Translating Behavioral Economics Evidence into Policy and Practice

A Report for the National Academies of Sciences, Engineering, and Medicine

Elizabeth Linos

Emma Bloomberg Associate Professor of Public Policy and Management

Harvard Kennedy School

79 JFK Street

Cambridge, MA 02138

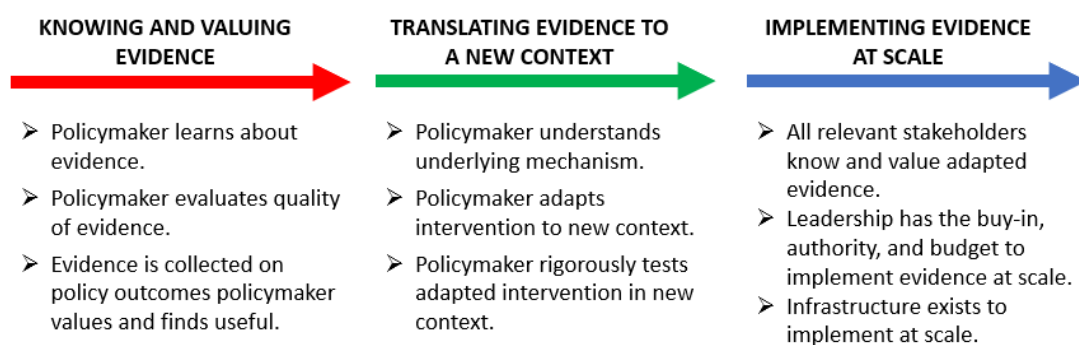elizabeth_linos@hks.harvard.edu

I.   Introduction

Over the last decade and a half, governments have increasingly harnessed evidence from behavioral economics. As the early success of Thaler and Sunstein's *Nudge* (2009) coincided with the end of a global financial crisis, cash-strapped governments grew interested in an evidence base that demonstrated that low-cost interventions could provide a clear return on investment. In parallel, a "credibility revolution" around causal inference encouraged policymakers to ask for a more rigorous evidence base around "what works" (Angrist & Pischke, 2010). Behavioral economics could provide both: early successes by "nudge units" set up in the UK and US focused on conducting high quality randomized controlled trials (RCTs) that tested low-cost interventions on outcomes with clear financial implications, such as savings and revenue collection (Halpern, 2016; Social and Behavioral Sciences Team, 2015). In recent years, the potential of behavioral economics to impact critical policy outcomes has expanded far beyond revenue collection and far beyond "nudges" (e.g., Bhargava & Manoli, 2015; Johnson & Goldstein, 2003; Lasky-Fink et al., 2021; Milkman et al., 2021; Schilbach, 2019). As of last count, there are more than 200 units across the globe specifically dedicated to the work of translating behavioral science evidence into policy (OECD, 2017).

Despite the proliferation of these units, an interdisciplinary literature documents how difficult it is to take academic evidence to scale in policy settings (e.g., Boggenshneider & Corbett, 2010; Haines, 1998; Kajermo et al., 2010). In development, Kremer et al. (2019) note that the evidence behind 41 RCTs funded by the US Agency for International Development was only adopted at scale in less than one-third of cases. DellaVigna, Kim, and Linos (2022) document a similar adoption rate of approximately 30% in behavioral interventions conducted by

US cities. Even when evidence *is* adopted, studies on the voltage drop estimate that between 50 and 90 percent of evidence-based programs will have lower effect sizes when taken to scale (List, 2022). Therefore, understanding and overcoming the barriers policymakers and practitioners face in using evidence from behavioral economics is critical for ensuring that research from behavioral economics produces the policy impacts that practitioners and scholars alike aim for.

I argue (like others before me) that the translation of evidence into policy is in itself a behavioral question with three major barriers or pain points: first, political actors need to know and value the evidence from behavioral economics; second, they need to translate the correct insight into a new context; and third, they need to act on the evidence by implementing the correct insight at scale (see Figure 1). Each of these barriers is the product of both individual- and organization-level factors. This paper describes each barrier, the factors that shape it, and discusses best practices and proposed next steps for how to address it.

Figure 1. Behavioral Barriers to Translating Evidence to Policy and Practice

KNOWING AND VALUING EVIDENCE	TRANSLATING EVIDENCE TO A NEW CONTEXT	IMPLEMENTING EVIDENCE AT SCALE
➤ Policymaker learns about evidence.	➤ Policymaker understands underlying mechanism.	➤ All relevant stakeholders know and value adapted evidence.
➤ Policymaker evaluates quality of evidence.	➤ Policymaker adapts intervention to new context.	➤ Leadership has the buy-in, authority, and budget to implement evidence at scale.
➤ Evidence is collected on policy outcomes policymaker values and finds useful.	➤ Policymaker rigorously tests adapted intervention in new context.	➤ Infrastructure exists to implement at scale.

II.   Barrier 1: Political Actors need to know and value the evidence

The first barrier in translating evidence to practice is informational: political actors need to know that evidence exists, understand its findings and level of rigor, and value it as useful. All the behavioral biases that may affect how an individual consumes new information – bounded rationality, inattention, ambiguity aversion, status quo bias, availability bias, and more – will also apply to policymakers and practitioners (see examples of these biases documented in public managers' and policymakers' decision making in Bellé et al., 2017; Moynihan & Lavertu, 2012; and Battaglio et al., 2018). As a result, in order to overcome these biases, behavioral scientists can apply some of the very tools they have developed to improve attention and understanding to communications with policymakers, such as simplifying and shortening messages, adjusting the messenger, or personalizing communications.

A growing literature suggests this may be effective. When policymakers are informed about new evidence, *how* that information is presented will impact their ability to update their priors. Nakajima (2021) shows that education leaders only update their priors when receiving evidence presented in brief, accessible explanations. Toma and Bell (2021) show that providing decision aids such as joint comparisons make policymaker preferences more elastic to evidence. Hjort et al. (2021) show that presenting easy-to-implement evidence-based strategies in revenue collection increases the likelihood of adoption, suggesting that policymakers are better able to understand and consume evidence when it is presented in behaviorally informed ways.

Beyond the types of biases that would affect any decision-maker, however, there may be other barriers to knowing about, understanding the merits of, and valuing research that are particularly salient in policy contexts. The political affiliation of the messenger (e.g., Peterson,

2018), the temporal proximity of research to a salient political issue (e.g., Henig, 2008), and the credibility of the information source (e.g., Bogenschneider & Bogenschneider, 2020) have all been shown to affect the likelihood with which a policymaker applies research to policy. In addition, political actors have been shown to vary widely in the degree to which they value research to begin with. Toma and Bell (2021) note that, at baseline, policymakers' preferences are relatively inelastic to evidence. Yet Hjort et al. (2021) show policymakers are willing to pay to receive the results of a study, suggesting some value placed on evidence. Qualitative research paints a similarly mixed picture: in a study of almost 200 state legislators, Boggenshneider & Corbett (2010) found that legislators range from "enthusiastic users of research" to "enthusiastic nonusers of research," with one-third of their participants falling into the latter camp.

Valuing research in a policy context, however, requires that evidence produces insight on the outcome that policymakers value. This seemingly tautological point has implications for the type of research that is more likely to be adopted. In some cases, the success metric presented in a research study matches the success metric a policymaker might value – a return on investment (ROI) calculation for an agency prioritizing revenue collection, for example. In other cases, however, a policymaker might have dual priorities, with only one of them being studied explicitly in the research. For example, a policymaker may need to know if a behavioral intervention is effective on average, as well as whether it is effective for a specific subgroup, to be able to speak to equity considerations. The evidence base on heterogenous effects of behavioral interventions on key policy outcomes is still relatively nascent. Moreover, there are more and more policy contexts where the success metric used in most of the research is no longer a good enough proxy for the outcome that today's policymakers prioritize. For instance, crime rates have long been used in many research contexts as a proxy for public safety. But as

the policy community has reconsidered whether crime rates are a good enough proxy for "public safety," the existing empirical evidence may not seem as valuable to policymakers who aim to show improvements in public safety. Complicating matters further, policymakers may be solving a different optimization problem all together. Although rarely described as such, policymakers may define success according to a change in a key behavioral outcome *conditional on* not surpassing a threshold of complaints, burdens placed on employees, or other negative externalities. This type of nuance is rarely provided or even studied in behavioral economics papers. As such, an open question is whether researchers (or translators of research) are able to present evidence clearly, but also on outcomes that policymakers value.

Perhaps more crucially, *what* studies are shared with policymakers will affect whether they know and understand the full evidence base on "what works". Recent debates on the average effect of behavioral interventions have shed light on the continuous challenge of publication bias in academia (see Simonsohn et al., 2014; Brodeur et al., 2016; Andrews & Kasy, 2019). Mertens et al. (2021) document a Cohen's $d = 0.43$ across 200 published studies on choice architecture. Closely following this paper, Szaszi et al. (2022) re-analyze the data and find an effect closer to zero after attempting to correct for publication bias using other methods. In contexts where publication bias is not a challenge because the whole "file drawer" is available for scrutiny, such as in DellaVigna and Linos (2022), the average effect of a behavioral intervention conducted by a US government agency is closer to a statistically significant 1.4 percentage points (Cohen's $d = 0.06$). Yet, when practitioners are asked to predict the average effect size of these government-run interventions, they predict closer to 6 percentage points – similar to the average effect in published papers (DellaVigna & Linos, 2022). As such, policymakers' understanding of the evidence is deeply affected by publication bias in a way that

influences future adoption. Put differently, if the only evidence that is shared widely is evidence of surprising success stories (that may or may not replicate), a policymaker's ability to understand what evidence to implement and what effect sizes to expect are severely limited. This could have longer term negative impacts: when insights from behavioral economics *are* implemented and found to have a lower-than-expected effect, this could create a negative feedback loop for future trust in – and use of – evidence.

Best practices and next steps

There have been many efforts to reduce informational barriers faced by policymakers by "bridging the gap" between academic writing and practitioners, some of which have been more successful than others. For example, a series of "intermediary institutions" such as think tanks and clearing houses explicitly play the role of translator, aiming to help policymakers parse through the evidence, weigh more rigorous studies more heavily, and consume the results of those studies in a manageable way. The UK government, for example, has invested heavily in "What Works" Centres that aim to increase both the supply of *and* the demand for evidence in priority policy areas. Perhaps the most successful is the Education Endowment Foundation (EEF), which is responsible for over 10% of all RCTs in education run globally, but is also central to summarizing and categorizing a broader evidence base in education policy in succinct and policy-oriented material (What Works Network, 2018). The EEF provides categorizations of different types of educational interventions based on three factors: the cost, the likely effect size, and the rigor of the evidence base. To my knowledge, a meta-evaluation of the impact of such efforts on policymakers has not been conducted.

To address the thorny challenge of publication bias, many see a wider commitment to transparency in both research design and methods as a reasonable best practice. *How* to encourage a wider commitment to transparency is still widely discussed: for example, solutions may range from more emphasis on pre-registering trials to academic journals allowing for acceptance of papers before results are collected, based purely on the design. Interestingly, one of the best examples of a commitment to transparency comes from within government. The Office of Evaluation Sciences (OES) publishes reports on all RCTs run in government, regardless of the outcome, thereby eliminating the "file drawer problem." More recently, OES has also committed to publishing all pre-analysis plans, thus joining a broader call for more transparent research.

Future research should consider the variable needs of policymakers and policy practitioners in learning about and understanding new evidence. For example, it is not yet clear *whose learning* in a government agency should be targeted in order to increase the likelihood of evidence adoption. Some theories of change place emphasis on the political leader (e.g., the mayor), others focus more on internal data or evaluation teams, while others still claim that "data and evidence literacy" should be critical for any government position and, as such, should be a bigger part of onboarding and training for all civil servants. More research is needed on whether training public sector workers at various levels of government to differentiate between rigorous and less rigorous evidence can impact the adoption of evidence through an informational channel. Future studies could also consider the impact of peer networks and communities of practice in the take-up of evidence, as there is much evidence in political science that policy ideas spread across policymakers in the same networks. Last, research on the impact of clearinghouses on how evidence is understood and shared would be particularly useful.

III. Barrier 2: Political Actors need to translate the evidence into their own context

Translating findings into a new context or for a new population is an ongoing challenge, even in purely academic circles. Camerer et al. (2018), for example, show that approximately two-thirds of social science experiments in Science or Nature replicate, with an average effect size half as large as in the original studies. When we ask policymakers to "translate evidence into policy", however, we are asking for something much more difficult. Rather than simply repeating the same intervention in a similar-enough context (e.g., conducting the same experiment on Amazon Mturk workers), we ask policymakers to produce a "conceptual replication" – parsing out the mechanism behind why something works, making an educated guess about whether the same mechanism would apply to a new population or context, potentially testing their educated guess, and using that new information to bring behavioral evidence to a new policy context. This process requires a deeper understanding of mechanisms – why it works, not simply "what works" – as well as an academic understanding of how to test hypotheses rigorously. It also requires that academic evidence be produced in a wider set of contexts so that policymakers can find and evaluate the likely relevance of results that have been tested outside of WEIRD (Western, Educated, Industrialized, Rich, and Democratic) settings.

Two challenges emerge. First, there is the question of who has the skills, authority, resources, and capacity to translate evidence into policy within government. This constraint includes having the time to learn about evidence (e.g., reading briefs, going to conferences, or reading academic findings). It also includes having the technical know-how to differentiate

between different research and evaluation methodologies, as well as between research reports of varying quality. Traditionally, governments have not tasked anyone with filling this role, or have relegated all data-related positions to IT departments, when in fact these are questions faced by program officers and policy leaders. This "skills deficit" – which should more appropriately be named a "time deficit" in many cases – has changed rapidly in multiple ways, which I outline below as best practice.

Second, until evidence on *what works, for whom,* and *when* is clear across a range of policy settings, any responsible "translation" of evidence would also entail rigorously testing the implementation of all "translated" insights from the academic literature in the new policy context. Put simply, all the practical challenges related to conducting RCTs in government are also barriers to translation of evidence (see Figure 2 for a more comprehensive set of challenges). In explaining why it is difficult to conduct RCTs in government, scholars often point to policy-makers' ethical hesitations or political constraints. While those barriers certainly may play a role, governments also face additional operational challenges in conducting RCTs that are not unique to government but may be exacerbated in policy contexts. For example, concerns about having a pure control group often come up as ethical considerations in early discussions with government partners. But even when a control group is acceptable, a key operational barrier governments face is how to collect information and outcomes on a group that does not receive treatment when the agency does not have the authority (or budget) to do independent primary data collection. Similarly, it is common to hear that RCTs are not preferred in policy contexts because it takes too long to get results that can be useful for decision-making. Yet, a main temporal challenge related to RCTs in government is often on the front end: being able to design and implement an evaluation alongside a pre-set timeline for the rollout of an intervention. The timelines on which

new policies and programs are implemented are sometimes too fast or slow for researchers, but they are certainly not dependent on research process approvals and designs. Last, there are layers of (warranted) oversight that exist when a government makes any change, including conducting an RCT. In practice, this means that conducting an RCT requires approvals from ethical or research bodies, as well as legal teams, communications teams, and political leaders. It also requires buy-in from operational teams that may not see translating evidence as part of their responsibility, nor have the technical know-how to appreciate why this methodology would provide different types of evidence than other methodologies. Taken together, these types of barriers do not eliminate the possibility of conducting RCTs in government. They do, however, limit the types of RCTs that can be conducted by a government, thus making it appear as though some evidence from academic settings is not being translated appropriately into government contexts.

Figure 2. Challenges when Conducting RCTs in Government

Defining the Outcome	- Agencies may lack authority or budget to collect data on outcomes in control group. - Agencies may not have clear visibility over outcome if intermediary providers or teams deliver services. - Agencies may not have data on or access to populations whose behavior governments want to change.
Implementing the Randomization	- Individual level data may be unavailable pre-intervention. - Technological, logistical, or operational constraints may hinder randomization. - Frontline worker discretion may limit options for randomization. - Key stakeholders may not approve of randomization.
Conducting the Intervention	- Monitoring fidelity to design may be difficult, especially when individual frontline workers are involved in implementation. - Frequent changes in staffing may affect implementation fidelity, especially for interventions that run over longer timeframes. - Timelines for evaluation design may not match timelines for intervention roll-out.

Best practices and next steps

In recent years, the increasing recognition that government needs full-time employees who understand and use data, research, and evaluation methods has led to both legislative changes and the creation of in-house evidence teams or fellowship programs. Legislative changes like the Foundations for Evidence-Based Policymaking Act (Evidence Act) establish the role of "Evaluation Officer" in each federal agency that will be tasked with playing a critical role in bringing evidence to practice. The Evidence Act also requires that all agencies create multi-year learning plans that will answer critical policy questions with more evidence. Many of these plans lend themselves to behavioral research. For example, the second priority of the Presidential Management Agenda's Learning Agenda (PMALA) includes questions on how to improve customer experience, an area where there is a wealth of behavioral economics insights. These reforms complement existing in-house teams with expertise in conducting RCTs in government, many of which have a strong behavioral slant. Following the strategic decision of the Behavioural Insights Team (BIT) in the UK and the Social and Behavioral Sciences Team (SBST) in the US to conduct RCTs on any academic idea that was translated into the public sector, cities ranging in size from New York to Scottsdale, Arizona have also developed in-house behavioral science teams that use RCTs as their primary methodology.

The OES model, in particular, is an interesting model to follow. OES operates as an in-government provider of technical assistance. The choice of project is primarily driven by the outcome, rather than the intervention, and very often demand-driven (by an agency). That is, rather than finding an opportunity to test a specific theory or behavioral insight, most such projects start by trying to move the needle on a policy outcome. The trained behavioral scientists are mainly seconded from academic institutions, providing a direct opportunity for academics to

translate their expertise into policy settings, but also to learn more about why it is so difficult to move the needle in government. When designing experiments, OES turns to both the academic literature and previous experiments within government, publishes full pre-analysis plans, and commits to publishing results after trials conclude, regardless of the results. By testing similar behavioral concepts in many settings, OES is creating an empirical evidence base on which behavioral insights translate, and documenting realistic effect sizes at scale, inclusive of any voltage drop.

Future work on translation of mechanisms to new contexts should focus primarily on reimagining how new interventions are designed, and by whom. This could include ensuring that those who are most impacted by the policy are involved in designing the adaptation, and the RCT, alongside behavioral researchers. Residents impacted by a policy design shift and front-line workers responsible for delivering a policy shift often have critical expertise in how to translate a behavioral idea into a new context and should participate in the design process as much as possible. Replicating existing findings in new and more diverse contexts also provides an opportunity to both understand mechanisms more fully and scale ideas that work for more types of populations.

IV.   Barrier 3: Political actors need to adopt the evidence at scale.

After translating evidence into a new context, policymakers face a final hurdle – scaling. The notion of a voltage drop – first made popular in implementation science as a description of why evidence produced in academic settings does not show similar effects in clinical practice – may be very relevant here (Kilbourne et al., 2007; List, 2022). For example, the strong evidence base around using social norms to reduce energy consumption shows a reduction in average

effects as the idea is scaled to more sites, due to "site selection bias." That is, the sites selected to first test a new behavioral intervention may be correlated with the likelihood of impact (Alcott, 2015). Wang and Yang (2021) show similar findings in China's policy experimentation where the site selection and resource allocation that allow for large effects in early experiments would not apply at scale.

The analogy in US government is that the first agencies to take on behavioral interventions may also be those where a behavioral intervention is most impactful, or are agencies best suited to implement interventions with fidelity. Similarly, the population that academics use in their research may be very unrepresentative of the population at large, leading to an over-weighting of behavioral interventions that work only for certain sub-populations, such as higher socioeconomic status households. Therefore, policymakers cannot – and should not – assume similar effects when they try to scale up behavioral interventions, especially if some part of the delivery of the intervention or the target population is fundamentally different than those of the original study.

Recent evidence by DellaVigna et al. (2022) suggests that even when political actors know the evidence and even when it has already been translated into their context, adoption of evidence at scale is still not guaranteed. The same patterns likely hold true in private and non-profit sectors (e.g., Athey & Luca, 2019; List, 2022). This suggests a third set of barriers that are less well explored in the behavioral economics literature, but commonly discussed in political economy, sociology, and other fields that study how organizations innovate. Across 73 RCTs in US cities, conducted in collaboration with the Behavioral Insights Team, DellaVigna et al. (2022) document that only one-third of treatments were adopted as the new status quo after the end of the RCT. To be sure, 100% adoption would not be the normative goal in this situation.

Rather, the normative goal would lie somewhere between a 78% adoption rate (the percentage of RCTs where a treatment arm had a positive effect, as measured by effect size) and a 45% adoption rate (the percentage of RCTs where a treatment arm had a positive effect, that was also statistically significant at the 5% level and the ROI was positive).[1]

The literature points to a host of factors that may influence why some organizations are better able to adopt evidence than others. Two important and related organizational distinctions emerge: first, some organizations appear more oriented to be "learning organizations" than others; and second, some organizations have more "organizational slack" to act on evidence than others. Previous evidence suggests that the main determinants of "organizational slack" are size, wealth, and personnel (see Argote & Miron-Spektor, 2011; de Vries, Bekkers, & Tummers, 2015; Besley & Persson, 2009). In particular, larger or wealthier organizations may have the resources to be able to adopt new evidence (Naranjo-Gil, 2009; Fernandez & Wise, 2010). Organizations that have set up routines and practices to transfer knowledge internally may also be better set up for organizational learning (see, for example, Moynihan & Landuyt, 2009; Zahra & George, 2002), and thus be more ready to adopt new evidence. Knowledge transfer may be particularly difficult in policy settings with political or employee turnover. A political transition makes knowledge transfer harder practically and adoption of previous administrations' "results" politically sensitive. As such, efforts to retain innovative career civil servants within government may increase the likelihood of an agency acting on the evidence it has produced.

---

[1] This calculation does not incorporate the scenario where the outcome measured is not in fact the outcome that policymakers value, because in the RCTs documented in the paper, the outcome was primarily selected by the policymakers. However, this is an important barrier in other contexts in determining what the normative "correct" rate of adoption would be.

The type of intervention attempted also affects whether policymakers will adopt it at scale. In DellaVigna et al. (2022), the main predictor of adoption was whether the treatment was adjusting an existing communication, rather than intervening in a new process. Put differently, incremental improvements to existing infrastructure were much more likely to be adopted than more large-scale interventions. If this is true in other contexts as well, this may shift how we think about the role of behavioral economics in policy. Rather than investing in insights that will yield the largest ROI or largest impact on the policy outcome, this suggests that behavioral scientists may want to consider what is most likely to get adopted when designing interventions for practice.

Best practices and next steps

To my knowledge, there is not enough research on the decision-making and behavioral barriers faced by policymakers who have already committed to using evidence-based practice, invested in conducting a rigorous test of a new idea in their own contexts, and then must turn to adoption at scale. This suggests an important behavioral question for future research: How do we make it easier for policy practitioners to implement evidence when they know about it, value it, and it is readily applicable to their contexts? Some potential avenues for further exploration may include testing ex-ante and ex-post adjustments to the RCT process. Ex-ante interventions may include setting up commitment mechanisms or other processes in advance of running an RCT that make adoption of any successful interventions more automated after the end of the trial. Ex-post interventions may include investing additional resources (either financial or human) in post-evidence adoption by creating professional roles that focus exclusively on this last stage of the process. In both cases, practitioners and behavioral scientists can lean on existing research on

organizational learning and organizational change to explore new approaches to close the post-evidence adoption gap. Addressing these issues through further research will also help inform best practices.

V.   Looking Forward

We are still in the early stages of understanding how evidence can inform practice (as well as how practice can inform evidence) and, in particular, how to effectively translate evidence from behavioral economics into meaningful policy outcomes. In order for many of the barriers cited here to be addressed, academics, policymakers, and the larger evidence-based policy-making ecosystem may have to shift their behaviors.

In academia, a new set of interdisciplinary academic collaborations between those trained in behavioral economics and those trained in implementation science or public management may help address some of the likely challenges. These types of collaborations are beginning to emerge: Beidas et al. (2021), for example, call for more integration of behavioral economics and implementation science in the field of mental health. A new sub-field of public administration scholarship dedicated to using insights from behavioral economics has also started to gain steam (e.g., Grimmelikhuijsen et al., 2017).

Similarly, there is a growing need for more researcher-practitioner collaborations where policymakers are directly involved in defining the question and outcome they care about. These new collaborations could rethink standard research designs and methods to be able to speak to the outcomes that are most valuable to policymakers. This may include placing greater emphasis on conducting high-powered studies that can detect heterogenous effects on marginalized populations, or collecting labor cost data alongside an intervention to be able to provide better

cost-benefit calculations at the conclusion of a study. To be clear, such researcher-practitioner collaborations already exist, but to become more frequent and effective in the future, new graduate-level training in how to "speak" government or how to conduct collaborative research with government partners may be required.

In policy and practitioner communities, there has been an immense investment in conducting rigorous behavioral experimentation in government, thus building an evidence base that is applicable to policy contexts. The next stage of this process is for policymakers to invest in *adoption* of evidence up-front. This could include building commitment tools or other mechanisms through which a new piece of evidence is adopted should it prove to be effective in testing. It also requires investments in the people, processes, and platforms in government that will make adoption more seamless once evidence exists. Philanthropic organizations and other funders will play a critical role in determining whether questions of implementation and adoption will be deemed as valuable as traditional behavioral economics questions and therefore, also have a critical role to play in this process.

Last, the broader ecosystem of evidence-based policy practitioners has a responsibility to ensure that evidence is truly more relevant to policy. This includes a commitment to transparency of results, regardless of outcome, and a growing comfort with small (but real) effects in policy. Actively expanding who is involved in the production of evidence, both by including more voices in the co-design of interventions and in the selection of outcome metrics, is critical. Ultimately, viewing the process of understanding, translating, and adopting evidence as a series of behavioral pain points allows the research community and the policy community to come together to address a common goal: better policy outcomes through research.

Works Cited

Allcott, H. (2015). Site selection bias in program evaluation. *The Quarterly Journal of Economics*, *130*(3), 1117–1165. https://doi.org/10.1093/qje/qjv015

Andrews, I., & Kasy, M. (2019). Identification of and correction for publication bias. *American Economic Review*, *109*(8), 2766-94.

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

Battaglio, R. P., Belardinelli, P., Bellé, N., & Cantarelli, P. (2018). Behavioral Public Administration *ad fontes:* a synthesis of research on bounded rationality, cognitive biases, and nudging in public organizations. *Public Administration Review*, *79*(3), 304–320. https://doi.org/10.1111/puar.12994

Belle, N., Cantarelli, P., & Belardinelli, P. (2017). Cognitive biases in performance appraisal: Experimental evidence on anchoring and halo effects with public sector managers and employees. *Review of Public Personnel Administration*, *37*(3), 275–294. https://doi.org/10.1177/0734371x17704891

Bogenschneider, K., & Bogenschneider, B. N. (2020). Empirical evidence from state legislators: How, when, and who uses research. *Psychology, Public Policy, and Law*, *26*(4), 413–424. https://doi.org/10.1037/law0000232

Bogenschneider, K., & Corbett, T. I. (2010). *Evidence-Based Policymaking: Insights from Policy-Minded Researchers and Research-Minded Policymakers*. Routledge.

Brodeur, A., Lé, M., Sangnier, M., & Zylberberg, Y. (2016). Star wars: The empirics strike back. *American Economic Journal: Applied Economics*, *8*(1), 1-32.

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Nave, G., Nosek, B. A., Pfeiffer, T., Altmejd, A., Buttrick, N., Chan, T., Chen, Y., Forsell, E., Gampa, A., Heikensten, E., Hummer, L., Imai, T., … Wu, H. (2018). Evaluating the replicability of social science experiments in nature and science between 2010 and 2015. *Nature Human Behaviour*, *2*(9), 637–644. https://doi.org/10.1038/s41562-018-0399-z

Dai, H., Saccardo, S., Han, M. A., Roh, L., Raja, N., Vangala, S., Modi, H., Pandya, S., Sloyan, M., & Croymans, D. M. (2021). Behavioural nudges increase COVID-19 vaccinations. *Nature*, *597*(7876), 404–409. https://doi.org/10.1038/s41586-021-03843-2

DellaVigna, S. (2009). Psychology and economics: Evidence from the Field. *Journal of Economic Literature*, *47*(2), 315–372. https://doi.org/10.1257/jel.47.2.315

DellaVigna, S., Kim, W., & Linos, E. (2022). Bottlenecks for evidence adoption. https://doi.org/10.3386/w30144

DellaVigna, S., & Linos, E. (2022). RCTS to scale: Comprehensive evidence from two nudge units. *Econometrica*, *90*(1), 81–116. https://doi.org/10.3982/ecta18709

Emanuel, E. J., Ubel, P. A., Kessler, J. B., Meyer, G., Muller, R. W., Navathe, A. S., Patel, P., Pearl, R., Rosenthal, M. B., Sacks, L., Sen, A. P., Sherman, P., & Volpp, K. G. (2015). Using behavioral economics to design physician incentives that deliver high-value care. *Annals of Internal Medicine*, *164*(2), 114. https://doi.org/10.7326/m15-1330

Grimmelikhuijsen, S., Jilke, S., Olsen, A. L., & Tummers, L. (2017). Behavioral public administration: Combining insights from public administration and psychology. *Public Administration Review*, *77*(1), 45-56.

Haines, A., & Donald, A. (1998). Making better use of research findings. *British Medical Journal (BMJ)*, *317*(7150), 72-75.

Halpern, D. (2016). *Inside the Nudge Unit: How Small Changes Can Make a Big Difference*. WH Allen.

Henig, J. R. (2009). *Spin Cycle: How Research Gets Used in Policy Debates--The Case of Charter Schools.* Russell Sage Foundation

Hjort, J., Moreira, D., Rao, G., & Santini, J. F. (2021). How research affects policy: Experimental evidence from 2,150 Brazilian municipalities. *American Economic Review*, *111*(5), 1442–1480. https://doi.org/10.1257/aer.20190830

Johnson, E. J., & Goldstein, D. (2003). Do defaults save lives?. *Science*, *302*(5649), 1338-1339.

Kajermo, K. N., Boström, A. M., Thompson, D. S., Hutchinson, A. M., Estabrooks, C. A., & Wallin, L. (2010). The BARRIERS scale--the barriers to research utilization scale: A systematic review. *Implementation Science*, *5*(1), 1-22.

Kilbourne, A. M., Neumann, M. S., Pincus H. A., Bauer M. S., Stall R. (2007). Implementing evidence-based interventions in health care: application of the replicating effective programs framework. *Implementation Science, 2(42)*. 10.1186/1748-5908-2-42

Kremer, M., Thomas, M., Gallant, S., Rostapshova, O. (2021). Is Development Innovation a Good Investment? Evidence on scaling and social returns from USAID's innovation fund. Working Paper

Lasky-Fink, J., Robinson, C. D., Chang, H. N.-L., & Rogers, T. (2021). Using behavioral insights to improve school administrative communications: The case of truancy notifications. *Educational Researcher*, *50*(7), 442–450. https://doi.org/10.3102/0013189x211000749

List, J. A. (2022). *The Voltage Effect: How to Make Good Ideas Great and Great Ideas Scale.* Currency

Mertens, S., Herberz, M., Hahnel, U. J., & Brosch, T. (2021). The effectiveness of nudging: A meta-analysis of choice architecture interventions across behavioral domains. *Proceedings of the National Academy of Sciences*, *119*(1). https://doi.org/10.1073/pnas.2107346118

Moynihan, D. P., & Lavertu, S. (2012). Does involvement in performance management routines encourage performance information use? evaluating GPRA and part. *Public Administration Review*, *72*(4), 592–602. https://doi.org/10.1111/j.1540-6210.2011.02539.x

Nakajima, N. (2021). Evidence-Based Decisions and Education Policymakers. Working Paper.

OECD (2017): "Behavioural Insights and Public Policy: Lessons From Around the World."

Peterson, M. A. (2018). In the shadow of Politics: The pathways of research evidence to health policy making. *Journal of Health Politics, Policy and Law*, *43*(3), 341–376. https://doi.org/10.1215/03616878-4366136

Roberto, C. A., & Kawachi, I. (2014). Use of psychology and behavioral economics to promote healthy eating. *American Journal of Preventive Medicine*, *47*(6), 832–837. https://doi.org/10.1016/j.amepre.2014.08.002

Schilbach, Frank. 2019. "Alcohol and Self-Control: A Field Experiment in India." *American Economic Review*, 109 (4): 1290-1322.

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-curve: a key to the file-drawer. *Journal of experimental psychology: General*, *143*(2), 534.

Social and Behavioral Sciences Team. (2015). *Annual Report*. Retrieved from https://github.com/gsa-oes/SBST-NSTC/blob/master/download/2015%20SBST%20Annual%20Report.pdf

Szaszi, B., Higney, A., Charlton, A., Gelman, A., Ziano, I., Aczel, B., Goldstein, D. G., Yeager, D. S., & Tipton, E. (2022). No reason to expect large and consistent effects of nudge

interventions. *Proceedings of the National Academy of Sciences*, *119*(31). https://doi.org/10.1073/pnas.2200732119

Thaler, R. H., & Sunstein, C. R. (2009). *Nudge: Improving Decisions About Health, Wealth, and Happiness*. Penguin Books.

Toma, M., Bell, E. (2022). Understanding and Increasing Policymakers' Sensitivity to Program Impact. Working Paper.

Vivalt, E., Coville, A. (2022). How Do Policy-Makers Update Their Beliefs? Working Paper.

Wang, S., & Yang, D. (2021). Policy experimentation in China: The Political Economy of Policy Learning. NBER Working Paper. https://doi.org/10.3386/w29402.

Weijers, R. J., de Koning, B. B., & Paas, F. (2020). Nudging in education: From theory towards guidelines for successful implementation. *European Journal of Psychology of Education*, *36*(3), 883–902. https://doi.org/10.1007/s10212-020-00495-0

What Works Network. (2018). *Five Years On*. Retrieved from https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/677478/6.4154_What_works_report_Final.pdf

Zahra, S. A. and George, G. (2002). Absorptive capacity: a review, reconceptualization, and extension. *Academy of Management Review*, 27, 185–203. https://doi.org/10.5465/amr.2002.6587995