would prefer (Thaler & Sunstein 2003).[6] Since governments always act to change behavior—by building a road here and not there, by subsidizing education for this person and not that person, etc.—policy makers find it easy to justify

---

[6]For some critiques of the concept of libertarian paternalism, see Hausman & Welch (2010) and Gigerenzer (2015); for a response, see Sunstein & Thaler (2003) and Sunstein (2015).

behaviorally informed approaches, especially if the program designers preserve the freedom of action and autonomy of the public.

The third reason why defaults are an appealing example of behavioral insights in policy is that they appear to operate through at least three behavioral mechanisms. First, many decisions require physical and mental effort, and so choosing the default (or making no choice at all) has lower transaction costs (Choi et al. 2003, Johnson & Goldstein 2013). Yet default effects are also found in experimental settings where such costs are absent (Samuelson & Zeckhauser 1988, Dinner et al. 2011). Second, some suggest the power of defaults can be attributed to psychological factors and cognitive biases, such as loss aversion (Tversky & Kahneman 1991), endowment effects (Kahneman et al. 1990), and time-inconsistent preferences (Kahneman & Tversky 1979, Pronin et al. 2008). The Save More Tomorrow program was a behaviorally informed intervention designed to leverage these principles to counteract the tendency of individuals to choose low retirement contribution rates in favor of more immediate access to money. Finally, some suggest that defaults provide an implicit endorsement by experts (McKenzie et al. 2006). As such, defaults may be most effective for individuals who lack expertise or experience in a particular area. For example, Löfgren et al. (2012) find that defaults had little effect on the decision to use carbon offsets for those attending an environmental conference.

The default example has been theoretically fruitful for social science even as it has been useful for government and improved the lives of people. The instances of positive and null effects have raised new questions for students of human decision making in both psychology and economics because there is no single clear answer about why the default effect works so well. This is an area in which an evidence base from the academy informed the creation of public policies, and the evaluation of, and experience with, those policies raised new questions for the academy in turn.

Broader Applications of Behavioral Insights

Defaults are just one example of a broader set of concepts, principles, and tools employed to conduct behaviorally informed policy making. Many of the underlying insights should be familiar to political scientists, as concepts such as framing, heuristics, cues, bounded rationality, social norms, and peer influence are commonly used to explain aspects of political behavior and politics more broadly (e.g., Kuklinski & Quirk 2000, Chong & Druckman 2007, Lodge & Taber 2013). Others, such as cognitive load (Sweller 1994) and ego depletion (Hagger et al. 2010, Carter et al. 2015, Friese et al. 2018), are more common in psychology and less common in political science.[7]

Practitioners are typically less concerned with specific models of cognition and more focused on the practical implications of behavioral theory for policy design. People pursuing an evidence-as-insight approach have often invented catchy mnemonic acronyms to encourage the application of these principles and focus attention on the psychology of the individual. MINDSPACE, for example, is short for Messengers, Incentives, Norms, Defaults, Salience, Priming, Affect, Commitments, Ego, and was developed by the Behavioral Insights Team of the United Kingdom's Cabinet Office as a guide for policy makers to common factors known to influence behavior (Dolan et al. 2010). Similarly, Behavioral Interventions to Advance Self-Sufficiency (BIAS)—a program focused on using behavioral insights to improve outcomes for low-income children, adults, and

---

[7]Many of these concepts are often situated within more general dual-system theories of human cognition that distinguish between forms of cognition that are "fast" (System 1) and "slow" (System 2) (Stanovich & West 2000, Kahneman 2011). Evans (2008) and Evans & Stanovich (2013) provide useful reviews in psychology, Lodge & Taber (2013) demonstrate applications to political science, and Brocas & Carrillo (2014) do so in economics. For some critiques of dual-process models, see Osman (2004), Keren (2013), Gigerenzer & Gaissmaier (2011).

families, sponsored by the Office of Planning, Research and Evaluation in the US Department of Health and Human Services with the contractor MDRC—developed the acronym SIMPLER to summarize various behavioral insights applied across 15 evaluations (Richburg-Hayes et al. 2017). SIMPLER stands for Social influence, Implementation prompts, Making deadlines, Personalization, Loss aversion, Ease, and Reminders. Finally, taking some of its own advice to heart, in 2014, the Behavioral Insights Team presented the EAST framework suggesting that policies should make the desired behavior Easy, Attractive, Social, and Timely.

The desire to simplify a long and growing list of cognitive biases and behavioral tendencies into a set of easy-to-apply principles is understandable, but it is also potentially problematic. The evidence base that scholars bring to issues is more contested and changeable than these prescriptive principles may suggest. On the spectrum between unfounded belief and scientific law, most behavioral insights fall somewhere in the middle. Even in the default example, where the evidence base is large and well established, there remain outstanding questions about the mechanisms that produce changes in behavior. Thus, the process of evidence-informed policy making requires more than mapping a set of "stylized facts" onto different policy problems (Hirschman 2016, Gelman 2018). Instead, it requires that we conceive of the task of learning what works in terms of both evaluation and insight, such that evaluations are used not just to answer the simple question "Did it work?" but also to explore "Did it work the way theory suggests it should?" Doing so can help address many of the common objections raised about evidence-informed policy making to which we now turn.

ROADBLOCKS ON THE WAY TO EVIDENCE-INFORMED POLICY

If every medium-sized city and county, US state, and OECD nation had a small behavioral insights team practicing evidence-as-insight approaches, or a small field experimentation team practicing evidence-as-evaluation approaches, or even a team like the OES that combines the two, would we see radically improved government? Would social science generate new theories and methods to grapple with new questions? Would the public be increasingly satisfied with the roles of both social science and government in their lives?

We think the answers to such questions can and should be yes but that the next stage of this movement will have to address a set of challenges before social science insights and methods become fully integrated into the practice of public policy and before the social sciences can fully benefit from the extra-disciplinary challenges provided by such collaboration. We present three general classes of related challenges: problems of principle, theory, and practice. For each, we argue that many common objections arise from a conception of evidence-informed policy making as solely focused on evaluation or insight and suggest that the concerns can be addressed by creating a stronger link between the two.

Problems of Principle and Politics

Problems with the principle of evidence-informed policy making are often couched in terms of concerns about paternalism. Critiques of paternalism can be either general or particular. General critiques argue that evidence-informed policy making will expand the government's ability to intervene in the lives of citizens in ways that necessarily constrain choice, limit freedom, or coerce behavior of at least some citizens. For example, in the context of evidence-based medicine, a critic might worry that refusing to cover some treatments not backed by a rigorous systematic review may limit medical innovation and prevent some people from receiving a potentially life-saving procedure or medicine. Particular critiques focus on the potential for evidence to be politicized

and used to support a particular political goal rather than provide an objective evaluation of what works. Here, the concern is that the evidence-informed policy maker is not an honest broker, dispassionately evaluating the facts, but a motivated salesperson, producing "policy-based evidence" guaranteed to support some preordained goal.

Proponents of evidence-informed policy making have both normative and practical responses to these concerns. They can highlight the extent to which many behaviorally motivated interventions are consistent with the principles of libertarian paternalism outlined by Thaler & Sunstein (2003, 2008); policies built around insights into how individuals are likely to behave under different scenarios (choice architectures) need not coerce behavior to improve welfare. More broadly, those who defend the behavioral-insights approach and evidence-informed policy making in general would make the following three points. First, even though the protection of individual freedoms is a role of government, critiques focusing on paternalism characterize this role either too narrowly or passively; contemporary governments clearly do more than preserve individual liberty. Second, there are other conceptions of a good government beyond the passive preservation of liberties, such as the solving of collective goods problems and the guarding of justice and equality. Third, contemporary government plays a large role in the lives of people whether or not all of its activities are carefully calibrated to accord with any given normative justification. Practitioners of evidence as evaluation would also note that these efforts may often produce evidence of what does not work, leading to the retooling of ineffective programs (e.g., Garner et al. 2013). Furthermore, practitioners could argue that the process by which evidence is generated can help insulate policy making from charges of political bias.[8] There is growing consensus about best practices that enhance the transparency and credibility of research in general—for example, developing a set of standard operating procedures (Lin & Green 2016), preregistering designs (Humphreys et al. 2013), and providing access to data and replication materials (McKiernan et al. 2016)—which increase the integrity of the policy-making process.[9]

Some of these approaches and principles are being modified for use in government. For example, the OES Research Integrity process involves a commitment to publish every study as well as multiple steps by which members of the team publicly preregister their own studies and review and replicate each other's work. Few academics commit to publicizing every study they begin, and most do not perceive their work as inherently the product of a team, yet these pieces of the OES process were added to the extant open-science processes of academia because the OES is a team of social and behavioral scientists within a government. Taken together, the OES process and others like it enable evidence-based policy teams to show that they are not producing policy-based evidence but rather credible evidence for evaluation as well as academic publication.

Problems of Theory and Insight

If we accept the premise that insights from social and behavioral science should inform policy and can do so while maintaining an epistemic authority independent of charges of political bias, the question then becomes whether academics actually have anything relevant to say. Consider,

---

[8]Highlighting the role of evidence-as-evaluation can also address some potential ethical concerns about behavior interventions. The act of evaluation provides an opportunity for the public and other stakeholders to assess not only whether the outcomes of the policy are desirable (e.g., more people saving for retirement or eating healthier foods) but also whether the means of achieving that goal (e.g., through our understanding of tendencies in individuals' subconscious or automatic behaviors) are acceptable.

[9]Indeed, such principles are embedded in how organizations like The Lab @ DC, which provides public access to projects through the Open Science Framework (**https://osf.io/institutions/thelabatdc/**), and OES function (see for example the OES Research Integrity Process at **https://oes.gsa.gov/methods/**).

for example, that most governments have thousands of forms collecting information from millions of people. Forms are one key way in which citizens interact with government, and we know that individual–government interactions can change how an individual acts and feels as a citizen (Skocpol 1995, Campbell 2003, Mettler & Soss 2004). Both civic and governmental efficiency gains can arise from improving forms—and there are so many forms, and so much time is spent on them, that small improvements should provide large benefits. It turns out, however, that few peer-reviewed articles grapple with this major way that the government interacts with its public, although the design of forms is of central concern for governments and may be addressed via direct input from the potential and/or past users of the forms (see, for example, The Lab @ DC's Form-a-Palooza at **https://osf.io/kf4r9/**). When asked to improve forms, most social scientists turn to the literature on the design of surveys and the cognitive science of asking and answering. Academics can also appeal to common sense and the basic science of communication: Plain language and simple graphic design ought to do a better job guiding the public and eliciting accurate information than legal language presented in small fonts. However, those researchers who have confronted these problems know that in form reform, past literature and theory provide a basis for reasoned improvisation but not the opportunity to directly translate some approach that worked in the lab into policy. The benefit of having academics participate in this process is that they help structure research designs and provide initial theory-driven intuitions to answer policy makers' immediate questions about what works while also helping to articulate, assess, and advance explanations for why these approaches worked.

The fact that the academic literature does not provide direct guidance for many, or even most, policy challenges need not stop efforts at collaboration. From the perspective of policy makers, as long as the policy improvisations based on related and better-established domains and existing governance expertise are paired with clear assessments, then little will have been lost and much gained by using collaborations to build a new base of findings, explanations, and hypotheses. From the perspective of academics, the fact that the questions about which knowledge is accumulating in academia are not always the burning questions of the day within government should be productive for science. For example, a synthesis of the research on survey response (e.g., Sudman et al. 1996, Tourangeau et al. 2000) with other research on graphic design and language could produce new insights into human communication or into the citizen–state relationship. An experimentally induced increase in enrollment in some program via better government processes in turn provides an instrument to study longer-term consequences of participation in programs and the effects of government in general.[10]

Problems of Practice and Learning

A final set of concerns arises from the general problems of learning from observation. The results of one study of one policy in one place at one moment may not teach us directly about that same policy as applied in another place and another time (Cartwright & Hardie 2012, Deaton & Cartwright 2017). Such warnings about a "crisis of generalizability" often arise in tandem with concerns about the primacy of randomized trials in the practice of evidence-informed policy making. Randomization provides multiple benefits to researcher–practitioner partnerships beyond the obvious benefits that it provides to all research designs—of ensuring no systematic differences between experimental groups and of guiding choice of statistical analysis procedures. Further, although RCTs can and do help social scientists answer "why" questions every day, if evidence equals

---

[10]See for example the somewhat surprising effects of college scholarships on degree completion detailed by Angrist et al. (2016).

an RCT, and RCTs are seen as only answering "what works" questions, then policy makers will be ill equipped to respond to changes in context, like the rise of the gig economy (De Stefano 2015) or changing climates (Gowdy 2008), and social scientists will struggle to use the collaborations to advance science itself. Giacomini (2009, p. 236) warns, in the case of medicine, that

> Equanimity about whether an intervention works prior to its test (equipoise) has lapsed into a tolerance for uncertainty about why an intervention should work at all. In this era of EBM's [evidence-based medicine's] maturity and considerable influence, one form of authority—expert opinion—has been replaced in many minds with another—evidence from well-designed RCTs.

To dramatize this problem, Giacomini (2009) considers the field of randomized studies of the health effects of remote prayer, in which people pray to God (or a god) for the healing of others without the others' knowledge. Giacomini cites 18 such studies as well as a systematic review by the Cochrane Collaborative (Roberts et al. 2009), most of which yield no evidence for an effect of remote prayer. The problem with this field, she argues, is not a lack of RCTs but rather "prayer researchers' reluctance to articulate any theory of how the prayer intervention is supposed to work" (Giacomini 2009, p. 244). More broadly, she cautions that "experimental evidence about unexplainable interventions may be not only pragmatically worthless, but even misleading or harmful" (Giacomini 2009, p. 246).

A lack of theory, black-box models of causality, problems of generalization, and arguments from authority or misunderstandings about RCTs are not simply technical concerns. Cartwright & Hardie (2012), for example, tell the story of the randomized field experiment in Tennessee showing that smaller class sizes had an effect on academic achievement there, paired with the story about a smaller class size policy backfiring in California. It turns out that the effect of class size on educational outcomes does not exist in isolation; a small class size with an underprepared teacher may produce worse outcomes than a large class size with an expert teacher. According to Cartwright & Hardie, the Tennessee results inspired the state of California to rush the hiring of new teachers to accommodate the class size policy change. Thus, classes were smaller but many were staffed by underprepared teachers. Cartwright & Hardie argue that the policy had negative consequences because of this difference in context and that, in general, the success or failure of any policy is crucially dependent on the background factors that constrain the actors. Causal processes always occur in, and depend on, context.

Of course, these concerns about how an inherently contextual, or local in time and space, set of observations can inform general statements are not new to social science (Guba & Lincoln 1994). The problem of a fetishized method is not new either; anyone can look to the history of their discipline and notice that certain approaches to observation and learning rise and fall in popularity. The problem of undervaluing answers to "why" questions may be more recent and even understandable as a focus on what works can be strategic to defuse political arguments. But we think there are good reasons why the practice of evidence-informed policy making can avoid some of the problems and concerns raised by Giacomini, Cartwright, and others.

First, the embrace of randomization and other tools for credible causal inference in observational studies with administrative data (Brady 2019) across organizations and governments will yield more clear and focused findings that attend to the context of their research because these studies are being done by a given organization to inform its own next actions. Such studies require the involvement of the people on the ground and a fair amount of "shoe leather" (Freedman 1991) in order to learn about the specifics of the problem. The social scientists in the OES have, for example, collaborated with experts in human-centered design to learn in depth about a few individuals or a few places before returning to the literature in psychology and economics and

the governmental administrators of the program under revision or creation. Most work in a government occurs in teams and is problem oriented (Watts 2017), so it is natural to deploy multiple modes of observation and expertise in the tasks of description and interpretation. Furthermore, most governments want to understand the populations and contexts in which the proposed policy may be implemented, and a pilot study may be fielded only months before overall implementation. In such studies, including those common in organizations like the OES, the context and population of the study are the context and population of the policy itself.

Second, public preregistration of analyses and designs can help limit the statistical problem of false discoveries, and increased access to data enables both replication that can detect errors and explorations that can direct further research. The fact that the results from such studies are increasingly being generated by collaborations between academics and governments carries some added benefits—larger sample sizes, fewer incentives to withhold null results, testing on populations of interest, the potential to follow changes over time as policies scale up—that directly speak to challenges of generalization.

Third, the growth in the number and quality of studies arising from academic–practitioner collaboration can in turn facilitate the meta-analyses and systematic reviews common in the fields of medicine and education. Likewise, studies themselves can be designed in a collaborative fashion to facilitate learning across contexts. EGAP has pioneered this approach, which they call the "metaketa" approach (the Basque word for "accumulation"), in which roughly five teams of researchers from around the world agree to implement the same experimental arm in each of their five different contexts, and also agree to collect the same key outcome data. Another team designs the meta-analysis and monitors individual projects from design to field to analysis, and publication of results occurs first with the meta-analysis and later with the individual teams. The first metaketa on information and accountability is now complete, and an in-depth description of the methods and procedures will be published in an edited volume (Dunning et al. 2018). This approach speeds learning about the relationship between contexts and policy interventions.

A final response to the idea that all observation is local even if we, as researchers, desire to learn in general, is to focus on what policy practitioners often call theories of change (Weiss 1997, Coryn et al. 2011). This focus on "why" avoids the misconception that RCTs are only useful for "what works" questions. If we can articulate why or how a given intervention may work, then (*a*) we can design research to target the explanation itself rather than the "Does it work?" question and (*b*) governments and organizations will be better prepared to respond to changes in the context. For example, if the policy works because people in neighborhoods know each other well, then when neighborhoods experience rapid change—perhaps because of climate change events local to the place, or an influx of newcomers due to such events elsewhere—the government can more easily predict and prepare for the changing functioning of the policy. Attention to theories of change also promises the most benefits for a theory-driven academia itself. The more evidence-based policy collaborations focus on why a particular policy might work better than another, the quicker the translation of the new research into the academic consensus and the more agile government will be in the face of change.

Lessons from Behavioral Insights for Evidence-Informed Policy Making

Many objections to evidence-informed policy making arise from the potential disconnects between evaluation and insights in the process of policy making. Evidence can seem paternalistic and political when the procedures for evaluation are not credible and transparent and when the mechanisms by which an intervention works are opaque or poorly explained. Insights can seem insufficient or ad hoc unless we conceptualize evaluations as an opportunity for learning, and the

process of evaluation can seem overly restrictive, costly, and narrow unless the results are situated within a broader learning agenda designed to articulate and clarify a theory of change.

In this section, we offer a brief discussion of what that process might look like in practice, drawing on our own experiences and the advice of others. Our goal is not to provide the definitive how-to manual but rather to highlight similarities to what scholars are already doing in their own research practice and draw attention to some features of policy collaborations that present both challenges and opportunities for learning. We focus particularly on the way behavioral insights have been applied to policy because we think it presents a clear case of how theory and insight can be more closely linked to credible evaluations.

Evidence-informed policy making often begins with a definition of problems and goals. The process is similar to that of clarifying a research question, except that the academic must be able to speak not just to existing literatures and theory but also to government agencies and stakeholders. Learning that language takes time and a considerable amount of relationship building. Collaborators must trust and understand each other, developing a set of shared goals and expectations often formalized in memorandums of understanding and data use agreements (which are also important for clarifying what data can be used for academic publication). The process can involve some salesmanship—convincing policy makers of the benefits of randomization and other tools for credible evaluation—as well as compromise so that both sides have a clear sense of how success will be measured and evaluated and what actions the agency might take if findings differ from what is expected in these planning stages.

Upon agreeing that behavioral insights might be applied to a particular policy problem, practitioners engage in diagnostic tasks: collection of evidence, reviews of past studies, ethnography, exploratory analyses of historical data, and discussions with agencies and stakeholders. Often the practitioners aim to produce a behavioral map—similar to what many in engineering and business describe as process mapping (Damelio 2016) and what Gray (2017) calls theory mapping—that outlines the various steps and bottlenecks in the policy process where insights into human behavior might be used to enhance outcomes.

The same set of practices and designs that produce credible academic research generally yield credible evidence for evaluation. For various reasons, the ideal design a researcher might implement is not always feasible in a particular context. Practitioners must be flexible and creative, ready with alternative strategies to address logistic and political constraints. For example, a researcher may need to articulate the appeal of a stepped wedge (Brown & Lilford 2006, Hemming et al. 2015) or adaptive designs (Hu & Rosenberger 2006) to policy makers concerned about the ethics of randomly assigning access to a program and be able to clarify the limitations and challenges that arise from randomizing over clusters rather than individuals (Raudenbush 1997). Perhaps the most important goal at this stage of the process is for stakeholders and policy makers to commit to a plan for how the evidence will be evaluated and interpreted before the data are collected. One path to finding agreement on what constitutes a meaningful effect or how a project's cost–benefit analysis will be used is to preregister the design of the program evaluation (Humphreys et al. 2013). While the benefits of preregistration are increasingly clear to academics, policy makers may need to be convinced of its benefits as a way to enhance both the scientific and political integrity of the results.

The final stages of this process, involving project management and analysis, are quite similar to what social scientists are likely to encounter in their own research. Issues may arise during implementation, although sometimes these problems are themselves theoretically fruitful.[11]

---

[11] For example, if teachers deviate significantly from some pilot curriculum, then future evaluations might also assess the effects of this curriculum conditional on further training and staffing (e.g., Banerjee et al. 2017).

Likewise, results may be clear and consistent with expectations, or they may be uncertain, equivocal, and only partially consistent with prior theory. Scholars and practitioners must neither oversell the results of a promising pilot nor completely abandon a project that has worked elsewhere but appears ineffective in a new context. Perhaps more than in academia, null results can hold considerable policy sway—for example, evidence that Drug Abuse Resistance Education (D.A.R.E.) had little if any effect on student behavior led schools to stop offering these programs (Weiss et al. 2008). Furthermore, because most policy interventions are conducted on populations of interest (rather than in a lab or with a convenient sample of willing participants), concerns about generalizing out of sample are muted if not moot, and questions about the ability of a promising program to scale up to serve a broader population are often the next step in the process. Banerjee et al. (2017) provides a rich discussion of this process, examining a program called Teaching at the Right Level that showed promising returns for closing educational gaps in early pilots in India, and how practitioners learned and adapted in response to both successes and failures as the program was implemented in different contexts and at greater scales across India.

Overall, the practice of evidence-informed policy making mirrors much of what scholars already do. It offers several unique opportunities in terms of access to data, populations, and experts and the ability to test theories in new contexts and over time.

A PROMISE OF BETTER SCIENCE, BETTER GOVERNMENT, BETTER SOCIETY

This reflection on evidence-informed policy making has led us to notice distinctions within the diverse movement. Different actors have deployed different strengths in pursuit of improved public policies and more efficient and compassionate governance. The focus on evaluation, on discovering what works, has received the most attention. But efforts to build a human-centered government using behavioral insights are now well established in some places and are growing at multiple levels of government across the globe. We have suggested ways to engage or even overcome the challenges facing both the evidence-as-evaluation and evidence-as-insight approaches. We recommend combining them and adding a focus on theories of change and explanation. We suggest a commitment to multi-year learning agendas that are shared between the government and academy, and we urge close ties between the government and academy—for example, a team can be anchored by full-time government employees and full-time academic researchers but also include academics and policy experts on one-year leaves as well as academics and other researchers who work on particular projects.

A focus on theory offers many benefits for science, government, and society. In particular, a greater emphasis on the "why" questions behind policies offers better incentives for scientists to participate in this process. Political science is a theory-generating discipline, after all, and publications depend on the assessment of and debates about theory. And while we have focused much of our discussion on insights, broadly defined, from the behavioral sciences, we believe there are some specific ways in which the field of political science can benefit from these collaborations.

First, we think that political science methodology will grow as it is challenged by the need for new research designs and statistical inferences in new contexts. For example, the field experiments common in government may involve the measurement of many outcomes, interventions, and/or treatment arms. The social sciences have not engaged very deeply, to date, with the related problems of testing many hypotheses or estimating many effects in such situations. The scope of interventions (with numbers of subjects often in the tens to hundreds of thousands) and the

connection to administrative data sets present a chance to combine clever strategies for causal identification—for example, adaptive designs that allocate more subjects to treatments that appear more effective over time (Murphy 2003, Kuleshov & Precup 2014)—with applications from the field of machine learning to let the data help identify heterogeneous treatment effects (Imai & Ratkovic 2013, Wager & Athey 2017).

Second, collaboration provides the opportunity to advance several fields of substantive interest in political science. For scholars of bureaucracy and policy change, collaboration presents an opportunity to study actors and processes firsthand. Likewise, the study of policy feedback can be expanded to new domains and populations of interest, and administrative data from multiple agencies can fill out the picture of how multiple interactions with different arms of government shape citizens. The behavioral focus of many collaborations can provide scholars of political behavior the chance to test theories of psychological information processing at a much grander scale, over multiple periods of time, with more dynamic measures of attitudes and/or behavior, in realistic settings. And given the global scope of this movement, the opportunity is ripe for comparative scholars willing to help governments and agencies coordinate interventions across countries and contexts.

More broadly, one of the central premises of the evidence-informed policy movement is that using evidence to inform and evaluate programs is not only good policy but also good politics. However, this is an open claim in need of evaluation at both the institutional and individual levels. Scholars need not participate directly in collaborations to learn from them whether policies informed and evaluated by evidence are more likely to overcome partisan gridlock, diminish polarization, or spread from one jurisdiction to another. Similarly, scholars of political communication and trust have an opportunity to try to understand what works in communicating information about what works. Do citizens understand and value the principles of open science? Does a commitment to rigor, transparency, and impartial evaluation make a difference in how citizens interpret controversial findings that may directly impact their daily lives? Can this commitment improve more general sentiments about government? And if the relationships between the public and government are improved via the efforts of this movement, what are the political consequences? What theories would relate trust and confidence in institutions to what other outcomes?

Since the efforts of actors in the evidence-informed policy movement are to produce studies that are difficult to refute on methodological grounds—for example, by using RCTs—political scientists and students of human behavior in general are gaining a new evidence-informed beginning for explanation. Our existing theories have implications for what we should be seeing in these studies. And perhaps these studies will lead us to confirm, discard, or elaborate our existing understandings of fundamental mechanisms of human behavior. That is, even as we encourage evidence-informed policy teams themselves to focus more on theories of change and explanation—to make research design easier, to enable a more adaptive government as context changes—we think that those who study the relationship between individuals and institutions are receiving the gift of evidence about that relationship as a side-effect of the teams working to change government. Notice also that governments themselves are asking academics to help them vary how they relate to the public; this offers a chance to learn about the operation of institutions.

A greater focus on theory is not just a self-interested ploy to create more opportunities for academics to publish. Rather, an evidence-informed policy-making process focused on theory is in the interests of government and society as well, for at least two reasons. First, it is often easier and cheaper to test implications of theory than to evaluate a program in its entirety. In some cases, the process of evidence-informed policy making yields definitive answers (e.g., smart defaults can increase savings for retirement), but often it does not. Will providing free tuition

increase the number of college graduates and boost economic growth? It turns out that programs that decrease the costs of college increase enrollment and, to a lesser extent, persistence in degree programs, but their effects on time to degree, degree completion, and subsequent employment outcomes are more mixed and uncertain (e.g., Deming & Dynarski 2010, Angrist et al. 2016, Harris et al. 2018, Nguyen et al. 2018). An evidence-informed approach to policy making need not (and often will not) provide simple yes or no answers to be useful to governments and society. By combining multiple evaluations testing components of a well-articulated theory of change, evidence-informed policy making can offer more than just a simple answer to "Did the intervention work?" It can tell us something about why it worked, or why it worked for some and not others. In the case of education, evidence-informed policy making might highlight the need to pair aid with college-prep programs or draw our attention to the structure of merit or performance requirements in such programs. Further, by leveraging the benefits that come from access to administrative data, scholars and policy makers can assess further downstream effects without having to field a completely new randomized intervention (Brady 2019).

Finally, an evidence-informed policy-making process focused on theory will help governments in the long run adapt as the world changes. If the causal effect of a given policy depends on the social cohesion of a neighborhood, and the neighborhood changes, then the policy will no longer succeed. Having a set of plausible, even competing, explanations for why a policy is working would help a government respond to the changes in the world that will depress or augment the causal effects found during evaluation processes. When efforts to learn what works are centered on both producing evidence for evaluation and insights into mechanisms and theory, they are more adaptable to changes in context. Changes in climate, technology, demography, and the economy pose significant challenges to our governments and society. We think that the kinds of collaborations modeled so far have shown great results and even greater promise to improve the lives of people, make government better, and teach us more about the social and political world.

DISCLOSURE STATEMENT

ACKNOWLEDGMENTS

LITERATURE CITED

Angrist J, Autor D, Hudson S, Pallais A. 2016. *Evaluating post-secondary aid: enrollment, persistence, and projected completion effects*. Tech. rep., Natl. Bur. Econ. Res., Cambridge, MA

Banerjee A, Banerji R, Berry J, Duflo E, Kannan H, et al. 2017. From proof of concept to scalable policies: challenges and solutions, with an application. *J. Econ. Perspect.* 31:73–102

Baumgartner FR, Jones BD. 1991. Agenda dynamics and policy subsystems. *J. Politics* 53:1044–74

Benartzi S, Beshears J, Milkman KL, Sunstein CR, Thaler RH, et al. 2017. Should governments invest more in nudging? *Psychol. Sci.* 28:1041–55

Benartzi S, Thaler RH. 2013. Behavioral economics and the retirement savings crisis. *Science* 339:1152–53

Beshears J, Choi JJ, Laibson D, Madrian BC. 2008. The importance of default options for retirement saving outcomes: evidence from the USA. In *Lessons from Pension Reform in the Americas*, Vol. 1, ed. SJ Kay, T Sinha, pp. 271–307. Oxford, UK: Oxford Univ. Press

Bettinger EP, Long BT, Oreopoulos P, Sanbonmatsu L. 2012. The role of application assistance and information in college decisions: results from the H&R Block FAFSA experiment. *Q. J. Econ.* 127:1205–42

Bhatti Y, Dahlgaard JO, Hansen JH, Hansen KM. 2015. Getting out the vote with evaluative thinking. *Am. J. Eval.* 36:389–400

Bhatti Y, Dahlgaard JO, Hansen JH, Hansen KM. 2017. Moving the campaign from the front door to the front pocket: field experimental evidence on the effect of phrasing and timing of text messages on voter turnout. *J. Elect. Public Opin. Parties* 27:291–310

Bluhm R, Borgerson K. 2011. Evidence-based medicine. In *Philosophy of Medicine*, ed. R Bluhm, K Borgerson, pp. 203–38. Amsterdam: Elsevier

Brady HE. 2019. The challenge of big data and data science. *Annu. Rev. Political Sci.* 22:297–323

Brocas I, Carrillo JD. 2014. Dual-process theories of decision-making: a selective survey. *J. Econ. Psychol.* 41:45–54

Brown CA, Lilford RJ. 2006. The stepped wedge trial design: a systematic review. *BMC Med. Res. Methodol.* 6:54

Cairney P. 2016. *The Politics of Evidence-Based Policy Making*. New York: Palgrave Macmillan

Campbell AL. 2003. *How Policies Make Citizens: Senior Political Activism and the American Welfare State*. Princeton, NJ: Princeton Univ. Press

Carter EC, Kofler LM, Forster DE, McCullough ME. 2015. A series of meta-analytic tests of the depletion effect: self-control does not seem to rely on a limited resource. *J. Exp. Psychol. Gen.* 144:796–815

Cartwright N, Hardie J. 2012. *Evidence-Based Policy: A Practical Guide to Doing It Better*. Oxford, UK: Oxford Univ. Press

CASBS. 2018. CASBS in the history of behavioral economics. Center for Advanced Study in Behavioral Science. **https://casbs.stanford.edu/news/casbs-history-behavioral-economics**

Choi JJ, Laibson D, Madrian BC, Metrick A. 2003. Optimal defaults. *Am. Econ. Rev.* 93:180–85

Choi Laibson D, Madrian BC, Metrick A. 2004. For better or for worse: default effects and 401(k) savings behavior. In *Perspectives on the Economics of Aging*, ed. D Wise, pp. 81–126. Chicago: Univ. Chicago Press

Chong D, Druckman JN. 2007. Framing theory. *Annu. Rev. Political Sci.* 10:103–26

Comm. Evidence-Based Policymaking. 2017. *The promise of evidence-based policymaking*. Rep. Comm. Evidence-Based Policymaking, Washington, DC

Congdon WJ, Shankar M. 2015. The White House social and behavioral sciences team: lessons learned from year one. *Behav. Sci. Policy* 1:77–86

Congdon WJ, Shankar M. 2018. The role of behavioral economics in evidence-based policymaking. *Ann. Am. Acad. Political Soc. Sci.* 678:81–92

Coryn CL, Noakes LA, Westine CD, Schröter DC. 2011. A systematic review of theory-driven evaluation practice from 1990 to 2009. *Am. J. Eval.* 32:199–226

Damelio R. 2016. *The Basics of Process Mapping*. New York: CRC/Productivity Press

De Stefano V. 2015. The rise of the just-in-time workforce: on-demand work, crowdwork, and labor protection in the gig-economy. *Comp. Labor Law Policy J.* 37:461–71

Deaton A, Cartwright N. 2017. Understanding and misunderstanding randomized controlled trials. *Soc. Sci. Med.* 210:2–21

Deming D, Dynarski S. 2010. College aid. In *Targeting Investments in Children: Fighting Poverty When Resources Are Limited*, ed. PB Levine, DJ Zimmerman, pp. 283–302. Chicago: Univ. Chicago Press

Dinner I, Johnson EJ, Goldstein DG, Liu K. 2011. Partitioning default effects: why people choose not to choose. *J. Exp. Psychol. Appl.* 17:332–41

Djulbegovic B, Guyatt GH. 2017. Progress in evidence-based medicine: a quarter century on. *Lancet* 390:415–23

Dolan P, Hallsworth M, Halpern D, King D, Vlaev I. 2010. *MINDSPACE: influencing behaviour through public policy*. Rep., Institute for Government and Cabinet Office. **https://www.instituteforgovernment. org.uk/sites/default/files/publications/MINDSPACE.pdf**

Druckman JN, Green DP, Kuklinski JH, Lupia A. 2006. The growth and development of experimental research in political science. *Am. Political Sci. Rev.* 100:627–35

Dunning T, Grossman G, Humphreys M, Hyde S, McIntosh C, Nellis G. 2018. Metaketa I: information, accountability, and cumulative learning. Evidence in Governance and Politics. **http://egap.org/ metaketa/metaketa-information-and-accountability**

Evans JSBT. 2008. Dual-processing accounts of reasoning, judgment, and social cognition. *Annu. Rev. Psychol.* 59:255–78

Evans JSBT, Stanovich KE. 2013. Dual-process theories of higher cognition: advancing the debate. *Perspect. Psychol. Sci.* 8(3):223–41

Fisher RA. 1925. *Statistical Methods for Research Workers*. Edinburgh/London: Oliver & Boyd

Fisher RA. 1935. *The Design of Experiments*. Edinburgh/London: Oliver & Boyd

Freedman DA. 1991. Statistical models and shoe leather. *Sociol. Methodol.* 18(2):291–313

Friese M, Loschelder DD, Gieseler K, Frankenbach J, Inzlicht M. 2018. Is ego depletion real? An analysis of arguments. *Personal. Soc. Psychol. Rev.* **https://doi.org/10.1177/1088868318762183**

Gaines BJ, Kuklinski JH, Quirk PJ. 2007. The logic of the survey experiment reexamined. *Political Anal.* 15:1–20

Gale WG, Iwry JM, Orszag PR, Lucas L. 2005. The automatic 401(k): a simple way. *Tax Policy Center* 401:1207–14

Garner S, Docherty M, Somner J, Sharma T, Choudhury M, et al. 2013. Reducing ineffective practice: challenges in identifying low-value health care using Cochrane systematic reviews. *J. Health Services Res. Policy* 18:6–12

Gelman A. 2018. The failure of null hypothesis significance testing when studying incremental changes, and what to do about it. *Personal. Soc. Psychol. Bull.* 44:16–23

Gerber A, Green DP. 2017. Field experiments on voter mobilization: an overview of a burgeoning literature. In *Handbook of Economic Field Experiments*, Vol. 1, ed. AS Gerber, DP Green, pp. 395–438. Amsterdam: North Holland

Gerber AS, Green DP. 2000. The effects of canvassing, telephone calls, and direct mail on voter turnout: a field experiment. *Am. Political Sci. Rev.* 94(3):653–63

Gerber AS, Green DP. 2012. *Field Experiments: Design, Analysis, and Interpretation*. New York: W.W. Norton

Giacomini M. 2009. Theory-based medicine and the role of evidence: why the emperor needs new clothes, again. *Perspect. Biol. Med.* 52:234–51

Gigerenzer G. 2015. On the supposed evidence for libertarian paternalism. *Rev. Philos. Psychol.* 6:361–83

Gigerenzer G, Gaissmaier W. 2011. Heuristic decision making. *Annu. Rev. Psychol.* 62:451–82

Gowdy JM. 2008. Behavioral economics and climate change policy. *J. Econ. Behav. Organ.* 68:632–44

Gray K. 2017. How to map theory: reliable methods are fruitless without rigorous theory. *Perspect. Psychol. Sci.* 12:731–41

Guba EG, Lincoln YS. 1994. Competing paradigms in qualitative research. *Handb. Qual. Res.* 2:105

Gueron JM, Rolston H. 2013. *Fighting for Reliable Evidence*. New York: Russell Sage Found.

Hagger MS, Wood C, Stiff C, Chatzisarantis NLD. 2010. Ego depletion and the strength model of self-control: a meta-analysis. *Psychol. Bull.* 136:495–525

Harris DN, Farmer-Hinton R, Kim D, Diamond J, Reavis TB, et al. 2018. *The promise of free college (and its potential pitfalls)*. Rep., Brown Center on Education Policy at Brookings. **https://www.brookings. edu/research/the-promise-of-free-college-and-its-potential-pitfalls/**

Hausman DM, Welch B. 2010. Debate: to nudge or not to nudge. *J. Political Philos.* 18:123–36

Hemming K, Haines TP, Chilton PJ, Girling AJ, Lilford RJ. 2015. The stepped wedge cluster randomised trial: rationale, design, analysis, and reporting. *BMJ* 350:h391

Hirschman D. 2016. Stylized facts in the social sciences. *Sociol. Sci.* 3:604–26

Hu F, Rosenberger WF. 2006. *The Theory of Response-Adaptive Randomization in Clinical Trials*, Vol. 525. Hoboken, NJ: Wiley

Humphreys M, de la Sierra RS, van der Windt P. 2013. Fishing, commitment, and communication: a proposal for comprehensive nonbinding research registration. *Political Anal.* 21(1):1–20

Imai K, Ratkovic M. 2013. Estimating treatment effect heterogeneity in randomized program evaluation. *Ann. Appl. Stat.* 7:443–70

Iyengar S. 2011. Laboratory experiments in political science. In *Cambridge Handbook of Experimental Political Science*, ed. JN Druckman, DP Green, JH Kuklinski, A Lupia, pp. 73–88. Cambridge, UK: Cambridge Univ. Press

Johnson EJ, Goldstein D. 2003. Do defaults save lives? *Science* 302:1338–39

Johnson EJ, Goldstein DG. 2013. Decisions by default. In *The Behavioral Foundations of Public Policy*, ed. E Shafir, pp. 417–27. Princeton, NJ: Princeton Univ. Press

Kahneman D. 2011. *Thinking, Fast and Slow*. New York: Farrar, Straus & Giroux

Kahneman D, Knetsch JL, Thaler RH. 1990. Experimental tests of the endowment effect and the Coase Theorem. *J. Political Econ.* 98:1325–48

Kahneman D, Tversky A. 1979. Prospect theory: an analysis of decision under risk. *Econometrica* 47(2):263–291

Keren G. 2013. A tale of two systems. *Perspect. Psychol. Sci.* 8:257–62

Kinder D, Palfrey T. 1993. On behalf of an experimental political science. In *Experimental Foundations of Political Science*, pp. 1–39. Ann Arbor: Univ. Mich. Press

Kingdon JW. 1984. *Agendas, Alternatives, and Public Policies*. New York: Longman. 2nd ed.

Kuklinski JH, Quirk PJ. 2000. Reconsidering the rational public: cognition, heuristics, and mass opinion. In *Elements of Reason: Cognition, Choice, and the Bounds of Rationality*, ed. A Lupia, MD McCubbins, SL Popkin, pp. 153–82. Cambridge, UK: Cambridge Univ. Press

Kuleshov V, Precup D. 2014. Algorithms for multi-armed bandit problems. arXiv:1402.6028

Lin W, Green DP. 2016. Standard operating procedures: a safety net for pre-analysis plans. *PS Political Sci. Politics* 49(3):495–500

Lodge M, Taber CS. 2013. *The Rationalizing Voter*. Cambridge, UK: Cambridge Univ. Press

Löfgren Å, Martinsson P, Hennlock M, Sterner T. 2012. Are experienced people affected by a pre-set default option—results from a field experiment. *J. Environ. Econ. Manag.* 63:66–72

Madrian BC. 2014. Applying insights from behavioral economics to policy design. *Annu. Rev. Econ.* 6:663–88

Madrian BC, Shea DF. 2001. The power of suggestion: inertia in 401(k) participation and savings behavior. *Q. J. Econ.* 116:1149–87

Manning WG, Newhouse JP, Duan N, Keeler EB, Leibowitz A. 1987. Health insurance and the demand for medical care: evidence from a randomized experiment. *Am. Econ. Rev.* 77:251–77

McKenzie CR, Liersch MJ, Finkelstein SR. 2006. Recommendations implicit in policy defaults. *Psychol. Sci.* 17:414–20

McKiernan EC, Bourne PE, Brown CT, Buck S, Kenall A, et al. 2016. How open science helps researchers succeed. *eLife* 5:e16800

Mettler S, Soss J. 2004. The consequences of public policy for democratic citizenship: bridging policy studies and mass politics. *Perspect. Politics* 2:55–73

Morrissey M. 2016. *The State of American Retirement*. Washington, DC: Econ. Policy Inst.

Morton RB, Williams KC. 2008. Experimentation in political science. In *The Oxford Handbook of Political Methodology*, ed. JM Box-Steffensmeier, HE Brady, D Collier, pp. 339–56. Oxford, UK: Oxford Univ. Press

Morton RB, Williams KC. 2010. *Experimental Political Science and the Study of Causality: From Nature to the Lab*. New York: Cambridge Univ. Press

Munnell AH, Webb A, Golub-Sass F. 2012. *The national retirement risk index: an update*. Rep., Center for Retirement Research, Boston College. **http://crr.bc.edu/wp-content/uploads/2012/11/IB_12-20-508.pdf**

Murphy SA. 2003. Optimal dynamic treatment regimes. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)* 65:331–55

Newhouse JP, Ins. Exp. Group. 1993. *Free for All? Lessons from the RAND Health Insurance Experiment*. Cambridge, MA: Harvard Univ. Press

Neyman JS. 1990 (1923). On the application of probability theory to agricultural experiments. essay on principles, Section 9. *Stat. Sci*. 5(4):465–80. Transl. and ed. DM Dabrowska, TP Speed, from Polish

Nguyen T, Kramer J, Evans B. 2018. *The effects of grant aid on student persistence and degree attainment: a systematic review and meta-analysis of the causal evidence*. Work. Pap. 18–04, Stanford Cent. Educ. Policy Anal., Stanford, CA

OECD. 2017. *Behavioural Insights and Public Policy*. Paris: OECD

Osman M. 2004. An evaluation of dual-process theories of reasoning. *Psychon. Bull. Rev*. 11:988–1010

Peirce CS, Jastrow J. 1885. On small differences in sensation. *Memoirs Natl. Acad. Sci*. 3:73–83

Pichert D, Katsikopoulos KV. 2008. Green defaults: information presentation and pro-environmental behaviour. *J. Environ. Psychol*. 28:63–73

Prell M. 2013. *Participation in the Supplemental Nutrition Assistance Program (SNAP) and unemployment insurance: How tight are the strands of the recessionary safety net?* USDA-ERS Econ. Res. Rep. 157. **https://permanent.access.gpo.gov/gpo47447/err157.pdf**

Pronin E, Olivola CY, Kennedy KA. 2008. Doing unto future selves as you would do unto others: psychological distance and decision making. *Personal. Soc. Psychol. Bull*. 34:224–36

Raudenbush SW. 1997. Statistical analysis and optimal design for cluster randomized trials. *Psychol. Methods* 2:173

Reiss J, Sprenger J. 2014. Scientific objectivity. In *The Stanford Encyclopedia of Philosophy (Winter 2017 Edition)*. **https://plato.stanford.edu/archives/win2017/entries/scientific-objectivity/**

Richburg-Hayes L, Anzelone C, Dechausay N, Landers P. 2017. *Nudging change in human services: final report on the Behavioral Interventions to Advance Self-Sufficiency (BIAS) Project*. Tech. Rep. May, Off. Plan. Res. Eval., US Dep. Health Hum. Serv., Washington, DC

Roberts L, Ahmed I, Davison A. 2009. Intercessory prayer for the alleviation of ill health. *Cochrane Database Syst. Rev*. 1:CD000368

Sackett DL. 1997. Evidence-based medicine. *Sem. Perinatol*. 21(1):3–5

Sackett DL, Rosenberg WM, Gray JM, Haynes RB, Richardson WS. 1996. Evidence based medicine: What it is and what it isn't. *BMJ* 312(7023):71

Samuelson W, Zeckhauser R. 1988. Status quo bias in decision making. *J. Risk Uncertainty* 1:7–59

Shafir E. 2013. *The Behavioral Foundations of Public Policy*. Princeton, NJ: Princeton Univ. Press

Skocpol T. 1995. *Protecting Soldiers and Mothers: The Political Origins of Social Policy in United States*. Cambridge, MA: Harvard Univ. Press

Stanovich KE, West RF. 2000. Individual differences in reasoning: implications for the rationality debate? *Behav. Brain Sci*. 23:645–65

Sudman S, Bradburn NM, Schwarz N. 1996. *Thinking About Answers: The Application of Cognitive Processes to Survey Methodology*. San Francisco: Jossey-Bass

Sunstein CR. 2015. Nudges, agency, and abstraction: a reply to critics. *Rev. Philos. Psychol*. 6:511–29

Sunstein CR, Reisch LA. 2014. Automatically green: behavioral economics and environmental protection. *Harvard Environ. Law Rev*. 38:127–58

Sunstein CR, Thaler RH. 2003. Libertarian paternalism is not an oxymoron. *Univ. Chicago Law Rev*. 70:1159

Sweller J. 1994. Cognitive load theory, learning difficulty, and instructional design. *Learn. Instr*. 4:295–312

Thaler RH. 2016. Behavioral economics: past, present, and future. *Am. Econ. Rev*. 106:1577–600

Thaler RH, Benartzi S. 2004. Save More Tomorrow: using behavioral economics to increase employee saving. *J. Political Econ*. 112:S164–87

Thaler RH, Ganser L. 2015. *Misbehaving: The Making of Behavioral Economics*. New York: W.W. Norton

Thaler RH, Sunstein CR. 2003. Libertarian paternalism. *Am. Econ. Rev*. 93:175–79

Thaler RH, Sunstein CR. 2008. *Nudge: Improving Decisions About Health, Wealth, and Happiness*. New Haven, CT: Yale Univ. Press

Tourangeau R, Rips LJ, Rasinski K. 2000. *The Psychology of Survey Response*. Cambridge, UK: Cambridge Univ. Press

Tversky A, Kahneman D. 1991. Loss aversion in riskless choice: a reference-dependent model. *Q. J. Econ.* 106:1039–61

Wager S, Athey S. 2017. Estimation and inference of heterogeneous treatment effects using random forests. *J. Am. Stat. Assoc.* 113(523):1228–42

Watts DJ. 2017. Should social science be more solution-oriented? *Nat. Hum. Behav.* 1:0015

Weiss CH. 1997. How can theory-based evaluation make greater headway? *Eval. Rev.* 21:501–24

Weiss CH, Whiting B, Murphy-Graham E, Petrosino A, Allison W, Gandhi G. 2008. The fairy godmother— and her warts making the dream of evidence-based policy come true. *Am. J. Eval.* 29:29–47

Contents

Errata

An online log of corrections to *Annual Review of Political Science* articles may be found
at http://www.annualreviews.org/errata/polisci