Better government, Better Science: The Promise of and Challenges facing the Evidence-Informed Policy Movement

Jake Bowers[1] and Paul Testa[2]

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

[2]Department of Political Science, Brown University;
email:paul_testa@brown.edu

Xxxx. Xxx. Xxx. Xxx. YYYY. AA:1–27

https://doi.org/10.1146/((please add article doi))

Keywords

evidence-based policy, evidence-informed policy, behavioral insights, randomized field experiments

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. We show that the evidence-based policy movement consists of two main threads: (1) an effort to invent new policies using insights from the social and behavioral science consensus about human behavior and institutions and (2) 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 criticisms facing one or the other 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-based public policy practice that (1) involves cross-sector collaborations using the latest theory plus deep contextual knowledge to *design* new policies, (2) the latest insights in research design and statistical inference for causal questions, and (3) is focused on assessing explanations as much as on discovering what works. We see the evidence-based public policy movement as 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.

Contents

1. Evidence as opposed to what?

The appeal of an evidence-informed public policy is self-evident, 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 policymaking process and the politics that surround it (Kingdon 1984; Baumgartner and Jones 1991; Cairney 2016). Increasingly, however, policy makers are inviting political scientists and other social and behavioral scientists to participate in policymaking 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 amounts of the 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 policymaking, but may be less aware of the growing opportunities to participate in this process. We think that , 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 in turn helps us understand and address some of the most important criticisms of evidence-informed policymaking. We argue many of the of 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 and desire among policymakers that an understanding of human behavior can help improve policy outcomes, has opened the door for social scientist 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 we argue is crucial to realizing the full promise of evidence-informed policymaking for government, science, and citizens.

We begin by discussing the dual functions of evidence in policymaking. We review the the evolution of evidence as evaluation for policymaking, highlighting the central role randomized field experiments have played in this process. Next we turn to more recent example applications of behavioral insights to policy policy. Our review is by no means exhaustive[1]. Instead we use this discussion to help re-frame some common critiques of evidence-informed policymaking 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" — credible research designs, transparent and open evaluation — and argue that these collaborations are often well suited to address larger questions of mechanism, context, and generalization. In doing so we hope to convince scholars of the tremendous potential such collaborations hold both for addressing questions within our discipline and the problems that face our society.

2. The meaning of evidence in public policy

What does "evidence" mean? What does it mean to "base" a policy on evidence? Or create a policy "informed by" evidence? When academics collaborate with government experts to solve specific policy problems, "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 shows that social comparisons can reduce opioid prescribing among doctors.').[2]

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 — in ideal form 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

[1] See Shafir (2013) for an extensive review of recent applications within the U.S. and OECD (2017) for a summary of applications around the world.

[2] Systematic measurement and observation also provides "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 loose their benefits each year even when their income does not change (Prell 2013) many policy makers would ask both why this happens and how such "churn" might be prevented.

to avoid any personal or systemic bias. The evidence-as-evaluation approach also 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.[3]

This idea that policy should be created using knowledge that is collective as opposed to individualistic, and objective as opposed to subjective, is not new. The closest immediate ancestor of the evidence-informed policy (EIP) movement is the evidence-based medicine movement. To combat a growth in medical costs and to reduce medical 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 (Bluhm and Borgerson 2011; Djulbegovic and Guyatt 2017; Giacomini 2009; Sackett et al. 1996; Sackett 1997). As envisioned then, better health outcomes would result from doctors following guidelines derived from dispassionate syntheses of the results of pre-registered randomized controlled trials that had gone through blind peer review than from doctors following the evidence of their own idiosyncratic experience to guide clinical decisions. In this way, 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.

Proponents of an evidence-informed policy-making process, a decade or two behind evidence-based medicine in its growth, tend to emphasize *both* the idea 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 Evidence-Based Policy Commission, formed by the Evidence-Based Policymaking Commission Act of 2016 emphasizes the idea of evidence-creation: "The Commission on Evidence-Based Policy-making (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." (Evidence-Based Policymaking 2017, p. 1) 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 both 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 ideas[4]. In practice, however, many debates around evidence-informed policymaking sometimes obscure, conflate, or ignore these two roles, and so it is useful to consider the roles of evidence for evaluation and insight first separately to see how they can be productively linked.[5]

[3]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 themselves 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 and Sprenger (2014) on the epistemic authority of science and idea of scientific objectivity from the point of view of the philosophy of science.

[4]See Office of Evaluation Sciences, "About", accessed online at `https://oes.gsa.gov/about/`

[5]In this article we refer to "evidence-informed" rather than the more popular "evidence-based" public policy movement because no scientific consensus alone has been enough to dictate a public policy. Instead, the scientific

3. 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 like Abt and RAND worked with the US federal government to field large-scale randomized controlled trials (RCTs) (Manning et al. 1987; Newhouse, Group, Staff, et al. 1993; Gueron and 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 set the second kind of debate to rest: 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, Kuklinski, and Quirk 2007) and lab experiments (Iyengar 2011; Morton and Williams 2010). Field experiments (Gerber and Green 2012) soon followed in the late 1990s.[6] Randomized field experiments in political science involved collaborations between 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 non-academic partner.[7] Governments began to collaborate with political scientists on such projects, as well. For example, Bhatti et al. (2015) and Bhatti et al. (2017) present voter turnout experiments done in direct collaboration with the Danish government.[8] Groups within government like the the Behavioral Insights Team (BIT) in the UK, the OES in the U.S., and The Lab@DC in the Mayor's Office in the District of Columbia soon joined the big research consulting firms (like Abt, RAND, MDRC, Mathematica Policy Research), academic-practitioner collaboration oriented research NGOS (like J-PAL, EGAP, and ideas42) and private firms (like 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, in 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 (1923) each built a version of statistical inference on the basis of random assignment. Fisher (1935, Chap 2) famously showed that randomization could be a "reasoned basis" for statistical inference about causal claims, although the use

consensus and academics themselves tend to play a role in collaboratively creating new public policies — the "evidence-base" and its interpreters, the academics, "inform" rather than dictate.

[6] 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 and Green (2017) for a short history.

[7] See, for example, the early field experiments focusing on voter turnout in collaboration with civic groups such as (Gerber and Green 2000; Morton and Williams 2008)

[8] See the website of EGAP `http://egap.org` for many more randomized field experiments, mostly focusing on topics in developing countries, designed and fielded in collaboration with NGOs.

of randomization to make fair comparisons goes back further, perhaps to the psycho-physical experiments of Peirce and Jastrow (1885). The idea that an RCT provides clarity of comparison is what Kinder and 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 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 could 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, Chap 2) in which a statistical hypothesis test is introduced using eight cups of tea and in which the only assumption to be justified is that the cups of tea were presented in a random order.[9] This clarity of comparison and method has enabled discussion about "what works" to focus on the substance. If a large RCT showed that policy X was better than policy Y, then policymakers 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 clearly 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 and Hardie 2012).

4. Evidence as insights: 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. Students of human decision making (mostly from psychology and economics) began to influence policy on the creation side even as students of randomized experiments and causal inference worked on the evaluation side. The early work on decision making within psychology such as Kahneman and Tversky (1979) helped give rise to the field of behavioral economics.[10] Together, this research helped launch a movement to use insights from social and

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

[10] See Thaler and Ganser (2015) and Thaler (2016) and CASBS (2018) for more on the history of behavioral economics.

behavioral sciences to improve policy.[11]

The popular book, *Nudge*, by Thaler and Sunstein (2008) further inspired this effort in the policy world. The pioneering UK Behavioral Insights Team (BIT), 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. And the White House Social and Behavioral Sciences Team and Office of Evaluation Sciences (OES) founded in 2014 within the the U.S. General Services Administration and the Executive Order instructing the federal agencies to attend to behavioral science as a part of the policymaking process implemented a combined approach — testing nearly every one of its policy creations with an RCT (Congdon and Shankar 2018; Benartzi et al. 2017; Congdon and Shankar 2015).

The "behavioral insights" approach to evidence-based policymaking 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 insights 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, or any other government forms and processes, then the policy 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 help illustrate this approach in practice, we discuss the default example of how behavioral insights improve public policy: specifically, the role of defaults in retirement savings.

4.1. 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 savings for retirement using the tax code. Yet, even with tax incentives and matching contributions from employers, many individuals eligible for programs like 401(k)s and IRAs do not save enough and/or do not save at all (Munnell, Webb, Golub-Sass, et al. 2012). Further, even those who do use such programs, do not save at the rate that a rational actor would save and so do not save enough for a comfortable retirement without the need for extra assistance (Benartzi and Thaler 2013).

Human beings often do not behave the way that rational actors would. Thaler and Ganser (2015) explain how economics turned to psychology as the rational actor based psychological micro-foundations of earlier economics failed to explain a growing number of economically relevant outcomes including retirement saving. Today the list of "cognitive biases" by which actual human behavior diverged from the predictions of standard rational actor models is quite long: for only one example, Bettinger et al. (2012) found that,

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

although saving thousands of dollars on college tuition make the material costs of a four-hour effort to fill out a form worthwhile for any rational actor, an easier form-filling out process caused more young people to take advantage of federal college loans.

The idea of a "default option" arose from the psychology and economics that sought to develop theoretical understandings of seemingly anamolous behavior and provide practical advice how policies should be designed in 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 policymaker could set automatic paycheck deductions for retirement savings at five percent in the hopes that rational actors would switch away from the default if they thought it wasn't optimal for them, and that regular humans would find lack of action easier and thus achieve their own long term goals 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 and 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 given the option 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 two to four percent (Choi et al. 2004; Madrian 2014). Thaler and 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 told they need to be saving more for retirement, and about 25 percent chose to increase their contributions to the recommended five percentage points after meeting with the consultant. Individuals who said they couldn't afford to increase their contribution were offered the chance to enroll in a plan that tied increased savings rates to future pay raises. Three and half years later, participants in the "Save More Tomorrow" plan, had an average contribution rate of about 13.6 percent — 4.6 percentage points higher than those who had increased their savings rate after the initial consultation but without the appeal to tying subsequent contributions to future raises. Thus, in addition to automatic enrollment, the principle of automatic escalation of contributions (a form of process default) increasingly common feature of savings plans offered by U.S. employers (Benartzi and Thaler 2013)

Why are defaults often the default example of the way behavioral insights can work in public policy? One reason is that they work, in a wide array of settings. Comparing rates of organ donation Johnson and Goldstein (2003), find the lowest effective consent rate among countries with opt-out systems is 85.9 percent, nearly 60 percentage points higher than the highest consent rate among countries requiring explicit consent (27.5 percent 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 and Katsikopoulos 2008; Sunstein and Reisch 2014).

Second, compared to other policy tools such as incentives, sanctions, and mandates, defaults are a

relatively "low-touch" intervention—what Thaler and Sunstein call a "nudge" or "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." (2008, p. 6). A nudge is consistent with the principles of what they call "libertarian paternalism" when it is "choice set preserving" (i.e. doesn't change the possibilities for a person's action), cheap and easy to avoid or opt out of, and leads to outcomes that individuals themselves would prefer (Thaler and Sunstein 2003).[12] 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, in this way and not that way — policymakers find it easy to justify behaviorally informed approaches, especially if the program designers preserve the freedom of action and autonomy of the public.

Third, defaults are an appealing example of behavioral insights in policy because 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 (Johnson and Goldstein 2013; Choi et al. 2003). Yet default effects are also found in experimental settings where such costs are absent (Samuelson and 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 and Kahneman 1991), endowment effects (Kahneman, Knetsch, and Thaler 1990), and time inconsistent preferences (Kahneman and Tversky 1979; Pronin, Olivola, and Kennedy 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, Liersch, and Finkelstein 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 defaults had little effect on the decision to use carbon-offsets for those attending an environmental conference.

The default example, then, is an example that 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 itself.

4.2. Broader applications of behavioral insights

Defaults are just one example of a broader set of concepts, principles and tools employed to conduct behaviorally informed policymaking. Many of the underlying insights should be familiar to political scientists, as concepts like framing, heuristics, cues, bounded rationality, social norms and peer influence are commonly

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

used to explain aspects of political behavior and politics more broadly (e.g Chong and Druckman 2007; Kuklinski, Quirk, et al. 2000; Lodge and Taber 2013). Others — such as cognitive load (Sweller 1994) or 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.[13]

Practitioners are typically less concerned with specific models of cognition and more focused on the practical implications of behavioral theory for policy design. People working to use evidence as insights have often invented catchy mnemonic acronyms to help 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 (BIT) of the United Kingdom's Cabinet Office to provide a guide for policymakers of common factors known to influence behavior (Dolan et al. 2010). Similarly, the Behavioral Interventions to Advance Self-Sufficiency (BIAS) project — a program focused on using behavioral insights to improve outcomes for low income children, adults, and families, sponsored by Office of Planning, Research and Evaluation in the U.S. Department of Health and Human Services with the contractor MDRC — developed the "SIMPLER" acronym to summarize various behavioral insights applied across 15 evaluations (Richburg-Hayes et al. 2017, 2017). SIMPLER stands for Social influence, Implementation prompts, Making deadlines, Personalization, Loss aversion, Ease, and Reminders. Finally, taking some of their own advice to heart, in 2014, BIT presented the EAST framework suggesting that policymakers focus on policies which 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 and necessary. But is also potentially problematic. The evidence base that scholars bring to issues is more contested and evolving than these prescriptive principles may suggest. On the spectrum between unfounded belief and scientific law, most behavioral insights fall somewhere in between. Even in cases like the default example, where the evidence base is large and well-established, there remain outstanding questions about the mechanism that produce changes in behavior. Thus the process of evidence informed policymaking requires more than just the 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 used not just to answer simple questions — did an intervention have an effect — but also deeper questions — did an intervention work the way theory suggests it should. Doing so can help address many of the common objections raised about evidence-informed policymaking to which we now turn.

[13]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) (Stanovich and West 2000; Kahneman 2011). Evans (2008) and Evans and Stanovich (2013) provide useful reviews in psychology, Lodge and Taber (2013) demonstrate applications to political science, and Brocas and Carrillo (2014) do so in economics. For some critiques of dual-process models see Osman (2004), Keren (2013), and Gigerenzer and Gaissmaier (2011).

5. Roadblocks on the way to evidence informed policy

If every medium sized city and county, every U.S. state, and OECD nation had a small behavioral insights team practicing evidence-as-insights, or a small field experimentation team practicing evidence-as-evaluation, or even a team like the OES which combines the two approaches, would we see a radically improved government, a social science generating new theories and methods to grapple with new questions, and a public increasingly satisfied with the role 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 the 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 problems: problems of principle, theory, and practice, each of which requires us to expand upon the links between insights and evaluation in the practice of evidence-informed policymaking. For each, we argue that many of the common objections arise from a conception of evidence-informed policymaking as solely focused on evaluation or insight and suggest that concerns can be addressed by creating a stronger link between the two.

5.1. Problems of principle and politics

Problems with the principle of evidence informed policymaking are often couched in terms of concerns about paternalism. Critiques of paternalism can be either general or particular. General critiques argue that EIP 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 policymaker engaged in EIP is not so much an honest broker, dispassionately evaluating just the facts, but a motivated salesperson, producing "policy-based evidence" guaranteed to support some preordained goal.

Proponents of EIP have both normative and practical responses to these concerns. First, they can highlight the extent to many behaviorally motivated interventions in EIP are consistent with the principles of libertarian paternalism outlined by Thaler and 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 and EIP in general would note that: (1) 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); (2) there are other conceptions of a good government beyond the passive preservation of liberties, such as the solving of collective goods problems or the guarding of justice

and equality; and (3) 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 doesn't work leading to the retooling of ineffective programs (Garner et al. 2013, e.g.). Further, practitioners would also argue that the process by which evidence is generated can help insulate policymaking from charges of political bias.[14] That is 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 and Green 2016), pre-registering designs (Humphreys, Sierra, and Windt 2013) and providing access to data and replication materials (McKiernan et al. 2016) — which increase the integrity of policymaking process[15] 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 pre-register, review and replicate each other's work. Few academics commit to publicizing every study they begin, and most do not perceive of their work as inherently products 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.

5.2. 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 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 (Mettler and Soss 2004; Campbell 2003; Skocpol 1995) Both civic and governmental efficiency gains can arise from improving forms — and there are so many forms, and so much time spent on forms, 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. 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 questions.[16] Academics can also appeal to common sense and the

[14]Highlighting the role of evidence for evaluation can also address some potential ethical concerns of behavior interventions The act of evaluation provides an opportunity for the public and other stakeholders to assess not just whether the outcomes of policy are desirable (e.g. more people saving 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.

[15]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 `https://oes.gsa.gov/methods/`)

[16]See for example, The Lab @ DC's "Form-a-Palooza", `https://osf.io/kf4r9/`

basic science of communication: plain language and simple graphic design ought do a better job guiding the public and eliciting accurate information than legal language written in small fonts. However, those researchers who have confronted these problems know that form reform is a place where past literature and theory provide a basis for reasoned improvisation but not the opportunity to directly translate some approach that worked in the lab to 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 both policymakers' immediate questions about what works while also helping 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 policymakers, 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 and 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 itself. For example, a synthesis of the research on survey response (e.g. Tourangeau, Rips, and Rasinski 2000; Sudman, Bradburn, and Schwarz 1996) 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.[17]

5.3. Problems of practice and learning

A final set of concerns arise from the general problems of learning from observation. Cartwright and Hardie (2012) and Deaton and Cartwright (2017) warn us that the results of one study of one policy, in one place, and one moment, may not teach us directly about that same policy as applied in another place and another time. 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 policymaking. 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. Yet, there are many policy environments in which randomization is difficult, especially if the desire is to use an RCT to learn about "what works" rather than "why" or "how". Further, although RCTs can and do help social scientists answer "why" questions every day, if evidence equals an RCT, and RCTs are seen as only answering "what works" questions, then policymakers will be ill equipped to respond to changes in context like the rise of the "gig" economy (De Stefano 2015) or changing

[17] See for example the somewhat surprising results of college scholarships on degree completion detailed in Angrist, Hudson, Pallais, et al. (2016).

climates (Gowdy 2008) and social scientists will struggle to use the collaborations to advance science itself. Giacomini (2009) 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 [Evidence Based Medicine]EBM's maturity and considerable influence, one form of authority — expert opinion — has been replaced in many minds with another — evidence from well-designed RCTs (p. 236).

To dramatize this problem Giacomini considers the field of randomized studies of the health effects of remote prayer (people praying 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, Ahmed, and Davison 2009) most of which yield no evidence for 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" (p. 244). More broadly, Giacomini cautions that "experimental evidence about unexplainable interventions may be not only pragmatically worthless, but even misleading or harmful" (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 and 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 under-prepared teacher may produce worse outcomes than a large class size with an expert teacher. According to Cartwright and Hardie, the Tennessee results inspired the state of California to rush their hiring of new teachers for the class size policy change. This decreased class sizes but with much less well prepared teachers. Cartwright and Hardie argue 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 to operate.

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, et al. 1994). And 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 strategy to defuse political arguments. But we think there are good reasons why the practice of evidence informed policymaking 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 2018) 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 itself about next actions rather than only an academic interested in theory assessment. 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 multiple modes of expertise in the task of description and interpretation. Further 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, such as those common in organizations like the OES, the context and population of the study is the context and population of the policy itself.

Second, public pre-registration 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 knowledge from such studies are increasingly being generated from 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 often call "theories of change" (Weiss 1997; Coryn et al. 2011). This focus on "why" avoids the misunderstanding the 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) then 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 events local to the place, or an influx of new comers because of climate change elsewhere — the government can more easily predict and prepare for the changing functioning of the policy. This last approach also promises the most benefits for a theory-driven academia itself. The more evidence-based policy collaborations focus on "why might this new approach work better than this other approach", the quicker the translation of the new research into the academic consensus itself and the more agile government itself will be in the face of change.

5.4. Lessons from Behavioral Insights for Evidence-Informed Policymaking

Many objections to evidence-informed policymaking arise from the potential disconnects between evaluation and insights in the process of policymaking. 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. While we draw on our own experiences and the advice of others, our goal is not to provide the definitive how-to manual. Rather our aim is to highlight many of the similarities to what scholars are already doing in their own research practice and draw attention to some particular 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 policymaking 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 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 policymakers 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 given findings different from what is expected in these planning stages.

Upon agreeing that behavioral insights might be applied to a particular policy problem practitioners engage in a number of diagnostic tasks, from gathering evidence through 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.

Having generated a set of potentially promising, behaviorally informed interventions, it is generally the case that the same set of practices and designs that produce credible academic research 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 designs to address logistic and political constraints. For example, a researcher may need to articulate the appeal of a stepped wedge (Brown and Lilford 2006; Hemming et al. 2015) or adaptive designs (Hu and Rosenberger 2006) to policymakers concerned about the ethics of randomly assigning access to a program and be able to clarify the limitations and challenges that arise when randomizing over clusters rather than individuals (Raudenbush 1997). Perhaps the most important goal at this stage of the process is for stakeholders and policymakers 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 pre-register the design of the program evaluation (Humphreys, Sierra, and Windt 2013). While the benefits of pre-registration are increasingly clear to academics, policymakers can be convinced of the benefits of this process as way to enhance both the scientific and political integrity of the results.

The final stages of this process are quite similar to what social scientists are likely to encounter in their own involving project management and analysis. Issues can and do arise during implementation although sometimes these problems are themselves theoretically fruitful. 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 uncertain, equivocal and only partially consistent with prior theory. Scholars and practitioners alike walk the line between neither overselling the results of promising pilot, nor completely abandoning a project that has worked elsewhere but appears ineffective in a new context. But perhaps more so than in academia, null results can hold considerable policy sway — evidence that the Drug Abuse Resistance Education (D.A.R.E.) had little to 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 convenience sample of willing participants) concerns about generalizing out of sample are more 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 particular rich discussion of this process, examining how a program called "Teaching at the Right Level" that showed promising returns for closing educational gaps in early pilots India, and learned and

adapted from both successes and failures as the program was implemented in different contexts and at greater scales across India.

Overall, the practice of evidence-informed policymaking 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.

6. A Promise of Better Science, Better Government, Better Society

This reflection on the evidence-based 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 that the challenges facing both types of evidence-based policy making can be productively engaged or even overcome by combining the two main approaches and by adding a focus on theory, or assessment of explanations, or on theories of change (and perhaps on the idea of a multi-year learning agenda that is shared between government and University), and by ensuring close ties between government and academy (such as a team anchored by full time government employees and full time academic researchers, but including academics and policy experts on year long leaves as well as academics and other researchers who work on particular projects).

A focus on theory offers three distinct benefits for science, government, and society. First, a greater emphasis on the "why's" 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 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 and/or many interventions and/or treatment arms. The social sciences has 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 (often in the tens to hundreds of thousands of subjects) and the connection to administrative datasets 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 and Precup 2014) — with applications from the field of machine learning to let the data help identify heterogeneous treatment effects (Imai, Ratkovic, et al. 2013; Wager and Athey 2017).

Second, collaboration provides the opportunity to advance a number of fields of substantive interest

in political science. For scholars of bureaucracy and policy change this 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 isn't just good policy it's also good politics. But this is an open claim, in need of evaluation at both the institutional and individual level. Scholars need not participate directly in collaborations to ask whether policies informed and evaluated by evidence are more likely to overcome partisan gridlock, diminish polarization, or are more likely to 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 it make a difference in how they interpret controversial findings that may directly impact their daily lives? Can a commitment to rigor, transparency, and impartial evaluation 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 of this? 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 either 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 itself. 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 themselves.

A greater focus on theory is not just a self-interested play to create more opportunities for academics to publish. Rather an evidence-informed policymaking 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 rather than having to evaluate a program in its entirety. Will providing free tuition increase

the number of college graduates and boost economic growth? While the process of evidence-informed policymaking may sometimes yield definitive answers — smart defaults can increase savings for retirement — often it does not — in general programs that decrease the costs of college increase enrollment, and to a lesser degree 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 and Dynarski 2010; Angrist, Hudson, Pallais, et al. 2016; Nguyen, Kramer, and Evans 2018; Harris et al. 2018). An evidence-informed approach to policymaking need not (and often will not) provide simple yes or no answers to the asked of it for to be useful to governments and society. By pairing multiple evaluations testing components of a well-articulated theory of change, evidence informed policymaking offers more than just a simple answer to the question of did some 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, it 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 policymakers can assess further downstream effects without having to field a completely new randomized intervention (Brady 2018).

Finally, an evidence informed policymaking focused on theory will help governments in the long run adapt and respond 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 available for why a policy is working as it should, would help a government react to the changes in the world which will depress or even augment the causal effects found during evaluation processes. When efforts to learn what works are centered around both producing evidence for evaluation and insights into mechanisms and theory, they more likely the can be adapted to address changes in context. Changes to our in climate, technology, demography and the economy pose significant challenges to our governments and society. We think that the kinds of collaborations modeled for us so far have shown the 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.

References

Angrist, Joshua, Sally Hudson, Amanda Pallais, et al. 2016. *Evaluating Post-Secondary Aid: Enrollment, Persistence, and Projected Completion Effects*. Tech. rep. National Bureau of Economic Research.

Banerjee, Abhijit, et al. 2017. "From Proof of Concept to Scalable Policies: Challenges and Solutions, with an Application". *Journal of Economic Perspectives* 31, no. 4 (): 73–102. ISSN: 0895-3309. `http://pubs.aeaweb.org/doi/10.1257/jep.31.4.73%20http://www.nber.org/papers/w22931%7B%5C%%7D0Ahttp://pubs.aeaweb.org/doi/10.1257/jep.31.4.73`.

Baumgartner, Frank R, and Bryan D Jones. 1991. "Agenda dynamics and policy subsystems". *The journal of Politics* 53 (04): 1044–1074.

Benartzi, Shlomo, and Richard H Thaler. 2013. "Behavioral economics and the retirement savings crisis". *Science* 339 (6124): 1152–1153.

Benartzi, Shlomo, et al. 2017. "Should governments invest more in nudging?" *Psychological science* 28 (8): 1041–1055.

Beshears, John, et al. 2008. "The Importance of Default Options for Retirement Saving Outcomes: Evidence from the USA". In *Lessons from Pension Reform in the Americas*, 1:271–307. Elsevier Ltd. ISBN: 9780191710285. arXiv: arXiv:1011.1669v3.

Bettinger, Eric P, et al. 2012. "The Role of Application Assistance and Information in College Decisions: Results from the H&R Block FAFSA Experiment". *The Quarterly Journal of Economics* 127 (3): 1205–1242. http://www.jstor.org/stable/23251984%20http://about.jstor.org/terms%20http://www.jstor.org/stable/23251984%7B%5C%%7D0Ahttp://about.jstor.org/terms.

Bhatti, Yosef, et al. 2015. "Getting out the vote with evaluative thinking". *American Journal of Evaluation* 36 (3): 389–400.

Bhatti, Yosef, et al. 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". *Journal of Elections, Public Opinion and Parties* 27 (3): 291–310.

Bluhm, Robyn, and Kirstin Borgerson. 2011. "Evidence-based medicine". In *Philosophy of medicine*, 203–238. Elsevier.

Brady, Henry E. 2018. "The Challenge of Big Data and Data Science". *The Annual Review of Political Science*.

Brocas, Isabelle, and Juan D. Carrillo. 2014. "Dual-process theories of decision-making: A selective survey". *Journal of Economic Psychology* 41 (): 45–54. ISSN: 0167-4870.

Brown, Celia A, and Richard J Lilford. 2006. "The stepped wedge trial design: a systematic review". *BMC medical research methodology* 6 (1): 54.

Cairney, Paul. 2016. *The politics of evidence-based policy making*. Springer.

Campbell, Andrea Louise. 2003. *How policies make citizens: Senior political activism and the American welfare state*. Princeton University Press.

Carter, Evan C, et al. 2015. "A series of meta-analytic tests of the depletion effect: Self-control does not seem to rely on a limited resource." *Journal of Experimental Psychology: General* 144, no. 4 (): 796–815. ISSN: 1939-2222.

Cartwright, Nancy, and Jeremy Hardie. 2012. *Evidence-based policy: a practical guide to doing it better*. Oxford University Press.

CASBS. 2018. "CASBS in the History of Behavioral Economics". [ONLINE: 11 JULY 2018]. https://casbs.stanford.edu/casbs-history-behavioral-economics.

Choi, James J., et al. 2003. "Optimal Defaults". *The American Economic Review* 93 (2): 180–185.

Choi et al. 2004. *For Better or for Worse: Default Effects and 401(k) Savings Behavior*, 401:81–126. June. ISBN: 0226903052.

Chong, Dennis, and James N Druckman. 2007. "Framing Theory". *Annual Review of Political Science* 10, no. 1 (): 103–126. ISSN: 1094-2939.

Congdon, William J, and Maya Shankar. 2018. "The Role of Behavioral Economics in Evidence-Based Policymaking". *The ANNALS of the American Academy of Political and Social Science* 678 (1): 81–92.

— . 2015. "The White House Social & Behavioral Sciences Team: Lessons learned from year one". *Behavioral Science & Policy* 1 (2): 77–86.

Coryn, Chris LS, et al. 2011. "A systematic review of theory-driven evaluation practice from 1990 to 2009". *American journal of Evaluation* 32 (2): 199–226.

Damelio, Robert. 2016. *The basics of process mapping*. Productivity Press.

De Stefano, Valerio. 2015. "The rise of the just-in-time workforce: On-demand work, crowdwork, and labor protection in the gig-economy". *Comp. Lab. L. & Pol'y J.* 37:471.

Deaton, Angus, and Nancy Cartwright. 2017. "Understanding and misunderstanding randomized controlled trials". *Social Science & Medicine* (). ISSN: 0277-9536.

Deming, David, and Susan Dynarski. 2010. "College aid". In *Targeting investments in children: Fighting poverty when resources are limited*, 283–302. University of Chicago Press.

Dinner, Isaac, et al. 2011. "Partitioning default effects: Why people choose not to choose". *Journal of Experimental Psychology: Applied* 17 (4): 332–341. ISSN: 1076-898X.

Djulbegovic, Benjamin, and Gordon H Guyatt. 2017. "Progress in evidence-based medicine: a quarter century on". *The Lancet* 390 (10092): 415–423.

Dolan, Paul, et al. 2010. "MINDSPACE: Influencing behavior through public policy". *Institute for Government and Cabinet Office.*

Druckman, James N, et al. 2006. "The growth and development of experimental research in political science". *American Political Science Review* 100 (4): 627–635.

Dunning, Thad, et al. 2018. *Metaketa I: Information, Accountability, and Cumulative Learning.* available online at http://egap.org/metaketa/metaketa-information-and-accountability.

Evans, J.S.B.T. 2008. "Dual-processing accounts of reasoning, judgment, and social cognition". *Annu. Rev. Psychol.* 59:255–278.

Evans, Jonathan St B.T., and Keith E. Stanovich. 2013. "Dual-Process Theories of Higher Cognition: Advancing the Debate". *Perspectives on Psychological Science.* ISSN: 1745-6916.

Evidence-Based Policymaking, Commission on. 2017. *The promise of evidence-based policymaking: Report of the Commission on Evidence-Based Policymaking.*

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

— . 1935. *The design of experiments.* Edinburgh / London: Oliver & Boyd.

Freedman, David A. 1991. "Statistical models and shoe leather". *Sociological methodology*: 291–313.

Friese, Malte, et al. 2018. "Is Ego Depletion Real? An Analysis of Arguments". *Personality and Social Psychology Review* (): 108886831876218. ISSN: 1088-8683.

Gaines, Brian J, James H Kuklinski, and Paul J Quirk. 2007. "The logic of the survey experiment reexamined". *Political Analysis* 15 (1): 1–20.

Gale, William G, et al. 2005. "The Automatic 401(k): A Simple Way". *Tax Policy Center* 401:1207–1214.

Garner, Sarah, et al. 2013. "Reducing ineffective practice: challenges in identifying low-value health care using Cochrane systematic reviews". *Journal of health services research & policy* 18 (1): 6–12.

Gelman, Andrew. 2018. "The Failure of Null Hypothesis Significance Testing When Studying Incremental Changes, and What to Do About It". *Personality and Social Psychology Bulletin* 44, no. 1 (): 16–23. ISSN: 1552-7433.

Gerber, A.S., and D.P. Green. 2017. "Field Experiments on Voter Mobilization". 1:395–438. ISSN: 1098-6596. arXiv: `arXiv:1011.1669v3`.

Gerber, Alan S, and Donald P Green. 2012. *Field experiments: Design, analysis, and interpretation.* WW Norton.

— . 2000. "The effects of canvassing, telephone calls, and direct mail on voter turnout: A field experiment". *American Political Science Review*: 653–663.

Giacomini, Mita. 2009. "Theory-based medicine and the role of evidence: why the emperor needs new clothes, again". *Perspectives in Biology and Medicine* 52 (2): 234–251.

Gigerenzer, Gerd. 2015. "On the Supposed Evidence for Libertarian Paternalism". *Review of Philosophy and Psychology* 6, no. 3 (): 361–383. ISSN: 1878-5166.

Gigerenzer, Gerd, and Wolfgang Gaissmaier. 2011. "Heuristic Decision Making". *Annual Review of Psychology* 62, no. 1 (): 451–482. ISSN: 0066-4308. arXiv: `0402594v3 [arXiv:cond-mat]`.

Gowdy, John M. 2008. "Behavioral economics and climate change policy". *Journal of Economic Behavior & Organization* 68 (3-4): 632–644.

Gray, Kurt. 2017. "How to Map Theory: Reliable Methods Are Fruitless Without Rigorous Theory". *Perspectives on Psychological Science* 12, no. 5 (): 731–741. ISSN: 1745-6916.

Guba, Egon G, Yvonna S Lincoln, et al. 1994. "Competing paradigms in qualitative research". *Handbook of qualitative research* 2 (163-194): 105.

Gueron, Judith M., and Howard. Rolston. 2013. *Fighting for Reliable Evidence*. New York: Russell Sage Foundation. ISBN: 978-1-61044-813-0.

Hagger, Martin S., et al. 2010. "Ego depletion and the strength model of self-control: A meta-analysis." *Psychological Bulletin* 136 (4): 495–525. ISSN: 1939-1455.

Harris, Douglas N, et al. 2018. "The promise of free college (and its potential pitfalls)". *Brown Center on Education Policy at Brookings*.

Hausman, Daniel M, and Brynn Welch. 2010. "Debate: To nudge or not to nudge". *Journal of Political Philosophy* 18 (1): 123–136. ISSN: 0963-8016.

Hemming, Karla, et al. 2015. "The stepped wedge cluster randomised trial: rationale, design, analysis, and reporting". *Bmj* 350:h391.

Hirschman, Daniel. 2016. "Stylized Facts in the Social Sciences". *Sociological Science* 3:604–626.

Hu, Feifang, and William F Rosenberger. 2006. *The theory of response-adaptive randomization in clinical trials*. Vol. 525. John Wiley & Sons.

Humphreys, Macartan, Raul Sanchez de la Sierra, and Peter van der Windt. 2013. *Fishing, Commitment, and Communication: A Proposal for Comprehensive Nonbinding Research Registration*.

Imai, Kosuke, Marc Ratkovic, et al. 2013. "Estimating treatment effect heterogeneity in randomized program evaluation". *The Annals of Applied Statistics* 7 (1): 443–470.

Iyengar, Shanto. 2011. "Laboratory experiments in political science". *Handbook of experimental political science*: 73–88.

Johnson, Eric J., and Daniel Goldstein. 2003. *Do Defaults Save Lives?* Boca Raton. arXiv: `arXiv:1011.1669v3`.

Johnson, Eric J, and Daniel G Goldstein. 2013. "Decisions by default." *The behavioral foundations of public policy*. 417–427. ISSN: 978-0-691-13756-8 (Hardcover); 978-1-400-84534-7 (PDF).

Kahneman, Daniel. 2011. *Thinking, Fast and Slow*, 93:54–57. 202249. New York: Farrar, Straus / Giroux. ISBN: 9780385676519. arXiv: `arXiv:1011.1669v3`.

Kahneman, Daniel, Jack L. Knetsch, and Richard H. Thaler. 1990. "Experimental Tests of the Endowment Effect and the Coase Theorem". *Journal of Political Economy* 98, no. 6 (): 1325–1348. ISSN: 0022-3808.

Kahneman, Daniel, and Amos Tversky. 1979. "Prospect Theory: An Analysis of Decision under Risk". *Econometrica* 47, no. 2 (): 263. ISSN: 0012-9682.

Keren, Gideon. 2013. "A Tale of Two Systems". *Perspectives on Psychological Science* 8, no. 3 (): 257–262. ISSN: 1745-6916.

Kinder, D.R., and T.R. Palfrey. 1993. "On behalf of an experimental political science". *Experimental foundations of political science*: 1–39.

Kingdon, John W. 1984. *Agendas, Alternatives, and Public Policies, 2^{nd} Edition (Longman Classics in Political Science)*. New York: Longman. ISBN: 0321121856.

Kuklinski, J H, P J Quirk, et al. 2000. "Reconsidering the rational public: Cognition, heuristics, and mass opinion". *Elements of reason: Cognition, choice, and the bounds of rationality*: 153–182.

Kuleshov, Volodymyr, and Doina Precup. 2014. "Algorithms for multi-armed bandit problems". *arXiv preprint arXiv:1402.6028*.

Lin, Winston, and Donald P. Green. 2016. "Standard Operating Procedures: A Safety Net for Pre-Analysis Plans". *PS - Political Science and Politics*. ISSN: 1537-5935.

Lodge, Milton, and Charles S Taber. 2013. *The rationalizing voter*. Cambridge University Press.

Löfgren, Åsa, et al. 2012. "Are experienced people affected by a pre-set default option—Results from a field experiment". *Journal of Environmental Economics and Management* 63, no. 1 (): 66–72. ISSN: 0095-0696.

Madrian, Brigitte C. 2014. "Applying Insights from Behavioral Economics to Policy Design". *Annual Review of Economics* 6, no. 1 (): 663–688. ISSN: 1941-1383.

Madrian, Brigitte C, and Dennis F Shea. 2001. "The Power of Suggestion: Inertia in 401(k) Participation and Savings Behavior". *The Quarterly Journal of Economics* 116, no. 4 (): 1149–1187. ISSN: 0033-5533.

Manning, Willard G., et al. 1987. "Health Insurance and the Demand for Medical Care: Evidence from a Randomized Experiment". *The American Economic Review* 77 (3): 251–277. ISSN: 0002-8282.

McKenzie, Craig R.M., Michael J Liersch, and Stacey R Finkelstein. 2006. "Recommendations implicit in policy defaults". *Psychological Science* 17 (5): 414–420. ISSN: 0956-7976.

McKiernan, Erin C, et al. 2016. "How open science helps researchers succeed." *eLife* 5. ISSN: 2050-084X.

Mettler, Suzanne, and Joe Soss. 2004. "The consequences of public policy for democratic citizenship: Bridging policy studies and mass politics". *Perspectives on politics* 2 (01): 55–73.

Morrissey, Monique. 2016. "The State of American Retirement". *Economic Policy Institute, Washington, DC*.

Morton, Rebecca B, and Kenneth C Williams. 2010. *Experimental political science and the study of causality: From nature to the lab*. Cambridge University Press.

— . 2008. *Experimentation in political science*.

Munnell, Alicia H, Anthony Webb, Francesca Golub-Sass, et al. 2012. "The national retirement risk index: An update". *Center for Retirement Research at Boston College* 1:719–744.

Murphy, Susan A. 2003. "Optimal dynamic treatment regimes". *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 65 (2): 331–355.

Newhouse, Joseph P., Rand Corporation. Insurance Experiment Group, Insurance Experiment Group Staff, et al. 1993. *Free for all?: lessons from the RAND health insurance experiment*. Harvard University Press.

Neyman, Jerzy S. 1923. "On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9.(Tlanslated and edited by DM Dabrowska and TP Speed, Statistical Science (1990), 5, 465-480)". *Annals of Agricultural Sciences* 10:1–51.

Nguyen, Tuan, Jenna Kramer, and Brent Evans. 2018. "The Effects of Grant Aid on Student Persistence and Degree Attainment: A Systematic Review and Meta-Analysis of the Causal Evidence | Stanford Center for Education Policy Analysis, Working Paper No. 18-04". `https://cepa.stanford.edu/content/effects-grant-aid-student-persistence-and-degree-attainment-systematic-review-and-meta-analysis-causal-evidence`.

OECD. 2017. *Behavioural Insights and Public Policy*. OECD Publishing. ISBN: 9789264269071.

Osman, Magda. 2004. "An evaluation of dual-process theories of reasoning". *Psychonomic Bulletin & Review* 11, no. 6 (): 988–1010. ISSN: 1069-9384. `https://link.springer.com/content/pdf/10.3758%7B%5C%%7D2FBF03196730.pdf%20http://www.springerlink.com/index/10.3758/BF03196730`.

Peirce, Charls S., and Joseph Jastrow. 1885. "On Small Differences in Sensation". *Memoirs of the National Academy of Sciences* 3:73–83.

Pichert, Daniel, and Konstantinos V Katsikopoulos. 2008. "Green defaults: Information presentation and pro-environmental behaviour". *Journal of Environmental Psychology* 28, no. 1 (): 63–73. ISSN: 0272-4944.

Prell, Mark. 2013. "Participation in the Supplemental Nutrition Assistance Program (SNAP) and Unemployment Insurance: How Tight Are the Strands of the Recessionary Safety Net?" USDA-ERS Economic Research Report No. 157. Available at SSRN: `https://ssrn.com/abstract=2357447`.

Pronin, Emily, Christopher Y Olivola, and Kathleen A Kennedy. 2008. "Doing unto future selves as you would do unto others: Psychological distance and decision making". *Personality and social psychology bulletin* 34 (2): 224–236.

Raudenbush, Stephen W. 1997. "Statistical analysis and optimal design for cluster randomized trials." *Psychological Methods* 2 (2): 173.

Reiss, Julian, and Jan Sprenger. 2014. "Scientific objectivity". Ed. by Edward N. Zalta. *The Stanford Encyclopedia of Philosophy (Winter 2017 Edition)*. `%5Curl%7Bhttps://plato.stanford.edu/archives/win2017/entries/scientific-objectivity/%7D`.

Richburg-Hayes, Lashawn, et al. 2017. *Nudging Change In Human Services: Final Report on the Behavioral Interventions to Advance Self-Sufficiency (BIAS) Project*. Tech. rep. May 2017. Office of Planning, Research, Evaluation, US Department of Health, and Human Services,

Roberts, Leanne, Irshad Ahmed, and Andrew Davison. 2009. "Intercessory prayer for the alleviation of ill health". *Cochrane Database of Systematic Reviews*, no. 2.

Sackett, David L. 1997. "Evidence-based medicine". In *Seminars in perinatology*, 21:3–5. 1. Elsevier.

Sackett, David L, et al. 1996. *Evidence based medicine: what it is and what it isn't*.

Samuelson, William, and Richard Zeckhauser. 1988. "Status quo bias in decision making". *Journal of Risk and Uncertainty* 1, no. 1 (): 7–59. ISSN: 0895-5646.

Shafir, Eldar. 2013. *The behavioral foundations of public policy*. 511. Princeton University Press. ISBN: 1400845343.

Skocpol, Theda. 1995. *Protecting soldiers and mothers*. Harvard University Press.

Stanovich, Keith E., and Richard F. West. 2000. "Individual differences in reasoning: Implicaitons for the rationality debate?" *Behavioral and Brain Sciences* 23 (5): 645–65, discussion 665–726. ISSN: 0140-525X.

Sudman, Seymour, Norman M Bradburn, and Norbert Schwarz. 1996. *Thinking about answers: The application of cognitive processes to survey methodology*. Jossey-Bass.

Sunstein, Cass R. 2015. "Nudges, Agency, and Abstraction: A Reply to Critics". *Review of Philosophy and Psychology* 6, no. 3 (): 511–529. ISSN: 1878-5166.

Sunstein, Cass R., and Lucia A. Reisch. 2014. "Automatically Green: Behavioral Economics and Environmental Protection". *Harvard Environmental Law Review* 38 (1): 127–158. ISSN: 0272-9490. arXiv: 1011.1669v3.

Sunstein, Cass R, and Richard H Thaler. 2003. "Libertarian Paternalism Is Not an Oxymoron". *The University of Chicago Law Review* 70 (4): 1159. ISSN: 0041-9494. arXiv: arXiv:1011.1669v3.

Sweller, John. 1994. "Cognitive load theory, learning difficulty, and instructional design". *Learning and instruction* 4 (4): 295–312.

Thaler, Richard H. 2016. "Behavioral economics: past, present, and future". *American Economic Review* 106 (7): 1577–1600.

Thaler, Richard H., and Shlomo Benartzi. 2004. "Save More Tomorrow^{TM}: Using Behavioral Economics to Increase Employee Saving". *Journal of Political Economy* 112, no. S1 (): S164–S187. ISSN: 0022-3808. doi:10.1093/qje/qjr055. arXiv: NIHMS150003.

Thaler, Richard H., and Cass R. Sunstein. 2008. *Nudge : Improving decisions about health, wealth, and happiness*. 293. Yale University Press. ISBN: 9780300122237.

Thaler, Richard H, and LJ Ganser. 2015. *Misbehaving: The making of behavioral economics*. WW Norton New York, NY.

Thaler, Richard H, and Cass R Sunstein. 2003. "Libertarian Paternalism". *American Economic Review* 93, no. 2 (): 175–179. ISSN: 0002-8282.

Tourangeau, Roger, Lance J Rips, and Kenneth Rasinski. 2000. *The psychology of survey response*. Cambridge University Press.

Tversky, Amos, and Daniel Kahneman. 1991. "Loss Aversion in Riskless Choice: A Reference-Dependent Model". *The Quarterly Journal of Economics* 106 (4): 1039–1061. ISSN: 0033-5533. arXiv: 225 - RamaseshanRelatQuality-2013.

Wager, Stefan, and Susan Athey. 2017. "Estimation and inference of heterogeneous treatment effects using random forests". *Journal of the American Statistical Association*, no. just-accepted.

Watts, Duncan J. 2017. "Should social science be more solution-oriented?" *Nature Human Behaviour* 1 (1): 0015.

Weiss, Carol H. 1997. "How can theory-based evaluation make greater headway?" *Evaluation review* 21 (4): 501–524.

Weiss, Carol H, et al. 2008. "The Fairy Godmother—and Her Warts Making the Dream of Evidence-Based Policy Come True". *American Journal of Evaluation* 29 (1): 29–47.