Experimental Thinking:
A Primer on Social Science Experiments

by
James N. Druckman
druckman@northwestern.edu
Department of Political Science
Northwestern University
Scott Hall
601 University Place
Evanston, IL 60208
Phone: 847-491-7450

To Marj and Dan Druckman

Table of Contents

List of Figures

List of Tables

Preface

Experiments are a central methodology in the social sciences. Scholars from every discipline regularly turn to experiments. Practitioners rely on experimental evidence in evaluating social programs, policies, institutions, and information provision. The last decade has seen a fundamental shift in experimental social science thanks not only to their emergence as a primary methodology in many disciplines, but also to technological advances and evolving sociological norms (e.g., open science). This book is about how to "think" about experiments in light of these changes. It argues that designing a good experiment is a slow moving process (given the host of considerations) which is counter to the current fast moving temptations available in the social sciences. The book includes discussion of the place of experiments in the social science process, the assumptions underlying different types of experiments, the validity of experiments, the application of different designs (such as audit field experiments and conjoint survey experiments), how to arrive at experimental questions, the role of replications in experimental research, and the steps involved in designing and conducting "good" experiments. The goal is to ensure social science research remains driven by important substantive questions and fully exploits the potential of experiments in a thoughtful manner.

Acknowledgements

In some ways, I have been writing this book for much of my life. At an early age, my mom, Marj Druckman, taught me how to systematically address problems and think about them from multiple angles. Meanwhile, my dad, Dan Druckman, exposed me to the social sciences and taught me how to think about experiments. He even invited me to work with him on my first experiment. I thank them both and dedicate this book to them. I also thank the many students, teachers, and colleagues who have shaped my thinking on experiments. Dozens of students have challenged me to think of experiments in new ways; they also have shown me how to creatively use experiments to address a wide range of questions. Several current and former students assisted in the writing of this book including Robin Bayes, Adam Howat, Maya Novak-herzog, Richard Shafranek, Andrew Thompson, and Anna Wang. Jake Rothschild insightfully commented on the entire book and I am extremely appreciative.

I also thank my early mentors – particularly Ken Janda, Skip Lupia, and Mat McCubbins – for teaching me about social science methods and experiments. Mat read the first chapter and in his typical fashion, asked "what's your point?", thereby forcing me to make it clearer. I am indebted to several colleagues who have talked with me about experiments and/or encouraged me to flesh out my thoughts over the years. A partial list includes Adam Berinsky, Toby Bolsen, Tabitha Bonilla, Cheryl Boudreau, John Bullock, Ethan Busby, Dennis Chong, James Chu, Fay Lomax Cook, Charles Crabtree, Catherine Eckel, David Figlio, Eli Finkel, D.J. Flynn, Jeremy Freese, Laurel Harbridge-Yong, Larry Hedges, Lenka Hrbková, Shanto Iyengar, Cindy Kam, John Kane, Martin Kifer, Samara Klar, Jim Kuklinski, Thomas Leeper, Neil Malhotra, Leslie McCall, Rose McDermott, Mary McGrath, Luke Miratrix, Dan Molden, the late Becky Morton, Kevin Mullinix, Diana Mutz, Tom Nelson, Dan O'Keefe, Mike Parkin, Erik Peterson, Dave

Rand, Jenn Richeson, Josh Robison, John Barry Ryan, Monica Schneider, Libby Sharrow, the late Lee Sigelman, Nick Stagnaro, Ben Tappin, Alex Theodoridis, Paul Sniderman, Sophie Trawalter, Lynn Vavreck, Jan Voelkel, Robb Willer, and Teppei Yamamoto. I am particularly indebted to colleagues who read and commented on the book. Adam Levine did so, sending along nearly ten pages of profound comments. This led to substantial revisions which Yanna Krupnikov, Matt Levendusky, and Rune Slothuus were kind enough to read – offering extreme insight and encouragement. This is nothing new as these colleagues have been vital sounding boards for years. The same is true of Don Green who played a crucial role in supporting this endeavor.

I thank Robert Dreesen from Cambridge who guided me at every stage and the anonymous reviewers whose comments pressed me to think more deeply about many of the topics covered in the book. I am especially appreciative of Yoshi Ono who arranged for me to give several presentations on experimentation in Japan during the summer of 2019. He was a superb host and is a great colleague. I also thank the participants who attended a conference on advances in experimental political science at Northwestern University in May, 2019. Their presentations and conversations (and the edited volume, co-edited by myself and Don Green, to which they contributed) significantly shaped my thinking. I thank the National Science Foundation (SES-1822286), the Ford Motor Company Center for Global Citizenship at the Kellogg School of Management at Northwestern University (directed by David Austen-Smith), the Institute for Policy Research at Northwestern University, and the Department of Political Science at Northwestern University for supporting the conference. All of these experiences formed the basis for this book. Finally, my family Nikki, Jake, and Sam have been a constant source of support in all the best ways.

Chapter 1: Why a Primer on Social Science Experiments?

In the 1909 American Political Science Association's presidential address, A. Lawrence Lowell stated "We are limited by the impossibility of experiment. Politics is an observational, not an experimental, science…" (Lowell 1910: 7). One hundred and ten years later, the Association's president, Rogers Smith (2020: 15), raised the question of "whether an excessive emphasis on experiments will unduly constrict the questions political scientists ask…" Clearly, much has changed in political science.[1] The same can be said about many social science disciplines where experiments have evolved from a non-existent method to an accepted method to a primary method. Even psychology – where experiments have forever been a central approach – has experienced substantial changes in the last decade due to shifts in the social sciences. Specifically, massive technological advances have facilitated data access and analysis,

[1] While not my focus, another contrast between Lowell and Smith concerns their perspectives on race. Lowell served as the President of Harvard from 1909 to 1933, during which time he attempted to ban African-American students from living in freshman halls (Sollers et al. 1993). In contrast, a fair amount of Smith's work explores the incorporation of minorities into political life, such as pointing out the United States has an ascriptive tradition that involves sexism, racism, and nativism (Smith 1993). He states in his presidential address, "we are not going to be able to understand major political developments of the past, present, and future if we do not explore more deeply the politics of identity formation, using all methods that can help" (12). This comparison reveals the extent to which the discipline has changed from one that largely ignored race for the first part of its history (Blatt 2018) to one that is now recognizing the central role of studying race and identity.

and social scientists from all disciplines have called for more "open science" practices that involve transparency and replication. There is a concern, however, that these changes may cause experimentalists to become "methods driven," neither asking appropriate questions nor maximizing the potential of the method (e.g., Thelen and Mahoney 2015: 19). These apprehensions accentuate the need for careful discussion of the experimental method. That is the goal of this book: the hope is to provide readers with a way to "think" about experiments, both as users and consumers.

I do this from the perspective of a political scientist and thus I discuss the evolution of experiments in political science and use many examples from political science. That said, the arguments I make and the suggestions I offer apply to any social science application of an *experiment* – which as will be discussed in detail – I define as a study where an intervention (by a researcher or a natural event) provides the primary mechanism by which one attempts to make a casual claim. What follows can be read by those with no background and/or interest in political science.

To be clear, this manuscript is not a vigorous defense of experiments, although it will become apparent that experiments have far-reaching applications. Further, the book is neither a textbook on experimental design and analyses – many such treatments exist – nor an advanced discussion of new developments, which is available in Druckman and Green's (2021) edited volume. Instead, the book is advisory and cautionary. As social scientists forge ahead with experiments, it is crucial they do so in the most productive and careful manner, remembering what experiments are for, why and when they should be used, and how they should be designed, implemented and evaluated. In short, this book will help social scientists think about experiments more productively.

My argument can be summarized as follows: in some sense, the social sciences have become fast moving – computing technology and easily accessible data sources have led to an explosion of experiments. This follows since the historic challenges for experiments involved computing and data limitations. Today, experiments seem often to be designed and implemented quickly and not connected to the full scientific process. I argue the result is a problem. Specifically, experiments need to be thought of as one part of a scientific process and not the first part. They need to be used when appropriate and build on / have an interplay with questions, observations, and theory. Moreover, conducting a quality experiment requires thinking through a litany of decisions, ranging from how to think about problems of casual inference to considering various counterfactuals to how to interpret replications. *A good experiment is slow moving (given the host of considerations) which is counter to the current fast moving temptations available in the social sciences.* This book is about thinking through the parts that make it slow moving.[2]

I make this argument by discussing the following.

[2] In so doing, I hope to make clear that, counter to Smith's (2020) assessment, experiments when done carefully need to not constrict their reach – however, I simultaneously emphasize Smith's point that experiments have *a* place in the scientific process and the key to exploiting their power is to understand that place, including their limitations. Put another way, I follow Smith's (2020: 16**)** advice to "find ways to place our particular studies more explicitly in broader accounts of politics that can credibly indicate their importance." Smith (2020: 15) further states that the "contributions of this experimental turn are undeniable…" My hope is this book will help make them even greater.

- How to think about the place of experiments in the social scientific process – that is, addressing the question of what role experiments play in the accumulation of knowledge (e.g., relative to theory building and other steps in the scientific method).
- How to arrive at questions that experiments are useful to answer.
- How to think about the assumptions underlying different types of experiments.
- How to think about evaluating the realism and validity of experiments, as well as assessing experimental samples.
- How to think about new experimental designs.
- How to proceed after an experiment is completed, and particularly how to think about the replication of experiments.
- How to think about the process of designing and conducting a "good" experiment; by this I do not mean the technical design details but rather all of the steps one should take to ensure experiments connect to theory and advance knowledge.

My approach seeks to broaden conversations about experiments by placing them in the larger research process where one must consider issues that ostensibly have little direct connection to experimental design (e.g., questions of sampling and measurement) but are essential if one hopes to design optimal experiments. At the same time, it focuses discussion by highlighting the need to attend to precise causal inference assumptions and counterfactual thinking. Further, I offer somewhat contrarian views on experimental realism and validity – perspectives that also lead to some cautionary notes when it comes to open science practices such as pre-registration and replication. Some highlights of the points I make include the following.

- Experiments are useful only if there exists a substantively grounded question, a well-defined target population, carefully constructed measures, and clear points of comparisons. Many extant experiments fail to explicitly consider these issues. (Chapter 2.)
- All experiments – whether using random assignment or relying on experimental control – involve assumptions about causal inference that often receive scant consideration. (Chapter 2.)
- The goal of an experiment is to generalize a causal relationship – in many, but not all cases, the size of the relationship from a single experiment is less important, and the sample used to document the relationship is not crucial. (Chapter 3.)
- In most cases, assessing whether experimental treatments resemble the "real world" is misguided, as the focus should be on the theoretical construct of interest and ensuring successful delivery of the treatments. (Chapter 3.)
- Generalizing an experimental result is more complicated than evaluating the "representativeness" of the sample. In fact, the representativeness of the sample only matters when causal effects differ across relevant people (or there are clear applied goals). (Chapter 3.)
- Recent design innovations that use audit field experiments and conjoint survey experiments offer many opportunities, but these designs have limitations and are only useful under particular circumstances. Ultimately, an experimental design is only as good as the question being addressed and the hypotheses being tested. (Chapter 4.)

- The process of asking good questions for experiments come from assessing the world and the field, socializing with a diversity of people, and building on prior experiments that did not go as planned. (Chapter 5.)
- Implementing a good experiment requires the documentation of every decision in detail but that does necessitate the formal registration of a pre-analysis plan. Such a plan, if done, should not constrain an experimentalist from exploratory data analyses or incorporating theoretical ideas that had not been initially considered. (Chapter 5.)
- After an experiment is done, repeating it for replications sake has limited value; however, replication can be used as a route to innovation (by extending prior designs) and aggregation so as to isolate the size of an effect. (Chapter 5.)
- Despite all the innovations in experimental social science, the steps needed to design a quality experiment remain the same and requires situating experiments in the entire scientific process. This starts with asking a relevant substantive question and, from there, involves a lengthy iterative process, but one that is doable and rewarding. (Chapter 6.)

Who would want to read this book? I hope the material is relevant to any social scientist, including students who are just learning about social science. Those who regularly conduct experiments may find that some parts are familiar but that other parts provide novel views. Those who do not engage in experimentation, or even begrudge experiments, may learn about the logic of experimentation, novel applications, and/or how to interpret and generalize experiments. These have become requisite skills for reading social science literatures.

I proceed in this chapter with a discussion of the evolution of experiments, illustrating this development through the field of political science. I argue that the discipline currently finds itself in a new era, parts of which apply to all of the social sciences. This new era began around

2010 and reflects the confluence of experiments achieving widespread acceptance in the discipline, technological advances, and the open science movement (these latter two dynamics have affected all of the social sciences). The era introduces many opportunities but also novel challenges. Ironically, the ease of conducting experiments today has the potential to undermine their quality. I conclude the chapter by discussing the motivation for the primer and reviewing the remainder of the book.

The Evolution of Experiments

In their foundational text on quasi-experiments, Campbell and Stanley (1963: 3) explain that we must "justify experimentation on more pessimistic grounds – not as a panacea, but rather as the only available route to cumulative progress. We must instill in our students the expectation of tedium and disappointment and the duty of thorough persistence… We must expand our students' vow of poverty to include not only the willingness to accept poverty of finances, but also a poverty of experimental results." This pessimistic portrayal reflects the prevailing reality of experiments for much of social science history: experimentalists had to overcome the logistical challenges of and limited opportunities for data collection. For example, in the first random assignment experiment published in the *American Political Science Review*, Eldersveld (1956) relied on 50 students and four staff members to work for about 400 hours so as to study 500 subjects – that is, it was far from a straightforward process. Iyengar et al.'s (1982) seminal agenda setting experiment had all of 28 subjects while Druckman's (2001) study of framing had 264 subjects but took roughly five months to collect the data. In addition to the shortage of readily available experimental subjects (even student subject pools can be used for only so many experiments), experimentalists also faced the inevitable occurrence of null results. These results

were rarely published, which led Ioannidis (2005) to famously claim that "most published research findings are false."

In the last few decades, experimentation has dramatically changed. Data collection opportunities are plentiful thanks to crowdsourcing platforms, internet panels, social media contacts, and elite samples via e-mail. Computing advances allow for large-scale experiments, sometimes on literally millions of participants (e.g., Bond et al. 2012). Moreover, scholars no longer dismiss null results as inherently uninteresting thanks to the recognition that only publishing statistically significant results can skew the published research record. These developments bring with them new opportunities but also a new type of possible poverty. The ease of data collection and acceptance of non-findings means scholars might be less incentivized to design and implement quality experiments: there is much less at stake with each experiment given the relative ease of data collection and increasing acceptance of null results. On the latter, it has become essential to distinguish meaningful null results from a carefully constructed and implemented experiment as opposed to those from a poorly designed study. In short, the concerns are a poverty of poor designs, inappropriate analyses, limited use of data, and/or flawed interpretation. Even an infinite amount of data cannot compensate for a thoughtlessly designed experiment. This makes it all the more important to ensure that experimentalists design sound studies and properly analyze, interpret, and present the data from particular samples. To situate the relevance of the aforementioned concerns, I next turn to an overview of how political science (as an example) arrived at its current state when it comes to experiments.

The Expansion of Experiments in Political Science[3]

[3] Parts of this and the next section come from Druckman and Green (2021).

As mentioned, the lessons that follow apply to any social science discipline, but in this section I offer an example of how experiments have emerged in my home discipline of political science. Similar trends, albeit at different points, have occurred in other disciplines such as economics and sociology. While psychology is a clear exception, having used experiments since the start, the changes I discuss in the next section – regarding technological change and open science – are if anything most consequential in psychology.

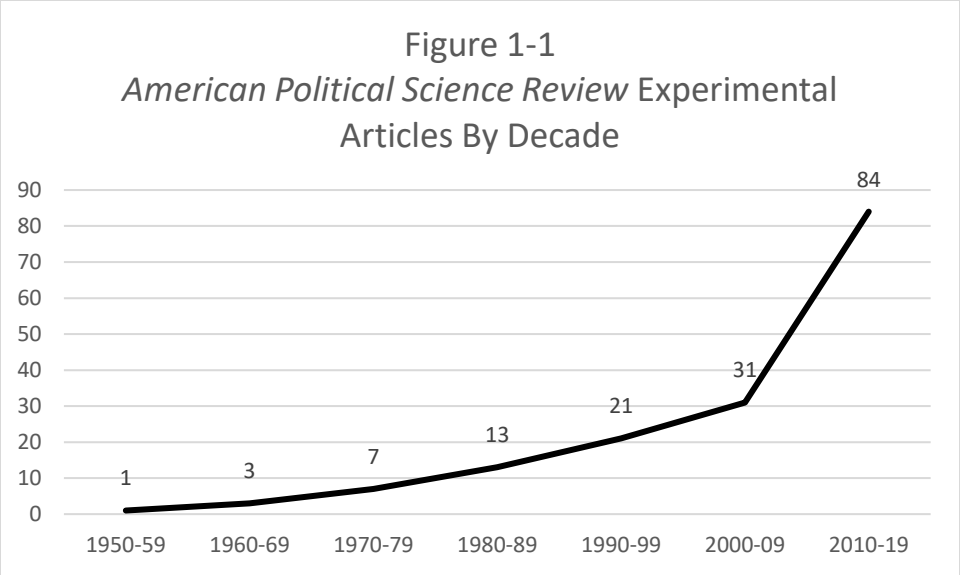
When it comes to political science, the Lowell quote with which I started the book makes clear that experiments were not present when the discipline launched. With a few notable exceptions – such as Gosnell's (1926) study of voting mobilization – experiments remained, at best, peripheral through most of the 20th century. The 1950s and 1960s saw some activity with a research program that used role-playing experimental simulations to test how situational factors affect decisions to go to war and international negotiations (e.g., Hermann and Hermann 1967, Mahoney and Druckman 1975, Guetzkow and Valadez 1981). A bit later, a short-lived journal titled *The Experimental Study of Politics* appeared. The status of experiments began to notably change, however, in the late 1980s and 1990s with experiments on Congressional committee decision-making (Fiorina and Plott 1978), media effects (Iyengar and Kinder 1987), elections (McKelvey and Ordeshook 1990), and public opinion (Sniderman et al. 1991) (also see Kinder and Palfrey 1993).

This slow emergence can be seen by charting the number of experimental articles in the discipline's flagship journal, the *American Political Science Review* (*APSR*), as documented in Figure 1-1. There were no experimental articles from 1906 to 1956 and thus the figure begins in

the 1950s and continues through 2019, reporting the number of experimental articles by decade.[4] (This is not a cumulative count of articles but rather the specific number by decade.) The figure reveals the aforementioned bump in the 1980s and 1990s and also shows the continuing increase in subsequent decades to 31 in 2000 to 2009 and 84 from 2010-2019.[5] The figure supports the claim that experiments moved from being a non-existent/marginalized method to an accepted method to a primary method.

[4] For the content analysis, I use a broad definition of "experiment" that includes experiments that rely on random assignment, that pay subjects based on their behaviors, and that exploit random or near random natural variations (i.e., "natural experiments"). I offer a formal definition of experiment in Chapter 2. I extend the timeline from Druckman et al. (2006) (also see Rogowski 2016). The total number of yearly articles in the journal remained fairly constant and thus any changes in the number of experimental articles does not reflect alterations in the total number of articles published. Finally, as with Druckman et al. (2006), I exclude Gosnell (1926) since he does not employ random assignment or other control mechanisms.

[5] Druckman and Green (2021) report results from the same content analysis. However, their coding ended as of May, 2019 (and any articled posed online at that point) whereas my analysis includes all published articles in 2019 (and thus includes 9 papers that had not been posted at the time of Druckman and Green's analysis).



While any division into particular "eras of experimentation" would contain an arbitrary element, I label three periods. First is the "pre-experimental period" that lasted until roughly 2000. As explained, the end of this period included the publication of some influential experiments; however, applications remained concentrated in a few substantive areas and experiments ostensibly were not a core part of political science curricula.[6] For instance, in their oft-used methods text from 1994, King and his colleagues (1994: 125) provide virtually no discussion of experimentation, stating only that experiments are helpful in so far as they "provide a useful model for understanding certain aspects of non-experimental design."

[6] That said, I agree that one could reasonably argue that the 1980s-1990s should be differentiated from the pre-1980s, given the growth of experiments in that time period (see Figure 1-1). Alternatively, one could merge the 1980s-1990s with 2000-2009, although, as explained shortly, two major events (i.e., Gerber and Green (2000) and the start of the Time-sharing Experiments for the Social Sciences program) signal a qualitative shift in the availability and use of experiments around 2000.

The "experimental political science 1.0" era, starting around 2000, saw the method become more mainstream/accepted as demarcated by two events. First, in 2000, Gerber and Green (2000) published their field experiment on voter mobilization. This study involved randomly assigning roughly 30,000 registered voters to receive non-partisan get-out-the-vote messages through personal canvassing, direct mails, or telephone calls. One of the more notable findings is that personal canvassing by far has the largest impact on mobilizing voters. The paper offered practical lessons for those interested in increasing turnout and spoke to an ostensible paradox at the time concerning the decline in voting turnout, linking it to the decrease in face-to-face mobilization. The paper accentuated the power of experiments for academics and practitioners. It also was only the third field experiment published in the *APSR*, and the first in nearly 20 years. The confluence of the widely discussed results along with the re-introduction of the field experimental method stimulated others to turn to field experiments. It sparked burgeoning literatures on voter mobilization (e.g., Nickerson 2008, Green et al. 2013) and vote choice (Wantchekon 2003), and more generally, ushered in the use of field experiments in other subfields (e.g. Findley et al. 2014, Hyde and Marinov 2014).[7] It also cohered with the expansion of field experiments in other disciplines (e.g., Bloom 2005).

Second, in 2001, Time-sharing Experiments for the Social Sciences (TESS) was established with support from the National Science Foundation. TESS capitalizes on economies

[7] Since 2000, roughly 30 field experiments have been published in the *APSR* and the *Annual Review of Political Science* has published several reviews of field experiments on a range of topics, including collective action (de Rooij et al. 2009), developmental economics (Humphreys and Weinstein 2009), political institutions (Grose 2014), and international relations (Hyde 2015).

of scale to enable scholars from across the social sciences, on a competitive basis, to conduct survey experiments on probability-based samples of the U.S. population (see Mutz 2011). Since its founding, TESS has supported more than 550 experiments. Clearly, the first decade of the 21st century saw the confirmation of experiments as a mainstream method.

Another change occurred about a decade into the 21st century, with the discipline establishing two major experimental institutions. This could be called "experimental political science 2.0." In 2009, Evidence in Governance and Politics (EGAP) formed as a network for those engaged in field experiments on governance, politics, and institutions. As it grew in membership and capacity, EGAP also expanded its worldwide outreach efforts to include instruction on experimental methods across the Global South. Then, in 2010, the first meeting of the American Political Science Association's section on Experimental Research took place, and a year later it voted to launch the *Journal of Experimental Political Science* (the first issue of which appeared in 2014).[8]

These institutions reflected and further stimulated the use of experiments, as is made clear in Figure 1-1 with 84 experiments published in *APSR* since 2010 (as noted). The number of

[8] Examples of other institutional developments in political science include the launching of subject pools in more than a dozen departments (Druckman et al. 2018a: 624) and the start of a Routledge book series focused on experimental political science. These institutional innovations too were accompanied by some notable publications. This list includes the explosion of experimental articles using Amazon's Mechanical Turk to furnish research participants (Berinsky et al. 2012, Mullinix et al. 2015) and, in 2011, the *Cambridge Handbook of Experimental Political Science* (Druckman et al. 2011).

experiments not only substantially increased but the reach of experiments also expanded to new domains. Examples include the study of elite responsiveness (Butler and Broockman 2011, Grose 2014, Costa 2017), social media communications (e.g., Settle 2018), governmental threats to use force (e.g., Tomz 2007, Kertzer and Brutger 2016), election monitoring (Hyde and Marinov 2014, Ichino and Schündeln 2012, Buzin et al. 2016), foreign aid (Beath et al. 2013), and governance and accountability (Dunning et al. 2019).[9] It is telling that, since 2010, roughly 44% of the experimental articles published in the *APSR* can be classified in the field of comparative politics (up from 19% during 2000-2009 and 2% during 1956-1999).[10]

In sum, the last decade has seen a dramatic growth of experimental approaches across political science. It is clear that political scientists think about and apply experiments in a very different manner than a decade ago: they think of experimentation as a primary methodology and apply it in novel domains. Understanding this new era of experimentation (starting around 2010) requires more than recognition of the growth of experiments, however. The basic nature of experimentation, across the social sciences, has changed due to technological advances and the open science movement.

Technological Change and Open Science

[9] The last decade also has seen new methods, with experiments increasingly using novel types of designs (e.g., conjoint survey experiments, audits) and samples (e.g., crowdsourcing platforms, social media) (see Druckman and Green 2021 for details). I discuss these developments in later chapters.

[10] These percentages come from a content analysis of the articles displayed in Figure 1-1.

The initial emergence of experiments (in the "pre-experimental era") followed on the heels of several technological advances. In the 1980s, the advent of computer-assisted telephone interviewing facilitated the implementation of phone-based survey experiments (Sniderman and Grob 1996). The pace of technological change has, if anything, accelerated in recent years. The costs and logistical challenges of data collection have dramatically dropped (e.g., Groves 2011), enabling researchers to access survey and behavioral data at a notably larger scale (e.g., Bond et al. 2012).

Consider four dynamics. First, as intimated, data are now much cheaper and easier to obtain thanks to the internet and the emergence of crowdsourcing platforms and commercial internet survey panels. These data can then be shared due to the growing use of public data repositories, such as Dataverse and Github. Second, social media offer researchers access to behavioral data and the opportunity to intervene experimentally (e.g., Kramer et al. 2016, Guess 2021). Third, the advent of portable computers with high resolution screens has made it easy for researchers to deploy surveys and lab-like treatments in field settings, which dramatically lowers logistical costs. Fourth, advances in computing allow researchers to analyze high-dimensional data, which is to say data with large numbers of predictors or measurements.

Apart from technological advances, the social sciences have become increasingly attuned to challenges of accumulating knowledge given perverse incentives to exaggerate the size and statistical significance of treatment effects or, conversely, to bury weak or unexpected findings. The tendency for journals to publish splashy, statistically significant findings is often termed publication bias (Brown et al. 2017). Evidence of this bias in many disciplines is not new, but political scientists have only recently begun to document it (e.g., Gerber et al. 2010, Malhotra 2021). One response to publication bias has been a call for more replications: emulating the

extant study's procedures but with new data (e.g., OSC 2015, Coppock et al. 2018). Scholars can undertake such replications in part because of a push for researchers to make their procedures, stimuli, surveys, and data publicly available. In political science, for example, most general and experimental journals require data access upon publication (Lupia and Elman 2014). The opportunities that come from data sharing and related initiatives have invigorated a call for "open science" (Nosek et al. 2015, Christensen et al. 2019) that includes standards for transparency, pre-registration of studies and analysis plans, and incentives for replication. In short, fundamental technological *and* sociological changes concerning open science have transformed social science experimentation.

Why a Primer?

The current state of experimentation can be characterized by the explosion of new data sources, the use of new measurement techniques, the introduction of previously underutilized designs, advancements in statistical methods, increased discussion about robustness and generalizability, and applications of experiments to more areas of study (Druckman and Green 2021). Conducting and publishing experiments is much easier than it was a few decades ago, and there are novel issues to consider (e.g., new sampling possibilities, design options, and open science considerations). Consequently, now is the time for a primer on experimental thinking – a call to step back and assess how to think about these opportunities. This approach will ensure that experimentalists do not move so quickly that they lose sight of (1) what experiments can contribute, and (2) the place of experiments in the larger social science enterprise. The ultimate goal is to help experimentalists conduct high quality, substantively relevant studies and avoid the aforementioned "new poverty" of experiments.

Book Outline

This book focuses on how to think about experiments, including the role of experiments in the research process, the validity of experiments, experimental designs, and the process of asking experimental questions and considering the role of experimental replications. These topics are fundamental to any experimental approach, and therefore sensible areas to cover in a primer on developments in the social sciences. This means, however, that I do not cover certain subjects, perhaps most notably ethics and advanced statistical techniques. The exclusions are not meant to minimize the importance of these (and other) areas. Rather, I opt for essential foci in thinking about experiments that can be suitably covered in a relatively short book.[11]

The book contains five additional chapters, which are written in an interlinked fashion to make the central point that, as stated, a good experiment is slow moving (given the host of considerations that are covered in this book) which is counter to the current fast moving temptations available in the social sciences. Chapter 2 starts by placing experiments in the scientific process – experiments are only useful in the context of well-motivated questions, thoughtful theories, and falsifiable hypotheses. I then turn to sampling and measurement since careful attention to these topics, despite being often neglected by experimentalists, are imperative. In the remainder of Chapter 2, I offer a detailed discussion of causal inference that I use to motivate an inclusive definition of "experiments." I view this as more than a pedantic exercise, as careful consideration of approaches to causal inference reveals the often implicit assumptions that underlie all experiments. I conclude the chapter by touching on the different goals experiments may have and the basics of analysis. The chapter serves as a reminder of the

[11] Druckman and Green's (2021) edited volume provides a comprehensive treatment of the full range of experimental topics. Also, on ethics, see Desposato (2018).

underlying logic of experimentation and the type of mindset one should have when designing experiments. A central point concerns the importance of counterfactual thinking, which pushes experimentalists to think carefully about the precise comparisons needed to test a causal claim.

Chapter 3 focuses on how to think when evaluating experiments. This includes a discussion of realism, particularly why mundane realism or resemblance to the "real world" receives far too much attention, as well as an overview of how to design experimental treatments. The chapter then turns to validity issues, offering a new way to think about external validity in assessing experiments. This includes a detailed discussion of sampling and why the onus should be more on critics of an experimental sample than on the experimentalist him/herself (i.e., to justify a sample).

In Chapter 4, I turn to experimental designs, focusing on three designs that have gained prominence in many social science applications in the last decade: audit field experiments, conjoint survey experiments, and lab-in-the-field experiments. These three designs also provide readers with examples of a type of field experiment, survey experiment, and lab experiment, respectively – the three conventional "types" of experiments employed in the social sciences (Druckman et al. 2011). The chapter reviews the basics of each design and provides prominent examples. Importantly, it also discusses limitations and challenges of the designs, or put another way, how to think about these new designs. This chapter includes a brief overview of "public policy experiments": it does so given the recent rise in studies of political elites which ultimately connect to policy-making and responsiveness. The chapter makes clear that the substantive questions being explored should drive experimental design choices and not vice versa.

Chapter 5 delves into the steps that occur prior to, during, and after an experiment – including arriving at questions to explore with an experiment, documenting the steps in the

process of conducting an experiment, and considering whether to replicate one's findings after an experiment. This discussion touches on the themes of the aforementioned open science movement, offering in many instances a cautionary perspective.

The final chapter (Chapter 6) touches on designing "good" experiments. The primary point is that regardless of changes, the fundamentals of conducing a sound experiment remain the same. I offer a list of steps that should be taken for any design. The ultimate goal is to place experiments – and recent developments in designs, sampling, and practice – in the larger landscape of the ever-changing social sciences.

Chapter 2: The Scientific Process and How to Think about Experiments

This chapter serves as a precursor. Most treatments of experiments begin with a definition; yet, I start with prior considerations including the place of experimentation in the scientific process, as well as discussions of sampling and measurement. While these latter topics are often covered in research design or statistics textbooks, many experimentalists ostensibly ignore them in practice. This leads to lower quality experiments and less progressive research agendas. I highlight essential points that need explicit attention prior to conceiving of an experiment. I then offer a discussion of causation which leads me to an inclusive definition of "experiment." I differentiate types of experiments and highlight variations in experimental goals. I conclude the chapter with a brief discussion of analyses as it pertains to design issues. My goal is to highlight a host of often overlooked considerations and assumptions that underlie any experiment. I also aim to situate experiments in the larger social science research process – by this, I mean the role experiments play in the accumulation of knowledge (e.g., relative to theory building and other steps in the scientific method).

The Scientific Process, Sampling, and Measurement

Science involves a process that produces systematized knowledge. For most, this involves the "scientific method." The basic method involves five steps: 1) ask a question, 2) develop a theory to answer the question, 3) derive testable hypotheses, 4) collect data to test those hypotheses, and 5) analyze the data. Other components involve external review, replication, and ensuring that hypotheses are falsifiable (i.e., can be proven incorrect) (Popper 1959, 1962).

Let us consider an example, one I will invoke several times throughout the book. It starts with a substantive research question: "does media coverage of campaign finance reform policies

affect opinions about those policies?" One might theorize that since people often depend on the mass media for political information, especially on issues that are more removed from their everyday lives, then media framing of the issue will shape individuals' opinions. It might be that media coverage focusing on the impact of special interests in "buying campaigns" will lead people to support policies limiting such interests (i.e., more restrictive campaign finance laws). The hypothesis is: "compared to those not exposed to media coverage, those exposed to coverage on how campaign finance laws will limit the influence of special interests will become more supportive of such laws, all else constant." The test might involve surveying people and asking them whether they were exposed to relevant media stories and about their opposition or support for campaign finance laws. One would then analyze the survey to assess whether those exposed exhibit greater support than those not exposed. If there exists a significant difference, it would constituent evidence consistent with the hypothesis; if not, the hypothesis can be rejected and one would have to dismiss or modify the theory. In that sense, one can never "prove a theory": instead, one continues to test a hypothesis that is assumed to be accurate until proven incorrect or falsified.

To be clear, my characterization of the scientific method is both debated and idealized; science in practice typically is messier (e.g., Sokal and Bricmont 1997). Failure to find evidence for a hypothesis need not lead one to abandon a theory entirely but instead can generate refinements and auxiliary hypotheses (e.g., Kuhn 1962, Lakatos 1970). Moreover, falsification – i.e., researchers make testable claims that future observations might reveal to be false (i.e., refutation) (e.g., Popper 1959, 1962, 1974, Miller 1985) – has evoked crucial debates concerning

its application in science (for general discussion, see Miller 1994, Oreskes 2019: 15-68).[12] Nonetheless, I will appeal to falsification at several points because it serves as a basic albeit idyllic standard to assess scientific progress, especially with respect to experiments (e.g., Cook and Campbell 1979: 25).[13]

Put another way, the scientific method and the idea of falsification serve as useful heuristics when it comes to thinking about social science experiments and their place in research programs. For political science, the questions asked come from subfields such as American politics, comparative politics, international relations, and political philosophy. As a discipline defined by context rather than methodology (Druckman and Lupia 2006), theories come in a host

[12] These include, for example, that scientific communities sometimes are unable to differentiate whether a given observation falsifies a theory, that it is not an accurate portrait of science in practice since many refuse to abandon falsified theories even in the absence of auxiliary hypotheses, that science is more of a community with norms than an individual endeavor, and that the idea of relying on corroborated rather than falsified theories is underdeveloped particularly since even a corroborated theory cannot be assumed to be correct.

[13] I also make arguments later in the book that may ostensibly challenge this portrait of the scientific method and falsification. For example, with regard to the former, I will discuss the invariable back and forth between theory and data. With regard to the latter, I will suggest that an experiment that refutes a hypothesis may do so for many reasons (including methodological ones), an approach that coheres with distinct philosophies of science. My goal though is not to deeply engage with alternative perspectives but rather to use the scientific method and falsification as baselines for discussion.

of guises ranging from formal theory to psychological models to historical narratives and more. The same can be said of other social sciences, to varying extents (e.g., even in economics, some rely on analytic narratives or behavioral theories). Further, there exists no agreement on what constitutes a "good theory" – deductive validity, parsimony, psychological realism, etc. Nonetheless, a primary goal, though by no means the only one, is to arrive at generalizable causal (falsifiable) statements relevant to the social, economic, and/or political world (King 1991: 1049).

The focus of this primer concerns how to test such statements (i.e., the fourth step in the idealized scientific method). There exists a multitude of approaches for conducting such tests including case studies, ethnography, focus groups, surveys, experiments, and so forth. Regardless of the method employed, questions of sampling, measurement, and analysis arise. While the discussion that follows may be review for many readers, the crucial point is that, even so, many researchers ostensibly do *not* sufficiently attend to these questions. This is especially true when it comes designing and conducting experiments.

Sampling

Most propositions involve assertions about some type of group – in the prior example, this might be all Americans, or all media consumers. Alternatively, one could study countries with the intent to understand the conditions that lead to war or trade agreements, or elected officials to isolate when they implement policies that align with public opinion, or international negotiators with the goal to discover when they display greater flexibility.

The first step involves identifying the population of interest – that is, to whom does the hypothesis apply? Identifying the population is a crucial, often missed step in experiments (Westreich et al. 2019: 439). As noted, the aforementioned hypothesis regarding media exposure

could concern all members of the U.S. population, only television viewers, or even only viewers of conservative media. Which population one studies can dramatically change the conclusions one reaches and should be determined by theory (i.e., step 2 in the scientific process). It may be that U.S. residents, on average, are moved by a special interests campaign finance story; that regular television viewers are less moved due to strong prior opinions developed from watching different news (e.g., coverage that focuses on campaign finance as a free speech issue); and that conservative viewers are not moved at all given a predilection to oppose government regulation. Researchers need to not only specify the population prior to data collection, but should also consider it at the theorizing stage – and be explicit in doing so.

If one collects data from all members of the population, it constitutes a census; unless one's population consists of a small number of relatively accessible units (e.g., a class, a small set of countries), obtaining a census quickly becomes infeasible given the time, cost, and the reality of non-response. Instead, then, one draws a sample – and as Piazza (2010: 139) states, the idea of sampling "is really quite remarkable." In essence, to make a statement about the entire population, a researcher need only collect data from a small subset, a sample of the population. In the ideal, one collects a probability sample where every unit (e.g., individual, country, elected official, negotiator) in the population has a known and non-zero chance of being in the sample. When this is done, in all likelihood, the sample will reflect the population and even a small sample allows for generalizable (causal) statements about the population.

For example, assume the population for the media example is all U.S. residents and the researcher obtains a random sample; this would require not only identifying and locating those chosen at random but also ensuring that they report their exposure to campaign finance media coverage (that focuses on special interests) and their level of support for campaign finance laws.

These data may show, for example, that those exposed are 10% more likely to be supportive of campaign finance laws (compared to those not exposed). That 10% effect constitutes a statistic one can use to infer the population parameter (i.e., the actual effect of special interests media exposure among all U.S. residents). Doing so involves basic statistics – i.e., engaging in hypothesis testing where one computes the probability of no actual relationship (i.e., the null hypothesis) in the population, given the observed relationship in the sample. If that probability is low, which often is taken to mean less than .05, one can reject the null hypothesis of no relationship and conclude that exposure to media frames of special interests correlates with the audience being more supportive of campaign finance laws.[14] Importantly, one can arrive at a conclusion about these types of relationships with fairly small samples (e.g., 120 individuals surveyed). Even finding a moderate relationship in a relatively small sample would allow one to conclude that that the relationship likely exists in the population (with some level of confidence).[15]

In practice, one can rarely, if ever, obtain a perfect probability sample given the challenge of identifying units and ensuring response – for these reasons, a large literature exists on sampling (e.g., Groves et al. 2009, Beimer 2010, Blair and Blair 2015). One topic, to which I will later return, involves how to use distinct weighting schemes with samples so that they better

[14] Some debate what probability should be seen as "significant" (e.g., Benjamin et al. 2018).

[15] The exact size of the sample needed depends on the size of the effect (i.e., whether the effect of one variable on the other – media exposure and policy support – is small, moderate, or large), and the confidence one would like to have in identifying the effect. However, the size needed is orthogonal to the size of the population.

reflect the population. For now, the key points for experimentalists are: (1) explicitly specify the population, (2) consider how the sample reflects the population, and (3) remember the goal is to make an inference to the population.[16] A crucial question concerns whether a feature of the sample leads to a distinct relationship from what one would find in the population. For example, does a sample characteristic mean one would be more or less likely to find a relationship between special interest media coverage and support for campaign finance law in the sample than in the population? As I will discuss, in some cases, the nature of the sample matters little for experimental (causal) inference, whereas in others it fundamentally shapes the conclusions one can draw.[17] To see why the former case often holds – consider a scenario where one finds special interests media coverage affects support for campaign finance laws among a sample of students. That same causal relationship would hold in the larger population of non-students unless some

[16] Doherty et al. (2006) serve as an exemplar of an experiment with an explicit discussion of the population. They explore how the amount of winnings from a state lottery affects attitudes towards estate taxes, government redistribution, etc. (They find lottery-induced affluence decreases support for estate taxes and, marginally, redistribution, but has no effect on broader issues concerning government and the economy.) Their target population is the U.S. public and they go to considerable lengths to show that their sample of lottery players generalizes to the target population (445-446).

[17] That said, in those latter cases, it could be that the goal is not a "representative probability sample" but rather a purposive sample aimed at a specific population of interest (Klar and Leeper 2019).

student characteristic such as age means students react differently than non-students to the coverage.

Discussions of sampling invariably focus on the units in the relevant study – i.e., the people or countries. This emphasis may reflect the central role of sample surveys, often with descriptive aspirations, in much of the social sciences. Here, ensuring representative samples is crucial to making statements about "all citizens" or "all voters" or "all members of a given group" (e.g., Robison et al. 2018). Yet, knowingly or not, scholars also sample the context, topic, and measures (e.g., treatments/outcomes) (Shadish et al. 2002: 23, 69-72).[18] For example, one could collect data on the impact of media exposure on campaign finance opinions during a time of national debate on the issue (e.g., in 2002 when the McCain-Feingold campaign finance law was passed, or 2009-2010 when the Supreme Court ruled on portions of the law) or during a time of little relevant discussion. In the former context, one may not find an effect because of wide-spread availability of countervailing messages, while in the latter one may find a strong effect since this exposure may be the only instance in which individuals receive relevant information.

[18] As I later discuss, these four areas of generalization cohere with Shadish et al.'s (2002: 20) external validity dimensions – they include units, treatments, outcomes, and settings. For this discussion, though, I have merged treatments and outcomes under the category of "measures," used the term "contexts" instead of "settings," and added "topics." Presumably, Shadish et al. (2002) would incorporate "topics" into "treatments," but, for sampling discussions, it is worth differentiating topics. Also, Cook and Campbell (1979) discuss generalizing over time. I incorporate this factor under context (although, as Shadish et al. (2002: 20) note, timing could be seen as relevant to all types of generalization).

One must always consider whether a finding generalizes to other contexts, including different times, distinct locations (e.g., states, countries), and varying data collection situations (e.g., laboratory, survey, field) (Druckman and Lupia 2006).[19]

Analogously, for many studies, one chooses a topic, such as campaign finance, with the goal of drawing inferences about relationships across topics – for instance, about a range of public policy issues beyond the laws of election funding. (Alternatively, one might wish to generalize across negotiation topics, voting systems, conflict situations, etc.) The topic choice can have notable consequences; indeed, Druckman and Leeper (2012a) point out that many experiments on public opinion suggest instability because they typically chose topics on which individuals do not have strong standing beliefs (e.g., campaign finance, urban sprawl, fictitious political candidates). Such issues contrast with the topics that appear in many national polls (e.g., the economy, national defense, partisan candidates with long histories) where opinions appear more stable. Finally, a given study chooses particular constructs, such as measures of media exposure and attitudes about campaign finance. Different measures may lead to divergent findings. For example, Ansolabehere et al. (2008) argue that including multiple measures for a given construct (e.g., three questions about campaign finance) will lead to more reliable findings.

In sum, sampling involves choosing not just units/individuals but also contexts, topics, and measures. Most ignore the latter three for two reasons. First, the populations of contexts, topics, and measures are ill-defined (Shadish et al. 2002: 23). Second, most studies have very small sample sizes on these dimensions since data come from a single context at a given time on

[19] A related issue for experiments is that typically participants have no choice but to be exposed to the treatment, which may not be a "typical" setting.

a single (or few) topics with select measures. Thus, it may seem infeasible to think about generalizing as it involves inferring to an unknown population from a tiny sample. The practical reality that one cannot run the same study with hundreds of distinct measures or in hundreds of contexts does not mean researchers should ignore those dimensions. Instead, they should thoughtfully consider their choices and how they fit into the extant research program. In so doing, they can also work to develop coordinated sets of experiments (Blair and McClendon 2021) and rely on design principles from small-N research (e.g., Geering 2001: 206-225, Seawright 2016: 76-106). Experimentalists typically aim to make generalizable causal statements and in so doing, they must reflect on how they chose their units, contexts, topics, and measures. This will stimulate research programs that introduce variations in context, topics, and measures, while also still attending to common concerns about having representative samples.

Measurement

The other crucial, often neglected, dimension in the design of an experiment (or any type of study) is measurement. As I will discuss, most experiments seek to measure the causal effect of a treatment (i.e., the independent variable) on an outcome (i.e., the dependent variable). Ensuring the validity of a treatment receives considerable attention and I return to this challenge below. The measurement of outcomes and variables that may condition an effect (e.g., partisanship might condition the impact of campaign finance coverage on support for reform) tend to get less attention.

In most cases, one cares about an abstract concept such as opposition or support for a public policy. There is then an inductive leap from that construct to the measure of that construct (e.g., the precise question asking respondents about their support for campaign finance laws on a 1-7 scale ranging from strongly oppose to strongly support). For many experiments, the

measurement approach involves using a survey that entails a systematic, standardized way to collect information via questions (Wright and Mardsen 2010). This requires understanding how best to ask questions and characterize an attitude, preference, emotion, or behavior with a quantitative indicator (i.e., a number). Fortunately, there exists a large literature on theories of survey response and how to write the most valid and reliable questions (e.g., Krosnick and Presser 2010, Vannette and Krosnick 2018) – on topics such as when to use open- versus close-ended questions, how many points on a scale, whether to label a scale, where and how to ask sensitive questions, etc. Generally, for any measure, be it on a survey or not, researchers need to carefully consider validity and accuracy.

Validity concerns the extent to which the measure/quantification reflects the abstracted concept. For example, if a question asks respondents to report whether they think candidates spend a lot of money it would not be a particularly valid way to capture opposition or support for a law limiting campaign spending. A measure that aims to gauge political knowledge but asks about sports trivia also would be problematic. These have poor "face validity" as they do not intuitively correspond to the concept. Also relevant is content validity, which refers to whether the measure covers the relevant aspects of the concept; for instance, does a measure of political knowledge capture all dimensions of knowledge, including that about institutions, processes, people, domestic politics, and foreign affairs (Delli Carpini and Keeter 1996: 68)? Or, for campaign finance policy, does the measure gauge attitudes about limits on both individual and organizational giving? In addition to dimensionality, a measure should have high construct validity in operationalizing the basic concept on a given dimension (Messick 1998); for example, do political knowledge items about institutions actually measure institutional knowledge? Asking if someone has heard of the Supreme Court would have low construct validity as a measure of

knowledge about what the Supreme Court does. Operationally, a measure with high construct validity also has convergent validity, meaning it correlates with measures of analogous concepts (e.g., political knowledge and political interest), and discriminant validity, implying no correlation to unrelated constructs (e.g., political knowledge and food preferences) (Campbell and Fiske 1959). Finally, criterion validity means the measure predicts, either concurrently or in the future, outcomes to which it theoretically relates (e.g., more knowledge increases the likelihood of voting or holding policy opinions that correlate with ideology; supporting campaign finance reform means supporting candidates who do so, all else constant). In Table 2-1, I offer a summary of these various types of measurement validity.

Table 2-1: Measurement Validity Concepts

Concept	Definition	Example
Face validity	Whether a measure intuitively corresponds to the concept.	A political knowledge measures that asks respondents sports trivia has low face validity.
Content validity	Whether a measure covers the relevant dimensions of a concept.	A political knowledge measure that only asks about institutions and ignores processes, people, domestic politics, and foreign affairs has low content validity.
Construct validity	Whether a measure operationalizes the concept on a given dimension.	A political knowledge question about the Supreme Court that asks if the respondent has simply heard of the Court has low construct validity.
Convergent validity	Whether a measure correlates with measures of analogous concepts.	A political knowledge measure that correlates with political interest has high convergent validity.
Discriminant validity	Whether a measure does not correlate with unrelated concepts.	A political knowledge measure that does not correlate with food preferences has high discriminant validity.
Criterion (or predictive) validity	Whether a measure predicts, either concurrently or in the future, outcomes to which it theoretically relates.	A political knowledge measure has high criterion validity if individuals with greater knowledge exhibit a stronger correlation between ideology and issue preferences.

Measures not only need to be valid but also accurate, meaning reliable and unbiased. Reliability entails arriving at the same value for the same unit (e.g., individual) if the measure is taken repeatedly (and the underlying construct remains unchanged). For example, presuming an individual does not learn more about politics between two points in time, a political knowledge measure exhibits reliability if that individual exhibits the same measured level of knowledge at

each point. Failure to do so introduces measurement error that can skew results. Fortunately, techniques to enhance reliability – such as using fully-labeled scales and the average from multiple measures for the same construct – increase reliability (e.g., Ansolabehere et al. 2008). A more perplexing challenge comes from measurement bias that occurs when the measure systematically under- or overstates the true value of the construct. Bias often occurs when it comes to items that query sensitive behaviors (e.g., drug usage) or normatively desirable actions (e.g., voting turnout). For example, the American National Election Study regularly reports that approximately 75% of the electorate votes in presidential elections, while actual turnout rates are closer to 60%. The mis-estimation reflects measurement bias. Approaches to reduce such bias exist (e.g., Tourangeau and Yan 2007, Rosenfeld et al. 2016), some of which rely on experimental methods such as list experiments (e.g., Blair et al. 2020).

The measurement issues discussed thus far do not solely apply to surveys. The increasing availability of behavioral data have led experimentalists to use outcome measures such as validated voting, social media searches, or campaign donations (e.g., Groves 2011, Peterson et al. 2017). The same considerations apply; for example, do social media posts validly reflect attitudes? Even administrative records that ostensibly seem accurate can be ridden with errors; for example, Berent et al. (2016) report that using administratively validated turnout figures ends up being as inaccurate as survey self-reports (even though the sources of inaccuracy differ). Behavioral data also can skew one's sample in unintended ways – for instance, using Twitter data to capture political rhetoric in the U.S. would be highly problematic since Twitter users tend to be ideologically extreme relative to non-users (e.g., Cohn and Quealy 2019). In short, with measurement, researchers need to carefully consider issues of validity and accuracy. This practice often requires piloting via focus groups and cognitive interviewing with individuals to

ensure they understand and view the measures as intended, or exploring how administrative/behavioral data are collected (e.g., Beatty and Willis 2007, Willis 2015).[20]

What does this challenge mean for experimentalists? The measurement issues discussed certainly do not only apply to experiments. Yet, while measurement concerns arise naturally in much observational research, particularly surveys, they usually receive less attention among experimentalists. In many cases, experiments may employ measures used in prior work. As a general practice, that may not be problematic; however, the stakes are quite high for experimentalists who typically go to great lengths to design studies meant to document the causal factors that explain a particular outcome. Failure to carefully consider the validity and accuracy of the outcome could render the experiment useless. Experimentalists should carefully specify the outcome of interest, how it will be measured, and whether that approach has high validity and accuracy. This practice might entail piloting, as discussed, and educating oneself on measurement best practices in the given substantive area.

To see how such challenges can become complicated, consider Ansolabehere et al.'s (1994) well-know and debated experimental study showing exposure to negative campaign

[20] To take one example, scholars often use "feeling thermometers" to gauge what partisans think of those from the other party. These measures ask individuals to rate those from the other party on a scale running from 0 (very cold) to 100 (very warm). Scholars had presumed this captured what voters thought of one another, but it turns out that the questions actually gauge voters' feelings about elected officials (elites) (Druckman and Levendusky 2019). This is a case where the measure did not capture what scholars had assumed.

advertising depresses intention to vote by 2.5% relative to seeing no advertisement.[21] The experiment involved randomly exposing individuals to seeing no political advertisement, an advertisement that was positive for a candidate, or an advertisement that was negative against the candidate's opponent. It measured vote intention by asking "Looking forward to the November election, do you intend to vote?" (832). Alas, as mentioned, voter turnout questions often contain considerable bias such that respondents overstate the extent to which they vote, sometimes by more than 20% (e.g., Burden 2000). One solution involves normalizing the question to make respondents feel less pressure to mis-report an ostensibly normatively preferred behavior, such as asking "In talking to people about elections, we often find that a lot of people were not able to vote because they are not registered, are sick, or they just don't have time. How about you? Do you plan to vote?" Such a question generates less bias (e.g., Belli et al. 2006, DeBell et al. 2018). So, if one wants to build on Ansolabehere's et al. study by introducing other versions of negativity (e.g., adding incivility) or conducting it a different time, should one use the same outcome measure as in the original experiment or a re-worded less biased one? The answer is not obvious; introducing a new measure means one can no longer directly speak to the replicability of the Ansolabehere et al. experiment since distinct results could stem from changes in the outcome measure. This occurs if alterations in responses to the changed measure correlate with reactions to the treatment (e.g., those who overstate vote intention are more or less sensitive to the original treatment). In most cases one would opt for what appears to be the "best" measure,

[21] Considerable debate continues about the impact of negative advertising (e.g. Lau et al. 1999, Lau et al. 2007, Krupnikov 2011).

but that choice has implications for comparability across experiments.[22] One needs to weigh these tradeoffs and *make the case explicitly,* well before designing the precise experiment. Experimentalists ought to attend to measurement: a point often neglected in practice.[23]

[22] In this case, it seems that the less biased measure uniformly increases accuracy across all groups of individuals (Persson and Solevid 2014) and thus it seems safe to use that measure. However, exploring that type of dynamic is exactly what an experimentalist needs to do prior to choosing a measure. One additional consideration is one must also consider measurement mode effects – for example, the original Ansolabehere et al. study was done in person and it seems that online measurement of vote intention is less biased (Holbrook and Krosnick 2010).

[23] In terms of surveys, a useful approach is total survey error (Biemer 2010). This refers to the difference between the population parameter (e.g., mean, proportion, regression coefficient) and estimate from the sample. It consists of sampling error – that is, the inherent uncertainty from sampling – and non-sampling errors that result from the design of the data collection (e.g., survey). These non-sampling errors are important to consider when thinking about data collection and include specification error (i.e., mis-measuring concepts a la face validity), frame errors (e.g., excluding relevant parts of the population from being eligible from the sample), nonresponse error, measurement error, and processing error. Processing errors are often underappreciated in an age of online survey software. Processing errors, though, are common due to mis-programming – such as typos in questions or assigning people to the wrong conditions/scenarios based on prior responses (e.g., a male is asked about his female identity strength due to a programing error). It is crucial that all aspects of an experiment be piloted and

Summary

In many ways, the discussion up to this point may read like a research design textbook. Yet, these issues often evade precise discussion or consideration in the design of experiments. Four takeaways include the following:

(1) Experimentalists need to always put their work in the larger context of the scientific process, which entails a substantial amount of theoretical work before the design stage.

(2) Counter to common practice, the population must be stated explicitly.

(3) Experimentalists need to recognize they are sampling not only the units but also the context, topic, and measures.

(4) Measurement warrants serious consideration and any measure must be justified and defended in light of validity and accuracy concerns.

Causal Inference and Experiments

The next question, after resolving sampling and measurement issues, concerns testing the causal hypothesis. The task of establishing that one variable leads directly to changes in another is far from straightforward and sits outside the realm of standard probability theory (Pearl 2000: 134). Much of the social sciences, over the past two decades, employs the Neyman-Rubin potential-outcome model (Neyman 1923, Rubin 1974, Imbens and Rubin 2015, Pearl and Mackenzie 2018: 269-280), the approach taken here (for formalizations of the framework, see, e.g., Druckman et al. 2011: chapter 2, Gerber and Green 2012).

---

proofed by several people who have not previously seen the study, and that the researcher check the flow of the survey.

Recall the example of the special interests campaign finance story as a cause of support for campaign finance law. Consider a study – with a representative sample and strong measures – that gauges exposure and attitudes. Regardless of method (e.g., experimental or not), a first step in establishing causation would be to reveal a correlation between frame exposure and reform support; of course, such a correlation may be spurious since those who support campaign finance laws may seek out stories consistent with that attitude. A second step then involves establishing that exposure preceded the formation of the attitude. Yet, time order does not address the well-known lurking or confounding variable problem (i.e., a variable that has an important effect on a relationship but is not included in the analysis): for example, another variable such as Democratic partisanship leads one to seek out such stories and be more supportive of campaign finance laws. As is well known, association is not equivalent to causation, even with an appropriate time sequence.

The ideal research design would be to take a given unit (e.g., person, country) and assess the impact of a variable (e.g., exposure = treatment; no exposure = control) at a single place and a single point in time: *Outcome(treatment, unit) versus Outcome(control, unit)*. For example, compare the individual's (unit's) campaign finance attitude (outcome) when exposed to the story (treatment) against his or her attitude when not exposed to the treatment (control). If the attitudes differ, this difference constitutes a causal effect of the special interests story. Alas, one cannot observe the same unit at the same time and place but under different scenarios (i.e., exposure and not). Hence, the Fundamental Problem of Causal Inference: it "is impossible to observe the value of [the treatment outcome] and [the control outcome] on the same unit… The implicit threat of the Fundamental Problem of Causal Inference is that causal inference is impossible" (Holland 1986: 947).

Holland (1986) outlines two solutions that, as will shortly become clear, are largely experimental in nature. First, the scientific solution entails observing a unit without treatment (e.g., no story) and measuring its value of the dependent variable (e.g., policy attitude). Then, at a later point in time, the same unit is exposed to the treatment (e.g., the story) and its behavior (e.g., policy attitude) is measured again. I using this technique, one must assume invariance: that the unit and all other aspects of the situation do not change from the time of the first (control) measurement to the time of treatment application.[24] Put another way, one assumes *temporal stability* such that any prior exposure does not affect future exposure (and timing sequence does not matter), and *causal transience* such that the measurement at one point in time does not affect measurement at a later point in time (Holland 1986: 948). These conditions would allow one, for instance, to measure people's campaign finance attitudes at one point in time, then expose them to the story and re-measure the attitudes to see if the story mattered. The researcher assumes nothing else happened over time to the individual or the context, and that the prior measurement did not influence the second measurement. These tests, then, typically involve within-unit or within-subject comparisons.

Another way to apply the scientific solution is to assume *unit homogeneity* – that two different units are equivalent in every (relevant) way (or that any difference is controlled for) – and expose only one of the units to the treatment. Thus, "we may seek homogenous units across time or across space" (King et al. 1994: 93). For example, one may find identical individuals who only differ in their exposure to the special interests campaign finance story. Notably, a

[24] In discussing underlying assumptions, I mostly follow Holland (1986), who notes these assumptions are not exhaustive.

presumption of unit homogeneity envelopes the idea that the units also did not self-select into treatment or control conditions as that would make them incomparable.[25]

The second solution Fundamental Problem of Causal Inference is the statistical approach that involves a focus on the average effects across an entire sample, such that some units receive the treatment and others the control. The crucial element entails random assignment of units to one of these two conditions (Holland 1986: 948) – for example, randomly assigning units in the sample to receive or not receive the special interests campaign finance story. By randomly assigning each unit, the experimenter can confidently conclude that any differences between the two groups, *on average*, stems from exposure to the treatment. This approach limits the causal interpretation to the treatment effect or random sampling variability, where the latter can be quantified. Just as with the scientific approach, the statistical approach entails assumptions. In this case, one must assume *independence*: assignment to the treatment or control is unaffected by any other relevant variable that would impact the outcome of interest (e.g., partisanship does not determine exposure to the treatment story or no-story control), including the outcome variable itself (e.g., campaign finance attitude does not determine exposure).

The benefits of random assignment result from an ability to assume on average equivalency of the randomly assigned groups (e.g., same percentage of women, same average income, same average ideology, etc.). Thus, any average difference in the outcome after one

[25] Unit homogeneity seems to implicitly assume the exclusion restriction and stable unit treatment value assumptions discussed below. These are relevant because the assumption is that units are comparable apart from exposure to the treatment which is the same presumption, via a distinct route, for the statistical solution discussed next.

group receives the treatment likely reflects a direct effect of that stimulus, if the relevant assumptions hold. It allows researchers to recover "sample average treatment effect" (SATE). Further, as I discuss below, one can explore differences in subgroups, if there is not a "constant effect" across groups (e.g., different partisans react different to the treatment) (Holland 1986: 949); this would provide the "conditional average treatment effect" (CATE). The statistical approach contains so much power that Cook and Campbell (1979: 5) state, "Random assignment is the great *ceteris paribus*—that is, other things being equal—of causal inference."

Alas, two other assumptions, beyond independence, underlie the statistical approach (Gerber and Green 2012: 39-44). First the *exclusion restriction* assumption means that outcomes vary as a function of receiving the treatment *per se* (e.g., and not due to knowledge of being in a treatment group, or different administrations/measures in the treatment versus control groups). Second the *stable unit treatment value assumption* (SUTVA) (or non-interference assumption) means "the [potential outcome] observation on one unit should be unaffected by the particular assignment of treatments to the other units" (Cox 1958, §2.4). For example, assignment to the campaign finance story by one individual should not affect the policy preference of another individual (e.g., which it could if they talked to one another about it).[26]

Table 2-2: Assumptions Underlying Solutions to the Fundamental Problem of Causal Inference

Type of Solution	Assumption	Definition	Violation Example
Scientific Solution	Temporal stability	Prior exposure does not affect future exposure (and timing sequence does not matter).	If an event, unrelated to the focus of the study, occurred between

[26] There also is an assumption of monotonicity such that there are no subjects who would receive the treatment if assigned to the control group, and would not receive the treatment if assigned to the treatment group.

			measurements that affect the outcome variable.
Scientific Solution	Causal transience	Measurement at one point in time does not affect measurement at another point in time.	If the process of obtaining the measure at one point alters the outcome at a later point (e.g., a respondent wants to ensure attitudes are consistent over-time).
Scientific Solution	Unit homogeneity	Two different units are equivalent in every (relevant) way (or any difference is controlled for).	If two respondents are very different from one another in how they might react to a treatment (e.g., a Republican and a Democrat are compared in their responses to a partisan message treatment).
Statistical Solution	Independence	Assignment to the treatment or control is unaffected by any other relevant variable that would impact the outcome of interest.	A variable that relates to the outcome variable affects which respondent receives treatment exposure (e.g., being a Democrat correlates with being exposed to a partisan message treatment).
Statistical Solution	Exclusion restriction	Outcomes vary as a function of receiving the treatment (e.g., and not due to knowledge of being in a treatment group, or different administrations/measures in the treatment versus control groups).	The experimenter measures those in the treatment in different ways than those in the control group.

Statistical Solution	Stable unit treatment value (SUTVA) (non-interference)	The outcome on a given unit is unaffected by other units.	If a respondent in the treatment group shares information from the treatment with a respondent in the control group.

In sum, the Fundamental Problem of Causal Inference presents a serious challenge for social scientists who hope to arrive at generalizable causal inferences. There exist two approaches to making causal inferences, although both require the satisfaction of assumptions. I review those assumptions along with examples of violations, for each approach, in Table 2-2. In thinking about experiments – the topic to which I next turn – one must always keep in mind the fundamental causal inference problems and the assumptions one makes.

Experiments

Perhaps ironically, despite a shared focus on the potential outcomes framework, scholars continue to disagree on what exactly constitutes a social science "experiment." For example, some focus on random assignment (e.g., Gerber and Green 2012: 8-17), while others attend to researcher intervention and control (e.g., Morton and Williams 2010: 42). Here, I define experiments in terms of how one addresses the Fundamental Problem of Causal Inference: an experiment is a study where an intervention provides the primary mechanism by which a researcher uses a procedure to resolve the Fundamental Problem of Causal Inference.

This inclusive definition contains two dimensions. First, an *intervention* refers to an event involving the researcher influencing the subjects or a natural event that facilitates inference. Second, there is a *procedure* to address the Fundamental Problem of Causal Inference: (1) the statistical solution that entails satisfying the independence, excludability, and SUTVA assumptions, or (2) the scientific solution that entails satisfying the temporal stability and causal

transience assumptions, or the unit homogeneity assumption.[27] In Table 3, I present a depiction of variations in approach to experimentation.[28]

Table 2-3: Experimental Approaches

	Statistical Solution	**Scientific Solution**
Researcher Intervention	(1) Random Assignment Experiment	(2) Within-Subject Experiment, Induced Value Experiment
Natural Intervention	(3) Random Lottery Natural Experiment	(4) Unit Homogeneity Natural Experiment (e.g., "As-if" Randomization)

In many disciplines, researcher intervention and random assignment constitute the dominant approach (cell 1). Here, the researcher *randomly* assigns values of the independent variable, such as exposure to the campaign finance story or not, and then measures the outcome, such as campaign finance attitudes (e.g., Shadish et al. 2002: 12, Morton and Williams 2008: 341, 2010: 47). For example, Druckman and Nelson (2003) randomly assigned respondents to read a news story depicting campaign finance laws as limits on special interest influence or to read no story. (Others randomly were selected to read a story that discussed campaign finance as

[27] The scientific approach is used in many physical sciences – it is what physicists do with electrons. Since electrons are all about the same and because they eliminate disturbances, researchers can isolate the control and treatment difference for each electron and for the population of electrons. This also was a common approach in the early days of psychology (e.g., the early psychophysical studies as well as behaviorism studies).

[28] Observational research is not captured in the table. That would involve situations where an intervention is not the primary mechanism for addressing the Fundamental Problem of Causal Inference.

a free speech issue.) They then had respondents complete a survey that measured their campaign finance attitudes, finding that exposure to the special interests story increased support for the laws. In other words, the average level of support in the story group exceed that in the control. When possible, this approach offers a straightforward approach since the researcher simply compares the average outcome values between the treatment and control. As discussed, it requires that satisfaction of the independence,[29] excludability, and SUTVA assumptions. Moreover, other problems can arise, such as non-compliance, where subjects do not receive a treatment to which they were assigned (i.e., they ignore the story), and, attrition where subjects who received a treatment never respond to the outcome measure (i.e., those exposed decide not to complete the follow-up survey) (Gerber and Green 2012).

The researcher intervention scientific solution (cell 2) brings with it the challenge of either observing the same unit over time and thus satisfying the assumptions that prior exposures and/or measurements do not matter (i.e., temporal stability and causal transience assumptions), or comparing non-randomly assigned units that are presumed to be identical in every way but for treatment assignment.[30] The former method would include some within-subject experiments where individuals experience multiple or all of the treatments with their behaviors being compared across exposures (see Gerber and Green 2012: 273-276). For example, Mutz (2007)

[29] Randomly assigned groups will, on average, be probabilistically equivalent, particularly if there is a sufficiently large sample (sample size concerns relate to statistical power, which I discuss below). There always exists a chance that the groups will not be on average the same.

[30] Here, the researcher is not intervening and assigning values of the independent variable as with random assignment but instead intervening to control background factors and time trends.

uses the approach to study the impact of uncivil discourse on physiological arousal; she has each experimental participant watch four different political debates that vary the civility of the exchange and the camera angle focus (i.e., whether it was a close-up). That is, every participant watches a video from each of the four possible conditions. Importantly, to address the possibility of temporal instability, she counterbalanced (i.e., randomly ordered exposure to the video from each condition) and hence could control for the order of condition exposure.[31] She finds that uncivil discourse significantly arouses subjects relative to civil discourse, and close-ups do the same (relative to angles further away). That the measures of arousal involved skin conductance tests beyond respondents' conscious control facilitated this approach (for more on within-subject designs, see Alferes 2012, Mutz 2021).

Another and perhaps more common approach in the researcher intervention scientific solution experiment (cell 2) involves addressing the unit homogeneity assumption by using financial incentives. Many economic experiments do this by employing induced value theory that makes respondents comparable sans the independent variable (Smith 1976, 1982).[32] Here,

[31] Mutz uses a Latin square design, where orders are randomized but in a way such that every condition follows every other as often as it precedes it and each condition appears exactly once in each of the four possible order positions. I thank Diana Mutz for discussing the details of the study.

[32] That said, many economic and political economy experiments also rely on the statistical/random assignment solution. For example, in his discussion of experimental economics, Guala (2005: 79) explains that in "many cases, one does not have resources to control for all background variations, because the required design would be too complicated, too

the experiment induces pre-specified characteristics in participants so that their "innate characteristics become largely irrelevant" (Friedman and Sunder 1994: 13) by offering an award medium (i.e., money) with: 1) monotonicity: subjects prefer more reward than less and do not become satiated; 2) salience: rewards for subjects depend on their actions in the study as defined by a set of rules in a way that more reward is received for a "good outcome"; 3) dominance: changes in subjects' utilities come from the reward with other influences being negligible; and 4) privacy: all subjects only learns about their own payoffs (Friedman and Sunder 1994: 12, Guala 2005: 232-236).[33]

---

costly, incompatible with the other controls, or perhaps ethically unacceptable, or simply because one does not know the full list of the relevant background factors. In such cases (i.e., in most cases), experiments rely on randomization…" (also see Guala 2009). Moreover, even where the scientific solution is crucial for causal inference, randomization might be used in assigning ordering of exposure or roles in the experiment. For example, the dictatorship experiment entails allowing a subject to decide how much of a fixed sum of money to keep for him- or herself and how much to give to another subject. This experiment is used to test predictions about self-interest (e.g., do subjects act entirely in their self-interest or do they split the difference?), yet subjects are randomly assigned to the roles (e.g., of giving or receiving money). The causal inferences of interest in such games often are about other factors affecting self-interest, as in the examples discussed below, rather than the role.

[33] These conditions mean the payment structures behaviors and is distinct from paying participants just for taking part in the study, as is often done in survey experiments and psychology-oriented lab experiments.

The approach assumes that payments contingent on subjects' decision-making makes any two subjects equivalent, except for alterations in the payment situation, or 1) some other factor that the researcher causes to vary or 2) knows to vary.[34] An example of the researcher intentionally causing variation comes from Habyarimana et. al. (2007), who explore public goods provision in ethnically heterogeneous settings (in Uganda). They do so by providing participants with money roughly equal to their per-capita daily income; participants then choose to divide it among themselves and two other individuals who varied in their ethnicities. They find that subjects do not allocate more to members of their own ethnic group than to members of other ethnic groups, except when the game is played without anonymity (i.e., their decisions were public). Here allocators give significantly more to co-ethnics, revealing the power of social norms and social sanctions. The validity of Habyarimana et al.'s causal results – regarding the power social sanctioning – rests on the assumption that subjects are virtually identical to one

[34] Incentives are assumed to wash out all potentially relevant non-comparabilities across units. In his seminal introduction of the idea, Smith (1976: 275) states that control "is the essence of experimental methodology, and in experimental exchange studies it is important that one be able to state that, as between two experiments, individual values (e.g., demand or supply) either do or do not differ in a specified way. Such control can be achieved by using a reward structure to induce prescribed monetary value on actions." As mentioned, many studies employ some mix of induced value control with randomization, such as randomizing the sequence of exposures to control for learning effects (Guala 2005: 78-80).

another and themselves each time they play the game because the payoff structure stays constant but anonymity does not.[35]

Examples where the causal factor is one that the researcher "knows to vary" come from studies that use economic games to look at behavior across contexts. For instance, Henrich et al. (2005) compare individual behavior across 15 different societies in one-shot ultimatum games: where an individual receives an endowment of money and offers an amount to a responder, who either accepts the amount or rejects it, in which case neither player receives money. They find notable variation in self-interest across societies driven largely by degree of societal market integration (with integration leading to more pro-social behavior/cooperation). Here, the payment structure makes participants equivalent other than the cultural setting in which they live (also see Cárdenas and Carpenter 2008, Eckel and Candelo 2021).[36]

The other two cells in Table 2-3 include studies where an intervention facilitates causal inference, but it does not stem from actions taken by the researcher. In the case of a "random lottery natural experiment" (cell 3), a "real world" event generates the random allocation of units

[35] The authors randomly vary the order in which participants are matched with co-ethnics or non-co-ethnics, but the key causal variable of non-anonymity is not randomly assigned. Also, Habyarimana et al. present additional experiments on the same topic (e.g., a puzzle game and a network game) where random assignment provides the key to causal inference. I thank Macartan Humphreys and Daniel Posner for discussing the details of the study.

[36] Burnham and Kurzban (2005) question the validity of these assumptions, suggesting the study is a quasi-experiment.

(individuals) into treatment and control groups (hence the statistical solution).[37] One example comes from Titiunik (2016), who relies on an intervention where some states (three in her case), after reapportionment, randomly assign state senators to serve either two-year or four year terms. They thus, on average, are equivalent to one another within a given state (i.e., the two year condition constituted the "control" and four year condition the "treatment").[38] She finds that the shorter terms lead Senators to campaign more aggressively (via spending) and do less legislatively (e.g., introduce fewer bills). Thus, short terms do not improve legislative performance. Another example is Ho and Imai's (2008) study that capitalizes on the random order in which California ballots list candidates. They show that, in general elections, minor party candidates gain .2 to .6 percentage points from being listed first on the ballot, while in primaries, partisan candidates gain 1 to 3 percentage points (with minor candidates sometimes doubling their vote share).[39] The dynamic reflects the use of a cognitive shortcut where voters

[37] Dunning (2012: 16) states that "the data used in natural experiments come from 'naturally' occurring phenomena – actually, in the social sciences, from phenomena that are often the product of social and political forces." My characterization differs from his insofar as he groups random lottery natural experiments with "as-if" random assignment natural experiments. I separate them since, in the latter case, most of the work requires establishing unit homogeneity (as I will discuss shortly).

[38] She engages in several checks to ensure the states actually use randomization.

[39] There is one complication to the randomization insofar as it is "systematically randomized treatment assignment." That is, the districts are ordered and the first one has a random ballot order. The others have specific orders that also are random but in particular ways contingent on

opt for the first candidate that seems acceptable to avoid the processing costs of analyzing them all. These types of random lottery natural experiments allow researchers to employ the statistical solution to the problem of causal inference, thanks to serendipitous random interventions over which they do not have direct control.

The final cell (cell 4) in the table likewise includes cases where the researcher does not directly impact the units. In this case, the natural intervention falls short of being strictly random, but nevertheless works in such a way that allows the researcher to make a strong case for unit homogeneity a la the scientific solution. Many refer to this method as "as-if" randomization; however, the inference relies on unit homogeneity and hence I refer to it as such. For instance, Meredith (2009) explores the impact of past voting eligibility (and possibly actually voting) on habitual voting. He does so by identifying a natural intervention of the eligibility to vote date. For example, people born on November 7, 1982 were eligible to vote in the 2000 presidential election while people born on November 8, 1982 were not. He then compares the over-time voting behavior of those born just before the cutoff to the behavior of those born just after. He finds past eligibility in presidential election years significantly increases the likelihood that a person will vote well into the future. Moreover, in 2000, eligibility increased the probability of people registering as Democrats rather than Independents. This constitutes a strong causal demonstration of early political experiences on subsequent behaviors. The key to the inference

---

the randomization of the first district. The authors adjust for this and a few other issues in their analyses. Additionally, Ho and Imai (2008) take steps to confirm randomization; they also address potential threats such as strategic candidate campaigning and/or entry as well as voters' consciously adjusting their vote choices to correct for potential ballot order effects.

involves establishing that "there are no other differences between individuals born pre- and post-election-week that may affect their subsequent participation" (Meredith 2009: 192-195). Meredith confirms this unit homogeneity assumption by comparing voting behaviors of people born around those dates in non-election years.

Another example is Hyde's (2007) study of the impact of international election observers on election-day fraud in the 2003 Armenian presidential election (measured by whether there was a reduction in the vote share of the incumbent who was expected to commit fraud). Hyde compares polling stations with observers against those without, arguing that they "were assigned to polling stations on election day using a method that I did not supervise but that comes very close to random assignment" (46). This "as-if" randomization approaches the statistical solution but ultimately it falls on Hyde to document unit homogeneity, again enveloping the independence assumption. She (48) states, since "the validity of this natural experiment rests upon this point [i.e., as-if random assignment and, in essence, comparability across groups], I will take some time to support it." She does so by detailing the arbitrary selection of polling places to the treatment, explaining the process by which observers traveled to polling stations, and arguing that stations per se, given the demographics and partisanship of the country, are unlikely to be predictable in terms of voting patterns. She further points out the authorities did not pre-announce the monitoring locations so candidates could not anticipate locations and react accordingly.[40] In short, Hyde goes to considerable length to establish the plausibility of unit homogeneity. While she appeals to "close to random" assignment (50), the inference rests on her

[40] She also offers statistical evidence for the comparability of polling places in the control and treatment groups (56-57) and checks the results with controls (57-60).

case for unit comparability a la the scientific solution of causal inference. With this method, she finds that the observers have a notable effect, reducing fraud.

The unit homogeneity natural experimental type stands out relative to other discussions of experiments that either limit the definition to strict random assignment (e.g., Gerber and Green 2012: 15-16) or group random assignment natural experiments with those that purportedly come close.[41] I use my classification, a la Table 2-3, for two reasons. First, in many of these designs, without strict random assignment, the intervention does much of the work that allows for causal inference, at least relative to statistical fixes. Accounting for this leverage allows for an inclusive definition of "experiment." Second, when the assignment lacks clear randomness, the burden falls on the researcher to establish unit homogeneity and in that sense it entails taking a scientific approach.[42] As Dunning (2012: 28) notes in these cases, "the onus is…on the researcher to make a very compelling case" to treat it as an experiment. He (2012: 239) details

[41] In his treatment of natural experiments, Dunning (2012) identifies three categories, including those that are truly randomized or "as-if random," regression-discontinuity, and instrumental-variables. Meredith (2009) is a regression discontinuity, while Hyde (2007) is an "as-if random," which I argue should be thought of distinctly from truly random. An instrumental-variables design occurs when the intervention is correlated with the topic of interest, such as using the Vietnam lottery to compare those who served in the military from those who did not – it is an instrument since those whose lottery number suggested conscription did not always end up serving (and service is the variable of interest in these studies) (Angrist 1990).

[42] As noted, though, satisfaction of the scientific assumption also must ensure independence such that units do not self-select into experimental groups (as well as excludability and SUTVA).

that such a case often requires demonstrating the comparison groups are balanced or homogenous; in essence, unit homogeneity.[43] To be clear, though, I acknowledge that in some cases the "as-if" randomization can be quite compelling – but, even there, from a technical perspective, one cannot fully rely on the statistical solution and must instead appeal to homogeneity (e.g., Titiunik 2021).[44]

[43] Sekhon and Titiunik (2012) raise another concern relevant to any type of natural intervention; specifically, the researcher needs to take considerable care to identify the proper treatment and control groups (i.e., it may not always be so clear). They offer an example of clear random assignment of gender quotas in elections. The original study focused on the random assignment that took place in one year and compared, at a later date, women candidates and election success from wards that had had the quota to those that did not (to see if having quotas at one point leads to more women representation once the quota is removed). However, those are not necessarily the correct groups to compare, since there had been randomized gender quotas at other times in various wards that may have affected outcomes at the later date.

[44] Titiunik (2021) uses a classification similar to mine, although she views cases in my final cell as observational data of a special type where the intervention facilitates casual inference. She takes particular issue with the presumption that the "as-if" randomization assumption approximates what is needed for the statistical solution (which requires equi-probable assignment). I agree with this perspective, only differing in terminology and labeling. Also, in these situations, balance tests between groups can be helpful as can auxiliary outcome "causal-process observation" (Mahoney 2010).

As an aside, I have avoided using the term "quasi-experiment." Many define a quasi-experimental study as one with an intervention without random assignment (Shadish et al. 2002: 12). In that sense, the scientific solutions in my characterization could be considered quasi-experiments. I largely avoid the term, other than a brief discussion in Chapter 4, since it seems to downgrade controlled studies where causal inference can be strong; moreover, scholars do not consistently use the term in the social sciences (Morton and Williams 2010: 25), sometimes using it to refer to studies where statistical controls dwarf the intervention in terms of making a causal inference (Dunning 2012: 19).

Summary

What makes an experiment unique relative to non-experimental work, according to my depiction, is that the intervention plays a more salient role in making a *causal* inference, relative to statistical adjustments (e.g., matching) or intensive qualitative exploration.[45] That said, in

[45] Matching entails identifying for every treated unit, a non-treated unit with similar observable characteristics against whom the effect of the treatment can be assessed. An example would be to take a sample of potential voters, find a partner for everyone – i.e., near twins, *except* the treated received a mobilizing message and the control did not, and to explore if those exposed voted. In this case, the bulk of the causal inference depends on the statistical ability to identify matches. The approach seems less powerful than experiments in identifying causal relationships (e.g., Arceneaux et al. 2006).

Also, my definition differs from some insofar as it is tied directly to causal inference. For example, I would not characterize behavioral decision making games that demonstrate technical

many cases, experiments require some statistical inspection, particularly when unit homogeneity needs to be established. Further, there exist "slippery" cases of studies where the importance of the intervention relative to statistical adjustments remains unclear; this complication accentuates the reality that the label "experiment," while useful, need not be definitive. This is the downside of my definition, but the upside includes not only being inclusive of how social scientists discuss experiments, but, more importantly, accentuating four often neglected lessons that should guide experimental design:

(1) Every experiment entails making untestable assumptions that allow one to resolve the Fundamental Problem of Casual Inference.

   a. For the statistical, random assignment approach, this involves independence, excludability, and SUTVA.[46]

   b. For the scientific approach, this involves temporal stability and causal transience or unit homogeneity.

   c. The onus is on the researcher to justify the assumptions. For this reason, many view random assignment as a stronger approach, as the assumptions ostensibly can be meet with more confidence. For instance, random assignment virtually ensures the satisfaction of independence whereas unit homogeneity requires

irrational decision making (e.g., Quattrone and Tversky 1988) as experiments, since they constitute descriptive exercises rather than causal tests.

[46] Campbell (1969) suggests the ideal is random assignment, but that is not always feasible or morally justifiable.

evidence that units in distinct "conditions" do not differ in a way that could impact the causal effect.[47]

(2) An experiment always involves an intervention that comes from the researcher or a naturally occurring event; the intervention either assigns values to units or introduces some type of control that facilitates comparison within or across units.

(3) Every experiment entails counterfactual thinking such that a comparison is being made for one value of an outcome (treatment) against another value (control). Experimentalists must carefully define what values they will compare.

(4) One hopes, when conducting an experiment, to generalize the causal relationship being tested (e.g., generalize across samples, contexts, treatments, and outcome measures).

The latter two points warrant discussion as they appear obvious in theory, but researchers typically ignore them in practice. The counterfactual thinking point pertains to all causal statements – "[e]ffects of causes are always relative to other causes (i.e., it takes two causes to define an effect" (Holland 1986: 959; also see Shadish et al. 2002: 5). This means any conjectured hypotheses must carefully specify the comparisons, and this often involves difficult choices (e.g., Sekhon and Titiunik 2012, Sniderman 2018: 261-262). For example, in the campaign finance experiment, is it relevant to compare exposure to the special interest framing to a situation with no message or a message that focuses on campaign spending as a free speech

[47] Some argue that the assumptions of the scientific solution in practice (e.g., using financial incentives) often are not met, arguing that individuals cannot be considered comparable and/or that repeated play creates a confound (Green and Tusicisny 2012; c.f., Morton and Williams 2010: 48).

issue? Researchers need to explicitly specify the comparison point before doing an experiment (e.g., they should not hypothesize that the special interests story will increase support without saying relative to what).

Generalizing the causal relationship often (but not always) means trying to generalize the existence of the relationship. This task differs from descriptive inferences such as trying to characterize the demographic and political features of a population. As I will later discuss, in the case of causal generalization, the crucial questions concern whether features of the sample, context, treatment, or outcome measures moderate the causal effect. In Popperian terms, one aims to continually test the causal proposition until rejecting it and then either abandoning it or revising it; in the ideal, one pits alternative theories against one another, but in the practice of the social sciences, one relies on multiple tests and typically revises the theory accordingly (Cook and Campbell 1979: 25-32).

Experimental Types and Goals

Beyond variations in experimental approaches to causal inference, there also exist a diversity of "types" and "goals." In terms of the former, most distinguish three types of experiments based on where the intervention occurs (Druckman et al. 2011: 6-7).[48] First are laboratory experiments, where the intervention (e.g., treatment, payment structure) occurs in a controlled setting: this could involve participants coming to the researcher's laboratory or the

[48] I use the word "type" differently from the word "approach." "Approach," as described in Table 2-3, refers to the source of the intervention (researcher or natural) and the way the Fundamental Problem of Causal Inference is addressed (statistical or scientific solution). "Type" refers to the location of the intervention.

researcher traveling to the participants to collect data in their town, such as at a community center (often called lab-in-the-field studies). Second, survey experiments occur when the intervention comes as part of the survey; for example, an experiment where some subjects receive a particular wording of a question (e.g., "Do you support assistance to the poor?") while others receive a different wording (e.g., "Do you support welfare?"). The surveys could be in-person, the phone, or via the web.[49] Third, participants in field experiments receive the treatment in naturalistic settings, as part of their daily lives, typically without knowledge that it is an intervention.[50] The natural experiments, highlighted in the second row of Table 2-3, nearly always occur in field settings (e.g., Gerber and Green 2012: 16), although in theory one could imagine a scenario where one occurs in a laboratory setting (e.g., an unexpected disruption in a controlled setting – such as fire alarm going off – might be studied in terms of its effects on performance).

Roth (1995: 22) identifies three non-exclusive roles – regardless of their type – that experiments can play. First, Roth describes "searching for facts," where the goal involves isolating "the cause of some observed regularity, by varying details of the way the experiments were conducted. Such experiments are part of the dialogue that experimenters carry on with one another." These types of experiments often complement other research methods and perhaps constitute the modal approach in much of the social sciences (other than perhaps economics)

[49] Many laboratory or field experiments use surveys to measure outcomes, but those are not survey experiments, which require that the intervention be part of the survey itself.

[50] Gerber and Green (2012: 10-13) distinguish field experiments further along four dimensions depending on the authenticity of the treatments, participants, contexts, and outcome measures.

where researchers test expectations derived from inductive reasoning/logic and build research agendas on a given topic (e.g., how media messages affect public opinion, how micro-finance programs influence well-being). A second role entails "speaking to theorists," where the goal is "to test the predictions [or the assumptions] of well-articulated formal theories... Such experiments are intended to feed back into the theoretical literature – i.e., they are part of a dialogue between experimenters and theorists." These experiments tend to test predictions from deductive formal models with precise expectations; much of this work tests game theoretic predictions on topics such as collective action, electoral systems, and coalition formation. The third role involves "whispering in the ears of princes" that facilitates "the dialogue between experimenters and policymakers… [The] experimental environment is designed to resemble closely, in certain respects, the naturally occurring environment that is the focus of interest for the policy purposes at hand." Examples of experiments with this goal would be studies to test the effect of a campaign's message on voting behavior or attitudes, or a curriculum intervention in a school. As I will later discuss, these types of experiments differ from others insofar as the precise size of the intervention's impact often matter more since those making policy (which includes government officials or leaders of non-governmental organizations) need to make cost-benefit calculations about implementation; policy oriented studies also face unique challenges regarding scaling up interventions for implementation (e.g., Al-Ubaydli et al. 2017, 2020a,b).

Summary

The goal of the experiment fundamentally affects how one interprets the results, as follows.

(1) When "searching for facts," one hopes to identify and generalize a causal effect but that precise size of the causal effect from a particular (single) study (e.g. how much does the

special interest story change campaign finance policy support relative to no story – 2%, 10%, 20%?) often is less important. The size may depend on context, timing, operationalizations that can be explored in subsequent studies. The size of effects is best explored across experiments to see if distinct circumstances vitiate or exacerbate (i.e., moderate) the size (I discuss this in more detail in Chapter 5.)

(2) When "testing theories," the precise size of the causal effect also is usually not crucial. Of particular relevance is that the experimental design closely matches the parameters of the theory.

(3) When "whispering in the ears of princes," the precise effect size – that is, the magnitude of the experimental treatment effect – often matters, as those making policies make investment decisions based on the impact.

Experimental Analysis

The analysis of an experiment ideally involves simple comparisons of the outcome measure(s) (e.g., average support for campaign finance laws, amount of donations to co-ethnics), across relevant variations in the treatment or independent variable (e.g., exposure to the special interests story or not, anonymity of donations or not). If one finds differences between the outcomes in the conditions such that the probability of those occurring by chance are very low – i.e., statistically significantly different – it constitutes evidence consistent with the causal hypothesis (such as a special interests message causing a change in policy opinion).[51] Analyses become more complex with even the partial violation of one of the aforementioned assumptions, or due to non-compliance or attrition. Gerber and Green (2012) offer a superb discussion of these

[51] See Gerber and Green (2012: 95-130) on the role of covariates in experimental analyses.

issues (also see Dunning 2012: 105-207). Here I offer a few points on analyses relevant to design and sampling issues.

In many cases, an experimentalist wants to not just document a causal relationship, but also isolate the mechanisms behind a causal effect. This involves mediation analysis, which refers to identifying the pathways through which the treatment influences the outcome. For example, the special interests campaign finance message may cause one to worry more about special interests corrupting government that in turn leads to increased support for strict campaign finance laws. The increased "worry" about special interests serves as the mediator, or the psychological process through which the message generates policy support. Alas, identifying mediational processes with most experiments proves exceptionally difficult – a point that has become widely recognized in the last decade. Gerber and Green (2012: 321) explain that it is "difficult…to extract reliable inferences about mediation from experimental data [and]… rare…to encounter a convincing demonstration in the social sciences." A common approach entails regressing the outcome variable on a variable indicating the treatment condition (e.g., from random assignment) along with possible mediators (Baron and Kenny 1986). This method falls vulnerable, however, to omitted variable bias or even reverse causation (e.g., participants rationalize their opinions on campaign finance by elevating their worry about special interests) (Bullock and Ha 2011).

There is a sizeable literature on statistical approaches (e.g., Imai et al. 2011, 2013, Imai and Yamamoto 2013; also see Glynn 2021, for an overview) and experimental design-based approaches for those hoping to isolate mediation (e.g., Gerber and Green 2012: 333-336, Bullock and Green 2020). Doing so means making explicit design choices so that mediation becomes a primary purpose of the design. For guidance on such designs, see Pirlott and MacKinnon (2016),

and for general advice on when it is justifiable to proceed with a mediation analysis, see VanderWeele (2015) and Glynn (2021). That said, researchers often neglect that while they may not be able to establish mediation, they can often rule it out. One can be confident that if a treatment does not have an effect on a potential mediator variable (either directly or moderated, which I next discuss), then that variable is an unlikely mediator. Put another way, rejecting potential mediators entails much less than corroborating mediators.[52]

The importance of identifying mediators depends the nature of a given project. For example, many scholars seek to explain the partisan polarization on climate change in the U.S. – Democrats believe in and hope to address human-induced climate change whereas Republicans typically express much more skepticism, even in the face of large amounts of information about climate change. One explanation is that individuals engage in directional motivated reasoning where they reject new information that contradicts their standing beliefs: Republicans dismiss evidence for climate change to protect what they already believe. Another explanation is that people seek to hold accurate views and Republicans simply distrust scientists and follow the advice of their party leaders who often question climate change and related policies. Here the mediational process matters– when an experiment shows Republicans do not alter their climate opinions when exposed to evidence for climate change, is the mediational process one that

[52] This is not quite as straightforward as it may sound. It is complicated by the possibility that a null effect of a treatment on a potential mediator could be hiding heterogeneity such that there is an effect but it is distinct for different sub-groups. Thus, to rule out mediation, one needs to also ensure there is not a moderated relationship between the treatment and potential mediator. I thank John Bullock and Don Green for discussion on this point.

involves motivated reasoning or the assessment of the sources? Answering that question has implications for effective messaging (e.g., try to change motivations or find distinct sources of information) (Druckman and McGrath 2019), and thus, experimentalists have sought to explore mediation by directly manipulating the potential mediators (i.e., motivations) (e.g., Bayes et al. 2020). Yet, in other situations, mediation matters less such as when it does not directly follow from a theory. For instance, the mediational processes underlying effects in experiments on voter mobilization may be intriguing but the ultimate goal involves pinpointing what messages stimulate turnout. In short, whether one should invest in designing experiments to identify mediators depends on the theory being tested (e.g., is the mediator crucial?) and the goal of the experiment (e.g., is it to test psychological mechanisms or to show the general impact of an intervention?).

Different from mediation is moderation. Moderation refers to a lack of uniformity of treatment effects. For example, it might be that Democrats, who tend to be less averse to government regulation, are persuaded by the special interests campaign message and thus the message affects (increases) their support for campaign finance laws. Yet, Republicans with their proclivity to oppose regulation remain unmoved by the message. In this case, partisanship moderates the treatment effects, or, put differently, there exist heterogeneous treatment effects based on partisanship. Another example is Malhotra and Popp's (2012) study of reactions to messages about the likelihood of future terrorist attacks in the U.S. They report an effect concentrated among Democrats who initially feared such an attack. Other examples include when a pharmaceutical drug exhibits a stronger effect on children than adults, political messages that have differential effects based on respondents' gender or race, and educational interventions that depend on students' socioeconomic statuses.

Experimental analysts often look first at the effects of treatments in the aggregate – that is, the SATE across the entire sample. Then, in cases where the researcher expects that variation between a treatment variable and an outcome variable may differ based on some third variable (such as partisanship, age, race, or socioeconomic status), he or she estimates the treatment effects by subgroup (i.e., CATE or the average treatment effect for a defined subset of units) (Gerber and Green 2012). The analysis appears ostensibly straightforward as it involves partitioning the sample into subgroups or conducting regressions where one regresses the outcome on the treatment interacted with the relevant indicator/measure for the relevant subgroups. That said, complications arise when the researcher lacks a theoretical basis for looking for moderation as some interaction is bound to be significant if one looks at enough without corrections for multiple comparisons. Generally speaking, researchers might be in one of two situations: they have a clear *a priori* expectation of a heterogeneous effect based on a precise subgroup (e.g., partisanship) and they directly test for it, or they do not in which case they might explore many potential subgroups. In the latter cases, advances in machine learning offer opportunities to soundly identify heterogeneous effects (see Green and Kern 2012; Egami and Imai 2015; Ratkovic and Tingley 2017, Ratokvic 2021).[53]

[53] Gerber and Green (2012: 311) explain that "interpretation remains ambiguous… Treatment-by-covariate interactions may provide useful descriptive information about which types of subgroups are most responsive to the treatment, but the theoretical question of whether these interactions are causal [i.e., the attribute caused the differential reaction] requires an experimental design that randomly varies what are believed to be the relevant subject attributes

Two final points on moderation concern "statistical power" and "blocking." Power concerns the probability of failing to reject the null hypothesis in the presence of a real effect (a Type II error) (see Glennerster and Takavarasha 2013: 241-297). For example, in the population, the reality could be that the special interests story affects opinions on campaign finance laws, yet in a given experiment (putting the nature of the sample – if it is representative – and measurement issues aside), there exists a chance that one will not find the relationship. Statistical power captures the chance of this *not happening* or, put another way, it refers to the probability of correctly rejecting a (false) null hypothesis. The example here would be properly rejecting the hypothesis of no relationship between the special interest message and campaign finance policy support. Statistical power is desirable, and just how much power one has in a given experiment depends on the sample size, anticipated effect size (e.g., does the message have a small, medium, or large effect on policy support?), and the statistical significance level one will employ (e.g., .01, .05, or .10). Larger effect sizes and sample sizes increase statistical power, while more stringent significance thresholds reduce it. Many computing programs allow experimentalists to identify the needed sample size in light of anticipated effect size and significance level.[54]

In assessing the amount of power one needs, an initial question concerns the importance of knowing a precise effect size versus knowing whether one has evidence consistent with a theory, regardless of the size. As I have mentioned, the effect size matters more for policy

---

or contextual characteristics." Of course, varying many moderators such as individual attributes (e.g., partisanship) can be difficult if not impossible.

[54] One such program that provides many other properties as well is DeclareDesign (Blair et al. 2019).

oriented experiments and in such cases, more power becomes essential to increase precision (Glennerster and Takavarasha. 2013: 262). Another consideration concerns whether one anticipates heterogeneous treatment effects since that requires distinct power calculations (e.g., Perugini et al. 2018, Kenny and Judd 2019); once one considers CATEs, statistical power dwindles and larger sample sizes become necessary (Gerber and Green 2012: 312). The bottom line though is, regardless, one needs to rely on prior work and pilot tests to anticipate the effect size for a given study and adjust accordingly to ensure sufficient power (e.g., design an experiment with fewer conditions if a larger sample size is needed to obtain sufficient power). Ideally, one would want power to identify a small effect but there exist invariable tradeoffs between one's confidence in the anticipate effect size (e.g., it will be medium or large) against data collection costs and design complexity. I suggest starting a research program with caution (i.e., anticipating a small effect) and build from there.[55]

In that vein, researchers sometimes design experiments as what Sniderman (2018: 262) calls "null by design," with the goal of showing that treatments others may imagine mattering in fact do not. Sniderman offers the example of Grimmer et al.'s (2015) study that sought to show (and does show) that legislative constituents respond as favorably to a legislator who only asks for a benefit for his or her constituency as to one who actually delivers a benefit (i.e., counter to what one might intuitively think). In these situations, researchers need high levels of statistical power to rule out that the null results occurred due to low power.

[55] While is reasonable to consider insufficient power as an explanation for non-significant results, one should not engage in retrospective power analysis to explain away an insignificant result and then simply ignore it (Hoening and Heisey 2001).

Finally, blocking or stratified random assignment refers to a process where the experimentalist partitions the sample into relevant subgroups, such as Democrats or Republicans or men and women, prior to randomization. Then, random assignment occurs separately within each group (Geber and Green 2102: 71-80, Glennerster and Takavarasha 2013: 153-158). Blocking can ensure that certain subgroups have a sufficient number of control and treatment units for analyses of subgroup-specific effects – for example, it can ensure enough Republicans receive the special interest story or not so as to look at that subgroup. Blocking becomes particularly useful when one has a strong *a priori* expectation of a moderating effect and the experiment's sample size is small.

Summary

(1) When scholars care greatly about identifying casual pathways or mediation, they need to refer to design approaches to mediation – few social scientists have conducted such experiments, but a growing literature offers guidance. The point is identifying mediation requires more than including potential mediator measures as outcomes and correlating them with the outcomes and treatments.

(2) Moderation refers to differential effects among subgroups – as I will later discuss, while these subgroups typically refer to variations among the units, one could also consider heterogeneity among settings, treatments, and outcome measures.

(3) The availability of programming to search for heterogeneous effects may deter scholars from theorizing in advance about moderators, but for reasons of generalization that I will later discuss, theorizing about moderators can be essential for the progression of knowledge.

Conclusion

Experiments are far from a magical elixir that leads to generalizable causal inferences. While most of the topics covered in this chapter receive attention in research design texts, they remain under-discussed in the practice of experimental social science. Indeed, the range of studies that use the label "experiment" remain hugely heterogeneous with little explicit discussion of what ties them together. Moreover, experiments seem to rarely define their target populations, and all too often experimental hypotheses fail to state, much less motivate, the comparisons involved (e.g., which experimental conditions are compared to which and why). Critiques of experiments often neglect the goal of generalizing the existence of a causal relationship, which differs from descriptively describing a population.

I have attempted to remedy these issues by laying out how experiments fit into the scientific process. I also offered a framework for thinking about different types of experiments and the assumptions underlying the causal inferences to which they lead. I emphasized the importance of counterfactual thinking in designing and presenting experiments, and how design and presentation depend on one's goals. I conclude with a reiteration of select points that an experimentalist should consider well before data collection.

- Identify the target population and justify the sampling approach (see the next chapter for further discussion on sampling).
- Recognize how the context, topic, and measures compare to related studies and the implications for generalization.
- Construct valid and accurate measures.
- Consider the assumptions underlying causal inference, given the experimental design, and whether these assumptions are met.
- Specify and theoretically justify the key comparison groups (i.e., the key counterfactual).

- State the goal of the experiment and consider implications for generalization (e.g., is the precise causal effect size important?).
- If heterogeneous effects are expected, consider blocking and ensure the sample size is sufficiently large.

Chapter 3: Evaluating Experiments: Realism, Validity, and Samples

I began the book with a discussion of the relative explosion of experiments in political science, and to varying extents, the other social sciences. I left unanswered, though, the question of what exactly drove resistance to the method for so many years (i.e., why experiments were so rare prior to the 1990s- 2000s). While it partially stemmed from the lack of technological opportunities, as discussed, it also reflected fundamental concerns about realism, external validity, and the nature of experimental samples. I turn to a discussion of these topics in this chapter. I do not re-visit past concerns, but rather, I highlight ways to think about the topics that continue to be misunderstood and mischaracterized.[56]

Realism

When it comes to assessing the contribution of a particular experiment, there are at least two ways to do so (Aronson and Carlsmith 1968, Aronson et al. 1998). First, experimental realism refers to whether "an experiment is realistic, if the situation is involving to the subjects, if they are forced to take it seriously, [and] if it has impact on them" (Aronson et al. 1985: 485). That is, experimental participants treat the situation as real, as they would approach any situation in everyday life (i.e., they are involved). Second, mundane realism concerns "the extent to which events occurring in the research setting are likely to occur in the normal course of the subjects' lives, that is, in the 'real world'" (Aronson et al. 1985: 485).[57]

[56] Parts of this chapter come from Druckman and Kam (2011).

[57] A third evaluative criterion is psychological realism, which refers to "the extent to which the psychological processes that occur in an experiment are the same as psychological processes that occur in everyday life" (Aronson et al. 1998: 132). The relevance of psychological realism

Much debate about experiments revolves around mundane realism. When participants do not match the population of interest or stimuli do not resemble something one encounters regularly in daily life, many conclude the study has limited value. If one's goal is to "whisper in the ears of princes" – that is, to directly test the causal effect of a precise intervention, such as a particular campaign message or a curriculum – then low-mundane realism indeed constitutes a problem. Such policy-oriented experiments aim to simulate a particular "reality." Yet, many (and in some fields, most) experiments do not have such a goal and in those cases, emphasis on mundane realism is misplaced (e.g., see McDermott 2002, Morton and Williams 2008: 345); *of much greater importance is experimental realism*. Failure of participants to take the study and treatments seriously compromises the basis of the causal inference, which in turn, renders the experimental results fairly meaningless (e.g., Dickhaut et al. 1972).[58] Moreover, scholars have yet to specify clear criteria for assessing mundane realism, and, as Liyanarachchi (2007: 57) explains, "any superficial appearance of reality (e.g., a high level of mundane realism) is of little comfort, because the issue is whether the experiment 'captures the intended essence of the

---

depends on one's philosophy of science (c.f., Friedman 1953, Simon 1963, 1979: 475-476; also see MacDonald 2003). I do not discuss it further as it has received relatively little attention (compared to experimental and mundane realism) in work on social science experiments.

[58] By "seriously," I mean analogous to how individuals treat the same stimuli in the settings to which one hopes to generalize (and not necessarily "serious" in a technical sense). Experimental realism may be of less importance in the case of policy oriented experiments since it may be reveal that in fact individuals do not take the stimuli being tested seriously which itself would be a relevant finding given the goal.

theoretical variables' (Kruglanski 1975: 106)" (also see Berkowitz and Donnerstein 1982: 249, Mutz 2011: 133-135).

When it comes to theory-oriented experiments, efforts to "match" to situations observed outside the study *per se* become problematic as the goal involves generalizing to the precise parameters put forth in the given theory. Plott (1991: 906) explains that the "experiment should be judged by the lessons it teaches about the theory and not by its similarity with what nature might have happened to have created." The same can be said for experiments with the goal of searching for facts; Mook (1983: 385) states that we "may assume that in order to *generalize* to 'real life,' the ... setting should *resemble* the real-life one as much as possible… This assumption is false" (italics in the original). This conclusion follows because experiments aim to isolate the impact of a causal variable, which means removing the "noise" of real life that then can be systematically built into subsequent studies. Put another way, given experiments aim to identify the effect of a given variable, it becomes inevitable they look different from the multi-dimensional nature of "real life." One often wants to establish what "can" happen; from a Popperian perspective, one cannot definitively prove what does happen. In short, unless one has strongly applied goals to assess an intervention implemented in real-time, mundane realism should play little to no role in assessing an experiment. Researchers should instead focus on experimental realism as well as the design of valid and accurate treatments, which I next discuss.

Designing Treatments

The discussion of realism raises the question of how one should construct experimental interventions that often serve as the causal variable of interest. One hopes that the treatment alters values of the independent variable (e.g., causes subjects to think about campaign finance in terms of free speech) or induces certain beliefs among participants (e.g., how much they will get

paid). The mundane realism of the treatment itself, unless one's goal is to "whisper in the ears of princes," is irrelevant. As Mutz (2021) states, "it is not necessary that people in the real world frequently encounter [the] treatment or experience... What is most important about a treatment is that it systematically and substantially changes the independent variable in the intended direction" (also see Mutz 2011: 84-99).

Mutz (2021) offers the example of a researcher interested in the influence of anxiety on political attitudes. An experimental approach requires a manipulation that induces anxiety (such as a stressful task or video clip). What matters is that anxiety changes – not how it is done. Mutz (2021) provides another example of using *Reader's Digest* stories to alter people's level of social trust (e.g. the story either described someone absconding with a wallet found on the ground or returning it) (Mutz 2005).[59] The intervention altered social trust – that it came from *Reader's Digest* did not matter.

These examples make clear that sound treatments do not depend on their mundane realism but rather on whether the relevant independent variable changes. When it comes to evaluating treatments, researchers should *not* trust themselves to validate them. For example, one should not assume that a news article that seems to talk about special interests and campaign finance will be read as such by experimental participants, that a particular audio recording will stimulate anxiety, or that a story will generate social trust. A crucial step taken in the design of an experiment entails validating the intervention with a sample that matches the experimental participants and/or the participants themselves.

[59] This study explored the impact of social trust on online purchasing behavior.

One approach to validation involves piloting the intervention – one need not test the outcome variables of interest but instead assess whether participants interpret and react to the intervention as presumed (e.g., increased anxiety or social trust). Piloting has the advantage of allowing one to evaluate different approaches before implementing the actual experiment. Ideally, one pilots on a sample drawn from the same population as the experiment. If that is not possible, however, one should carefully think about possible differences between the pilot sample and the experimental sample.[60] For example, many experiments use last names in treatments to signify race, such as exposing some people to vignettes about whites (e.g., using the name "Larsen") and others about African-Americans (e.g., using the name "Washington"). The common practice is to identify names using objective birth data or those used in other studies. Yet, as Crabtree and Chykina (2018: 21) note, the "potential problem here is that scholars often ignore the extent to which these choices accurately map onto how individuals perceive names." They show notable heterogeneity in race of names across U.S. counties (e.g., Washington is a common African-American name in some counties more than others). This finding exemplifies the importance of piloting treatments on the target populations when possible and, when not, carefully considering contextual confounds. Ultimately, a good experiment offers

[60] When it comes to applying induced value theory as part of the intervention in experiments, a common manipulation check is for researchers to have participants partake in practice sessions (e.g., Plott and Pogorelskiy 2017). Subjects are not included in the main experiment if they fail to act in accordance with the requisites of induced value theory given the rewards offered (e.g., Cooper et al. 1993: 1309). Indeed, if they fail to do so, it is a manipulation failure. This testing is done prior to any experimental treatments and so in a sense is a pilot.

corroboration for the intervention that instills confidence that the independent variable is a valid and accurate measure.

In addition to piloting, one can incorporate a manipulation check into the experiment itself to empirically assess whether respondents receive and perceive the treatment as intended. For example, this type of check would involve post-treatment questions about how one interprets an article or about their anxiety (see Mutz 2021 for detailed discussion). Kane and Barabas (2019) differentiate subjective manipulation checks from factual manipulation checks. The former involves asking respondents their perceptions of the treatment; for instance, does a treatment meant to change people's beliefs about the reinstatement of the draft actually change those beliefs (Horowitz and Levendusky 2011)? The latter entails testing for specific objective facts, such as asking respondents what news source a vignette came from (e.g., when it was clearly labeled as coming from a source such as CNN). Subjective checks ensure the treatment captures "the latent variables of interest" (Kane and Barabas 2019: 247). A factual information check can potentially not only confirm treatment validity but also be used to test for attention, which helps to establish experimental realism.

In practice, scholars should keep three lessons in mind. First, some researchers use attention checks that are unrelated to the experimental treatments, such as asking a trick question like "In the following question, please check only baseball even though it asks for 'all that apply.' What are your favorite sports? Check all that apply." These questions identify those not paying attention but provide no information about the manipulation.[61] Factual manipulation

[61] A robust literature on attention checks assesses the prevalence of inattentiveness (Maniaci and Rogge 2014) and its consequences (Oppenheimer et al. 2009, Hauser and Schwarz 2015).

checks have the advantage of serving as such attentional checks while also checking the manipulation itself. Second, if one includes a manipulation check within the experiment, rather than (or in addition to) relying on pilot tests, a prudent approach is to include the manipulation check questions after the key outcome measures to avoid interactions with the treatment itself (Berinsky et al. 2014, Hauser et al. 2018).[62]

Third is a significant caveat to the prior point. While manipulation checks within an experiment provide useful information, one should not selectively remove respondents from conditions based on their answers, and then proceed with analyses among that subset. The removal of respondents undermines the solutions to causal inference, and doing so requires non-trivial statistical adjustments (Angrist et al. 1996). Thus, when respondents fail a post-treatment manipulation check, it provides helpful information but one should not remove those who failed and then analyze the data as an experiment.

This point accentuates the usefulness of piloting to avoid discovering problematic manipulations too late. Alternatively, Kane et al. (2020) suggest an approach – applicable to vignette experiments – that involves having all participants read a mock vignette similar in structure to the experimental (treatment) vignette and then answer factual questions about it. Subsequently the experimenter randomly assigns participants, regardless of their success on the factual questions, to conditions. For example, participants read a story about scientific publishing (e.g., concerning rules about public access to federally funded research) and then answer

[62] Kane and Barabas (2019) offer evidence that placing manipulation checks before or after the main outcome measure makes little difference; however, placing them after eliminates the possibility that the manipulation check itself affects responses.

questions about the details of the story (e.g., how long did the article say publishers could wait before releasing federally funded research to the public?). Then the experimenter randomly assigns participants to conditions; Kane et al. offer the example of a student loan experiment where participants receive information critical of loan forgiveness for college students (a treatment condition) or no information (a control condition). Having the prior mock vignette allows the researchers to identify inattentive respondents prior to random assignment and then condition analyses on (pre-treatment) attention, likely isolating respondents who attend, perceive, and understand the stimulus itself.[63] The advantage here is that the check resembles the experimental manipulation and so those who pass the check likely process the experimental treatment as intended. Indeed, Kane et al. find that conditioning analyses on performance in the mock vignette leads to stronger treatment affects. The inclusion of a pre-treatment vignette need not preclude the inclusion of a post-treatment manipulation check.

Holding costs constant, an ideal study would include extensive piloting, pre-treatment checks, and post-treatment manipulation checks. Despite the crucial role of such checks, the trend, at least in political science, seems to be against using manipulation checks; based on a content analysis of political science journals (and comparing it with other content analyses), Mutz (2021) suggests that "current scholars may be *less* likely to employ manipulation checks than earlier experimentalists, even though there are more experiments in political science journals than there were in the past" (emphasis in the original). If true, it may reflect to researchers rushing to data collection without considering the elements needed for the study.

[63] This approach also avoids the use of seemingly tangential odd questions often used for attention checks that could confuse respondents.

Alas, even a well-piloted and checked treatment does not guarantee the elimination of confounds. Dafoe et al. (2018) point to the possibility of "informational equivalence" with treatments such that a given treatment operationalizes multiple concepts, only some of which are of theoretical interest. For example, treatments that vary whether a country is a democracy or not risk introducing perception not only of governance systems but also of geography, culture, demographic composition, and socio-economic standing (also see Shadish et al. 2002: 75). Another example comes from the aforementioned use of names to signal race – even putting aside the stated perceptual challenges, an individual may view a name meant to signify a Black individual (e.g., Latoya Washington) as not only revealing race but also class, and thus any impact of a treatment using that name could reflect the effect of race or class. In the ideal, an experimentalist varies only the key attribute across conditions to avoid confounds.[64]

Assessing Treatments

Since an experimental treatment constitutes a way to induce a value on the independent variable of interest, measurement validity and accuracy, as discussed in Chapter 2, become relevant. The validity question revolves around whether the manipulation alters levels of the theoretical construct (i.e., recall, measurement validity refers to the extent to which the measure reflects the abstracted concept). Ensuring validity may well involve deviating from what one

[64] That said, it may be that the relevant construct is multi-dimensional (e.g., the researcher wants to study the impact of democracies not just as institutional systems but also all the associated features). If not, the researcher should design the treatment to keep the confounding factors constant (e.g., clarifying the composition, socio-economic standings of the democracies and non-democracies) in the treatment or utilize one of the designs put forth by Dafoe et al. (2018).

may think is best for resembling the "real world" – or as having high mundane realism (i.e., the likelihood that the exact treatment will occur in the normal course of the participants' lives). This is clear in the *Reader's Digest* example from Mutz (2005).

Another example from a theory testing experiment comes from Lupia and McCubbins (1998). The authors develop a theory that absent particular external forces (e.g., penalties for perjury), persuasion can occur only if a receiver perceives a speaker to share his or her interests and perceives him or her to be knowledgeable. Other factors, such as actual common interests, likeability, or reputation are not necessary for persuasion. They design an experiment to test this, with the goal of manipulating perceptions of common interests and knowledge. They did this by having receivers predict the outcome of a coin toss (i.e., heads or tails) such that he or she made more money from correct predictions. The receiver received advice from a speaker who either had common interests with the receiver (e.g., made money when the receiver made a correct prediction) or conflicting interests with the receiver (e.g., made money when the receiver made an incorrect prediction). They also varied whether the speaker had knowledge (e.g., observed or did not observe the coin toss outcome). Lupia and McCubbins show that persuasion – the receiver believed the speaker's statement about the coin toss outcome (i.e., followed the advice) – occurs more with perceived knowledge and common interests. From the perspective of mundane realism, this looks very poor as it (i.e., predicting coin flips) does not resemble political or social persuasion scenarios. Yet, the treatment closely maps onto the theoretical concepts identified. If the authors had instead varied whether the speaker was likable or shared the receiver's partisanship, it would have lacked content validity, at the very least. That is, likeability and/or partisanship may not cover the dimensions of the construct of perceived common interests and knowledge. The authors designed the experimental treatments to operationalize independent

variables that map onto theory and ensure experimental realism (i.e., the situation is involving to the participants and they take it seriously) (Lupia and McCubbins 1998: 97-112, Bassi 2020). As explained in Chapter 2, in most circumstances, experimental realism takes precedence over mundane realism.

The other dimension of measurement concerns accuracy. Recall from Chapter 2 that accuracy refers, in part, to being unbiased such that a measure (or a treatment) does not systematically under- or overstates the true value of the construct. Bias occurs when a treatment induces participants to move too far in a given direction for reasons unrelated to the treatment itself. Consider work on inter-personal contact. A large literature explores whether intergroup contact can reduce prejudice; for example, does interacting with those from a different racial or ethnic group lead people to be more tolerant (e.g., Allport 1954)? The results are mixed (Paluck et al. 2018). Building on this idea, some explore whether having individuals "imagine" contact can reduce prejudice. The idea is that mentally simulating a positive interaction can have an effect, which, if true, would have profound potential since it does not entail overcoming systemic geographic and cultural forces that limit contact (e.g., Crisp and Turner 2009, Crisp et al. 2009). A common manipulation involves people thinking about interacting with another person (either who shares their demographic profile or not), offering details about where the imagined interaction occurs, and encouraging participants to think carefully about it and closing their eyes (e.g., Husnu and Crisp 2010, 2011). This work often finds that imaging contact with an out-group, relative to imagining something else or imaging contact with a distinct group, can significantly reduce intergroup bias (e.g., create/lead to more positive attitudes towards the other group) (Miles and Crisp 2014).

In assessing this treatment, one might worry about measurement bias or accuracy: the manipulation might lead to systematic over-statement of the positivity of the imagined contact relative to what would occur in non-experimental where the intervention may be applied.[65] This could occur because experimental participants anticipate that the experimenter has positive attitudes towards the outgroup leading to a demand effect – where participants want to please the experimenter. That is, participants imagine an interaction not as they typically would in, say, an educational setting (one domain where advocates suggest imagined contact can influence attitudes) but rather in a way that coheres with what they think the experimenter wants. There is social desirability bias in imagining contacts that over-state its positivity. Here the concern is not about measurement validity but bias due to participants trying to match the experimenter's desired views (Bigler and Hughes 2010: 132).

A similar concern arises in work that pays respondents financial incentives when asking them about political facts. The theory here is that partisans often misreport their factual beliefs so as to appear like "good partisans," such as reporting improved unemployment or inflation rates when their party controls the administration, regardless of objective reality. Some argue that an intervention to stimulate respondents to be more accurate and/or elaborative when thinking "reduce[s] partisan divergence and elicit[s] responses more informative of people's true beliefs by offering incentives to answer correctly" (Bullock et al. 2015: 526; also see Prior et al. 2015).

[65] One could also question construct validity if the construct was actual contact; however, that may be a mistake since those who study imagined contact make clear that the construct is not actual contact but the "concept of contact, mentally articulated in the form of an imagined interaction" (Miles and Crisp 2014: 3).

This work finds payment leads to less factually incorrect reporting. The question remains, however, if the financial incentive offered to the treatment group is a good proxy (measure) for elaboration (although see Jamieson and Weller 2020 on incentives and effort). An alternative possibility is that respondents want to earn money and answer questions in the way they think the study designers would like them to do so. That is, they may believe the experimenter has certain beliefs and to earn the money they need to match those beliefs. In short, the treatment generates an experimentally driven response reflecting a measurement bias in the treatment. Respondents misstate their actual beliefs – appearing more factually correct than they are – to make money since they think the experimenters have false beliefs themselves, which means these are not "people's true beliefs."[66]

My suggestions in the imagined contact and financial incentive experiments constitute nothing more than untested assertions. It is entirely possible that in both cases, there is no measurement bias in the treatments (e.g., Mummolo and Peterson 2019). Nonetheless, my point is that in assessing treatments, one needs to recall the lessons of good measurement, concerning validity and accuracy, and evaluate the treatments along those dimensions.

Summary

(1) Many assess experiments based on their mundane realism – that is, how much they resemble "the real world." If one's goal is to directly inform policy, this standard is reasonable. Otherwise, mundane realism does not constitute an important evaluative

[66] Bisgaard (2019) suggests that measuring factual knowledge itself may have insufficient content validity since other relevant dimensions include attributions of responsibility for the facts (e.g., economic situation).

criterion. Much more important are experimental realism and the construction of theoretically appropriate treatments.

   a. Experimental realism requires ensuring that participants take the study seriously (e.g., treat the situation as they would treat any other situation in life).
   b. Treatments should be constructed to operationalize the relevant theoretical construct, which is orthogonal to the question of mundane realism (unless the goal is to test a policy intervention).

(2) Pilot testing and/or manipulation checks are essential to ensure experimental participants perceive the treatment(s) as intended.

(3) Consider treatment confounds and whether they create problems for the causal inference under study.

(4) Assess treatments as a form of measurement – considering their measurement validity and accuracy (bias).

Validity

Few topics garner as much discussion when it comes to experiments as "validity," which, in this context, is distinct from measurement validity (as discussed in the prior section and in Chapter 2). In their classic text, Shadish et al. (2002: 38) distinguish four types of validity, summarized in Table 3-1: 1) statistical conclusion validity concerns the confidence one has in the covariation between the treatment and outcome, 2) internal validity concerns the confidence one has in concluding a causal relationship between the treatment and outcome, 3) construct validity concerns the confidence one has in the inferences about the constructs of interest (e.g., measurement validity, the setting reflects the focal setting or population of interest), and 4) external validity concerns the confidence one has that the cause-effect relationship "holds over

variation in persons, settings, treatment variables, and measurement variables."[67] Put another way, external validity refers to generalizability (Mutz 2011: 133). With construct and external validity, another dimension sometimes includes a finding holding over distinct time periods, although time can also be enveloped in the "different settings" category of external validity (Shadish et al. 2002: 20, 70; Cook and Campbell 1979).[68]

Table 3-1: Types of Validity

Type	**Definition**
Statistical conclusion validity	The confidence one has in the covariation between the treatment and outcome.
Internal validity	The confidence one has in concluding a causal relationship between the treatment and outcome.
Construct validity	The confidence one has in the inferences about the constructs of interest.
External validity	The confidence one has that the cause-effect relationship holds over variation in persons, settings, treatment variables, and measurement variables.

When it comes to statistical conclusion and internal validity, the key issues revolve around meeting the assumptions of causal inference previously discussed, in light of whichever approach one takes. Construct validity partially, although not entirely, involves questions of measurement and treatment design. In these senses, I have already touched on the topics; moreover, extended discussion of these types of validity are available in Shadish et al. (2002). Gerber and Green (2012) also offer superb guidance on ways to address statistical conclusion

[67] This concept is distinct from "ecological validity" which is akin to mundane realism.

[68] On timing, one particular salient consideration is pre-treatment effects – that is, whether experimental respondents had been exposed to information analogous to the treatments prior to the experiment (Gaines et al. 2007, Druckman and Leeper 2012b, Slothuus 2016).

and internal validity problems that arise from violating assumptions underlying the random assignment statistical approach to causal inference.

I focus here on external validity as it receives a disproportionate amount of attention in discussions of experiments. This attention imbalance likely stems from three misperceptions. First, many assume experiments possess high statistical conclusion and internal validity due to the intervention; yet, as explained, assumptions still must be met and experiments inherently do not guarantee high internal validity. Second, many perceive a finite tradeoff such that increased internal validity means lower external validity, when in fact addressing one type of validity does not invariably affect other types (see Shadish et al. 2002: 96-102). Third, there exists a widespread concern that experiments in the social sciences have low external validity; as McDermott (2011: 34) states, "political scientists tend to focus, almost exclusively, on problems associated with external validity" (also see Mutz 2011: 12). In what follows, I provide a constructive approach to thinking about external validity – by highlighting four key questions one should ask when generalizing – to reduce the reflexive presumption that an experiment has low external validity.

The first question one must ask when assessing external validity is "what is being generalized?" This ties back to my earlier discussion about experiments as an approach for generalizing casual relationships. It could be that one aims to generalize the existence of an experimental treatment effect (be it via the randomized intervention or controlled variation), *or* the precise size of a treatment effect.[69] The answer depends on the goal of the experiment. When

[69] Egami and Hartman (2020) draw a similar distinction in their formal framework where they put forth testing procedures for exploring "effect-generalization – generalizing the magnitude of

the underlying goal involves "whispering in the ears of princes" – i.e., speaking directly to policy interventions – then the size of the impact matters, since those making policy (e.g., government officials, leaders of non-governmental organizations) may use it to assess the costs and benefits of policy programs. This introduces a host of challenges coined as "threats to scalability" by Al-Ubaydli et al. (2017). These authors (282) explain that when policy programs are based on experimental findings, "the program (treatment) effects diminish substantially in size…" This occurs due to experimental findings being false positives, the population from which the experiment(s) drew the sample being amenable to larger effects (e.g., informed consent is obtained only from those likely to be affected, superior compliance in experimental studies than in policy programs), and loss of control once one leaves the experimental study setting (e.g., the treatment administration is no longer overseen by the researcher). The authors argue for more explicit consideration of scalability problems, pushing researchers to flip "the traditional model, calling on scholars to place themselves in the shoes of the people whom they are trying to influence and produce more policy based evidence. Our call is for policy research that starts by imagining what a successful intervention would look like fully implemented in the field, applied to the entire subject population, sustained over a long period of time, and working as it is expected because its mechanism is understood" (Al-Ubaydli et al. 2020b: 21).

When an experimentalist's goal involves searching for facts or testing theories, the precise effect size matters much less. Here, one looks for evidence consistent with or contradictory to a hypothesis or pattern. External validity questions concern whether the causal

---

causal effects [and] sign-generalization – assessing whether the direction of causal effects is generalizable" (1).

relationship itself generalizes, regardless of size. For social scientists, this is the case for many experiments. External validity concerns then come down to assessing whether features of the experimental sample, context, treatment, and measure preclude generalizing the relationship beyond the single study.[70]

In addition to assessing whether a given relationship generalizes beyond a study, one needs to ask an often ignored question: "to what is one generalizing?" As emphasized in Chapter 2's discussion of sampling, one must state to what population, settings, and causal variables one wants to speak (e.g., Tipton 2019 et al.).[71] With these statements in place, one needs to systematically assess whether there exist features of the experiment that would undermine the existence of a finding in the targets of generalization. To see this in practice, consider sample

[70] This is not to say effect sizes are completely irrelevant without a policy goal. As discussed, expected effect sizes play a role in identifying the appropriate sample size with regard to statistical power. In Chapter 5, I also explain that while a single effect size from a given study should be taken with great caution, the aggregation of effects across studies play a vital role in the accumulation of knowledge – my hesitancy here concerns generalizing an effect size from a given study.

[71] Westreich et al. (2019: 439) explain that "in few, if any, studies – randomized trials or otherwise – do authors report the target population for their causal effect." They introduce the concept of "target validity," which refers to "the total difference between the true causal effect in the target population and the estimated causal effect in the study sample" (438). This metric is useful insofar as it accentuates the importance of clearly defining the population.

generalization. For many experiments, at least in the field of American political behavior, the presumed population includes all U.S. citizens (or permanent residents) of voting age. An experiment may be done on a non-representative convenience sample and the issue then is whether a causal inference found in the experiment would hold in the population. The only reason it would not hold is if the causal relationship depends on individual attributes. Or put another way, if no individual level characteristics moderate the treatment effect, the same causal dynamic holds for all individuals in the population and thus what one finds in any sample from the population would be the same (putting aside measurement error, etc.).

To see this point in action, consider one of the most notable findings in the social sciences about the impact of conformity – people follow descriptive norms by adapting their attitudes when the majority of those around them hold particular opinions (e.g., Sherif and Sherif 1953, Asch 1956, Dwyer et al., 2016, Davis et al. 2018). Scholars invoke this finding to explain voting behavior (e.g., Gerber et al. 2008, Sinclair 2012), the formation of policy beliefs (e.g., Kahan 2017), and conservation behaviors (Bayes et al. 2020), *inter alia*. Yet, much of the direct causal (i.e., experimental) tests of conformity have relied on student samples, raising the question of whether the dynamic generalizes to non-student populations. It may not, because younger individuals have more malleable attitudes, meaning they conform more easily (Bond and Smith 1996). Such a rationale is plausible *and necessary* for one to question whether the finding generalizes. That is, one must state a reason why the experimental effect may not hold in the target population; the onus falls on the critic – as a Popperian perspective makes clear, nothing can be proven and thus an argument and evidence should be accepted until shown otherwise.

Yeager et al. (2019) replicate the classic conformity experiment (randomly assigning participants to a condition that said a majority of Americans support a public policy or a majority opposed it) on a probability sample of Americans.[72] Yeager et al. replicate the conformity effect such that those who learn a majority of Americans support a policy are more supportive than those who are told a majority opposed it; this finding reveals the generalizability of the conformity effect across samples. More interestingly, they also report that the effect size in their probability sample is significantly smaller than that found in the student samples, and the smaller size comes from conformity having a larger impact on respondents who resemble students in terms of age, education, income, race, gender, and region. The Yeager et al. results highlight: 1) the importance of distinguishing whether one wants to generalize the existence of a causal effect or its size, 2) when generalizing the existence (which I have suggested is the norm for much of social science), one must identify sample features that would undermine generalization to a specified target population, and 3) generalizations of well-developed and previously-tested causal relationships often hold across samples (Yeager et al. replicate 6/7 studies; also see Klein et al. 2014, Mullinix et al. 2015, Coppock 2019b).

Assessing external validity thus entails specifying what one is attempting to generalize and *to* what one is attempting to generalize. This means experimentalists need to carefully state their targets of generalizations but also that critics who cite low external validity need to specify what aspect of the study limits its generalizability. Much of the discussion – including the example from Yeager et al. – focuses on the generalizability of the units or sample. A third point

[72] They also explore six other classic psychological experiments that had been conducted mostly with student samples.

is that, as mentioned, external validity means generalizing across 1) samples, 2) settings, 3) treatments, and 4) outcome measures (Shadish et al. 2002: 83; see Egami and Hartman 2020 for a formalization that incorporates these four dimensions into the potential outcomes framework).[73] This often-missed point reveals a double-edged sword for experimentalists. On the one hand, scholars rarely explicitly consider how a study might generalize to distinct settings, treatments, and outcome measures – with some notable exceptions being explicit meta-experiments that look at contextual effects (e.g., Dunning et al. 2019, Blair and McClendon 2021).[74] On the other hand, the same holds for most non-experimental empirical work.

While scholars sometimes obsess about the random selection from a population of units and sometimes talk about settings of the experiment (e.g., laboratory, field, survey), they rarely discuss the representativeness of the larger context, treatments, and outcome measures.[75] As

[73] These four dimensions appear to differ from what I emphasized in my discussion of sampling in Chapter 2, where I pointed to sampling of units, contexts, topics, and measures. However, "settings" is synonymous with "context." Measures can be broken out into treatments and outcome measures, and "topics" are enveloped in treatments. I opted for the distinct characterization earlier since those are the elements substantively most relevant when thinking about sampling (i.e., units, contexts, topics, measures), but here I stay consistent with Shadish's (2002) classic treatment.

[74] For an example of generalizing over time, see Twenge et al. (2008).

[75] Shadish et al. (2002: 23-24) state, "[r]andom selection occurs even more rarely with treatments, outcomes, and settings [and timing] than with people… sampling methods of any kind are insufficient."

Mutz (2011: 131-132) emphasizes, the "tendency to make assertions about the generalizability of research results based solely on either the *setting* in which the study was done (i.e., laboratory versus field) or the representativeness of the *subjects* studied reflects a lack of considered thought on [the] much broader and ultimately more interesting subject… methodology alone provides little to no guidance when judging external validity" (emphasis in original). Mutz's point makes clear that assessments of generalizability involve more than a mundane realism standard or the sample, but also include the setting and the operationalizations, and how well they match the target context, outcome measures, and treatments.

One telling study is Landy et al.'s (2020) crowdsourcing initiative where more than a dozen research teams independently created stimuli to test the same research questions (on topics including moral judgment, negotiation, and implicit cognition). For instance, the implicit cognition study built on initial research measuring awareness (self-reports) of automatic prejudice with questions like: "Although I don't necessarily agree with them, I sometimes have prejudiced feelings (like gut reactions or spontaneous thoughts) that I don't feel I can prevent." Another team operationalized the construct with questions such as "Regardless of my explicit (i.e. conscious) beliefs about social equality, I believe I possess automatic (i.e., unconscious) negative associations towards members of stigmatized social groups." Yet, a distinct team took the approach of first offering an introduction about automatic triggered associations. Then they had respondents report their "first automatic reaction when you think about" various distinct groups (e.g., African-Americans, Latin Americans, Gay people, etc.) on a scale ranging from more negative automatic associations to more positive automatic associations. The study overall finds that across the five hypotheses tested there is substantial variation in results based on the sets of materials (i.e., very different results from distinct research teams tasked with testing the

same hypothesis due to the use of variations in operationalizations). This highlights that generalization across operationalization is far from a foregone conclusion.[76]

This discussion makes clear that once one starts to consider tests over all these dimensions (i.e., samples, settings, treatments, and outcomes measures) "a strong case can be made that external validity is enhanced more by many heterogeneous small experiments than by one or two large experiments" (Campbell and Cook 1979: 80). Experimentalists must consider all of these dimensions necessitating the need for multiple studies to capture them all.[77]

Fourth, just as causal inferences require points of comparison, so should statements of external validity; put another way, they should be precise. If one assesses the generalization of a sample to a larger population, then the relevant question is: holding all other aspects of the

[76] Interestingly, they also find that scholars were able to successfully forecast which of the hypotheses being tested were more likely to have consistent results (suggesting strong a priori expectations of hypotheses robust to variations in operationalizations).

[77] One also needs to be precise about the causal inference being generalized. For example, some critique experimental studies that show exposure to a given media story changes attitudes on the issue covered in the story. They argue that in other settings where people can choose the media to which they are exposed, the exposure effects diminish or disappear (e.g., Barabas and Jerit 2010, Arceneaux and Johnson 2013). The problem here is that the initial media studies were trying to generalize a relationship that assumed exposure (Mutz 2011:150-151), whereas the critic involves a causal relationship between media choice environments and attitudes. This is valuable and reveals a contextual limitation of the initial studies, but it does not undermine the value of those studies since the casual relationships being studied is distinct.

experiment constant, would the experiment have higher external validity if carried out on a different sample (and what kind of different sample)? As explained, a gain in external validity may occur less often than presumed, but, regardless, holistic statements that an experiment "has low external validity" not only provide little constructive guidance, but also unknowingly employ an illusionary idealistic study that randomly samples across units, settings, treatments, and outcomes. A more useful approach entails evaluating each dimension of external validity and considering how a different study may have increased generalizability relative to the extant study. Every empirical study is confined in terms of generalization and assessments need to be precise across dimensions of validity and points of comparison.[78] In Table 3-2, I offer a summary of the four key questions one should ask when thinking of generalizing.

Table 3-2: Generalization Questions

Generalization (External Validity) Questions	**Details**
What is being generalized?	Generalizing the existence of an experimental treatment effect, *or* the precise size of a treatment effect.
To what is one generalizing?	State to what population, settings, and causal variables one wants to generalize.
Which dimensions of generalization are most important or problematic?	One could aim to generalize across samples, settings, treatments, and outcome measures.
What is the counterfactual / comparison point for a generalization statement?	Evaluate each dimension of external validity and consider how a different precise study may have increased generalizability relative to the extant study.

[78] This is similar to the idea of "parallelism" in experimental economics – "it should be *presumed* that results carry over to the world outside the laboratory. An honest skeptic then has the burden of stating what is different about the outside world that might change results observed in the laboratory. Usually new experiments can be designed and conducted to test the skeptic's statement" (Friedman and Sunder 1994: 16; italics in original).

This brings me back to falsification. Every empirical test of a causal proposition, if done carefully, accumulates evidence and leads to knowledge progression – either by offering evidence consistent with a hypothesis or not, in which case one refines the theory (or starts over).[79] No study should be dismissed due to vague statements about external validity or mundane realism – every test helps and we need "*many* tests to determine whether a causal proposition has or has not withstood falsification" (Cook and Campbell 1979: 31; italics in original). This accentuates a corollary point that assessments of external validity ideally occur over a range of studies on a single topic (Mutz 2011: 135). The validity of any single study, regardless of the nature of its participants, context, and operationalizations, should be considered in light of the larger research agenda to which it hopes to contribute.

In sum, discussions of external validity need to consider: 1) the goal of the research, 2) the targets of generalizations and sources of heterogeneity, 3) multiple dimensions of generalizability, 4) the comparison points, and 5) the larger research agenda. External validity has long been the bane of experimental studies, and this is for good reason at times. Yet, when one critiques the external validity of an experiment, the onus falls on that person to specify why the causal relationship (or lack thereof) would not generalize to a target population, context, treatments, and outcome measures. On the flip side, experimentalists must take these considerations into account when they attempt to generalize, which means clearly stating the

[79] Causal hypotheses are never confirmed and evidence accumulates via multiple tests, even if all of these tests have limitations. Campbell (1969: 361) states, "…had we achieved one, there would be no need to apologize for a successful psychology of college sophomores, or even of Northwestern University coeds, or of Wistar staring white rats."

goal of the study, the target population/context/treatments/measures, and why the causal relationship or size thereof would generalize. More precise and sustained discussion about external validity will lead to improved experiments and, ultimately, intellectual progress.

Summary

(1) Most take statistical conclusion and internal validity for granted in experiments but one must assess the causal assumptions underlying claims from experiments, as enunciated in Chapter 2.

(2) Experiments do not inherently have lower external validity then other methods.

(3) The assessment of external validity requires answering the following questions.

   a. What is being generalized? In many cases, it may be the existence of a causal relationship whereas, in applied studies, the size of the relationship may be important.

   b. To what is one generalizing? This requires specifying the precise population of interest, as well as the target setting, treatments, and outcome measures. With regard to the sample, generalization depends partially on the existence of heterogeneous effect sizes.

   c. Which dimensions of external validity are most relevant/problematic? External validity entails generalizing across samples, settings, treatments, and outcome measures (and timing). Most experiments have no variation in the latter three dimensions, necessitating the need for multiple studies. A promising path involves systematically introducing variation in the three dimensions (drawing on approaches in the case study literature; see, e.g. Gerring 2001).

d. In assessing the generalizability/external validity of an experiment, what is the comparison point? Scholars need to be precise about the "counterfactual" study to which one compares an experiment.

(4) External validity should not be a reason to dismiss an empirical result; all empirical evaluations of falsifiable hypotheses, if soundly conducted, contribute evidence that accumulates over time.

Experimental Samples

In the prior section, I discussed the some of the conditions that determine whether a given experimental sample of units is problematic for generalizing a causal interference. Here I explicate with more detail; I do so because when it comes to critiques of the external validity of experiments, the central culprit has long been the sample of units. This stems from the prevalence of student samples used in experiments throughout the social sciences and concomitant suspicion of such samples (Sears 1986). Writing a little more than a decade ago, Kam et al. (2007: 421) state there exists a "simplistic heuristic of 'a student sample lacks external generalizability.'" Similarly, Gerber and Green (2008: 358) write, if "one seeks to understand how the general public responds… the external validity of lab studies of undergraduates has inspired skepticism" (also see McDermott 2002, 2011 for discussion).[80]

[80] An intriguing issue that has received relatively less attention concerns the comparability of student samples from different institutions. Lupton (2019) studies this issue, finding demographic and attitudinal differences across student samples from different schools but largely consistent treatment effects. She also finds variation in participant dropout rates depending on whether the school has a dedicated research pool.

Much changed in the ensuing decade thanks to the rise of online data sources that facilitate, in particular, the implementation of survey experiments. These samples can be roughly categorized into one of three types. First are probability internet panels where a company draws or recruits a probability sample (i.e., every unit has a known and equal chance of being sampled) of the population (e.g., a country), and individuals then agree to participate in periodic surveys for compensation. While coverage and non-response remain challenges, these samples constitute the gold standard; however, they come at a high financial cost. Second are non-probability but purportedly representative internet panel samples. In these cases, individuals opt-in (e.g., via advertisements) to receive compensation for taking surveys over time. The companies that oversee these panels then, when hired by a researcher, use a version of quota sampling to draw a set of respondents that matches a specified population on key demographics – such as having percentages of women and minorities that equal those found in the U.S. census. These samples vary in their ability to hit benchmarks, but some have been demonstrated to do quite well (e.g., Vavreck and Rivers 2008).[81] These data vary in cost but almost always come cheaper than probability samples.

Third, the most notable innovation comes from crowdsourcing labor market platforms, with the best-known being Amazon's Mechanical Turk (MTurk). Researchers can directly use MTurk or analogous platforms to hire individuals to complete tasks, including taking survey experiments for direct compensation. These platforms provide researchers with cheap convenience samples that ostensibly offer more heterogeneity than student samples. A

[81] One challenge for many of these samples, strictly speaking, is they often have poor joint distribution issues, such as lacking a sizeable percentage of low-income minorities.

substantial literature assesses the demographic nature of these crowdsourcing platforms and whether experimental results from them match those from other samples (Paolacci et al. 2010, Buhrmester et al. 2011, Berinsky et al. 2012, Huff and Tingley, 2015, Krupnikov and Levine 2014, Mullinix et al. 2015, Levay et al. 2016, Coppock 2019b). These samples also bring with them ethical concerns in terms survey participants not receiving a fair wage (e.g., Williamson 2016). Krupnikov et al. (2021) offer an overview of this work, showing the rise in these samples, as well as offering advice on how to best use student, community, crowdsourcing, and other types of convenience samples.

Technological innovation provides experimentalists with access to other types of samples as well. For instance, while experiments have long been done with elected officials – sometimes intentionally (e.g., Grose and Wood 2020) and sometimes naturally (e.g., Cirone and Van Coppenolle 2018) – the ability to contact them via e-mail has stimulated a relative explosion in "elite" experiments (e.g., those who work in government; see Neblo et al. 2018, Grose 2014, 2021, Nathan and White 2021). Of particular note are audit experiments, which I discuss in the next chapter (Costa 2017). The expansion of social media also provides experimentalists with a data source, sometimes to recruit subjects (e.g., Munger et al. 2019), and other times to experimentally study social media behavior (e.g., Bond et al. 2012, Guess 2021).

While all of these sampling opportunities (e.g., internet panels, crowdsourcing platforms, social media samples) allay some of the concerns about student samples, many still question: do

findings from a given sample generalize (e.g., have external validity)?[82] Ironically, as I next explain, the answer should be thought of in terms of theory rather than empirics.

Three Sampling Scenarios

In essence, there exist three scenarios – as formalized in Druckman and Kam (2011) – that could be the "truth" in a given target population. First, it may be that the impact of the treatment on the outcome variable (e.g., the impact of a persuasive message on attitudes, the effect of electoral observers on fraud, the influence of non-anonymity in allocation of funds to co-ethnics, etc.) is homogenous, meaning the treatment has the same impact on all individuals in the population. For instance, a persuasive message moves policy attitudes by 10% among everyone, regardless of their gender, income, race, partisanship, etc. The implication is that if one tests the treatment only on young people, the treatment would push attitudes by 10%, or if one tests it only on older people, one would also find a 10% movement. The same is true if the sample only consists of men or only included women, only low-income or only high-income, only Democrats or only Republicans, etc. Since the persuasive message treatment moves all segments of the population the same amount, the nature of the sample becomes irrelevant. In short, with such homogeneity, *any* sample from the population will, with the typical uncertainty, provide an unbiased estimate of the causal treatment effect (and the same estimate will come

[82] The trend in using these new sampling opportunities can be seen by counting the number of experimental articles in the *APSR* (as displayed in Figure 1-1) that employ such "non-traditional samples" (i.e., non-student convenience samples, social media, or elites). The use of these samples jumped by 6 percentage points in articles from 2010 to 2019, relative to 2000 to 2009 (52% to 58%).

from any sample). There is no threat to external validity from any sample even when it comes to the precise size of the effect.

A second sampling scenario is that the treatment effect in the population is heterogeneous such that individual characteristics moderate the effect. For example, a persuasive message about Medicare may differ across segments of the population; it may be that the effect is larger among older people and smaller among younger people. If one uses a sample that skews towards older (younger) people then the treatment effect would be wrong – i.e., it would be over-stated (under-stated) since the effect is larger (smaller) among older people. However, if the skewed sample has at least some mix of older and younger people (i.e., some variance), then one can in essence control for age by interacting it with the treatment variable. This is the technique discussed in Chapter 2 for modeling heterogeneity by interacting the treatment with the moderator (e.g., exposure to message * age). Doing so would result in an unbiased estimate (with typical uncertainty) of the size of the average treatment effect, even if the sample comes from a crowdsourcing platform, community volunteers, or an opt-in panel. Druckman and Kam (2011) use simulations to show that as long as there is just some variance in the moderator (e.g., age), the estimate will be on-target.

Here, an experimentalist controls for the sample bias that would skew a treatment effect by interacting the source of the bias with the treatment (with one caveat discussed below). Even if one uses a perfectly representative sample but does not control for age via an interaction, the treatment effect will be misleading as it may look moderate when in fact it is large for older people but small for younger people; the heterogeneity would be masked. This would be akin to having an omitted variable.

It follows that, in most cases, the external validity or generalizability of a sample depends not on the composition of the sample but rather on the strength of one's theory underlying the causal effect and then modeling the theory (i.e., the moderator) correctly. Scholars spend much time empirically comparing demographic and political benchmarks of samples when the issue ultimately concerns having stronger theory about moderators, and then constructing appropriate samples when needed.

The third sampling scenario provides a caveat (e.g., the "in most cases" phrase in the prior paragraph). A non-representative (non-probability) sample becomes problematic if it lacks variance in the moderator since in that case one cannot model the heterogeneous effect. A student sample, for example, would be problematic in the Medicare example given no variance on age to control for the sample skew.[83] That said, even in that case, the key is to develop a theory behind the causal effect, state the population of interest, and choose an appropriate sample. The example accentuates a common point of confusion – when it comes to estimating the experimental causal effect, it is not essential that the sample be representative of age *per se* (i.e., the goal is to not describe the demographics of the population), but rather that there exists

[83] This also is the case if one anticipates differential treatment reactions across cultures but collects data from only one culture. This is a particular concern given that many studies only collect data from Western, Educated, Industrialized, Rich, and Democratic (WEIRD) societies (Henrich et al. 2010).

variance on age.[84] The demographics of experimental samples only matter when it comes to moderators, given the goal to generalize the causal effect. While this means the researcher needs to carefully develop the *theory* behind an experiment, it also suggests that those who criticize the generalizability/external validity of a sample must do more than point to its lack of representativeness of the population; they must explain what moderator was ignored and why the given sample was inadequate (i.e., lacked variance) to test for heterogeneous effects.[85] A sample is only problematic for external validity if someone theorizes a moderator and there exists almost no variance on that moderator in the sample.

One could argue that much of social science theory is not suitably developed to identify key moderators, given a common focus on general tendencies among populations and data that lack sufficient power. My response is twofold. First, if / when this is the case, it highlights the importance of the inevitable conversation between theory and exploratory data analyses to isolate moderators with the latter exploiting methodological innovations I discuss in the next section. Second, at best, it remains debatable where the burden lies – my position is that it lies on the critic to explain why a particular sample may be problematic in terms of moderator dynamics as the alternative allows anyone to dismiss a sample with virtually no explanation.

---

[84] As Coppock et al. (2018: 12) states, "simply noting that convenience and probability samples differ in terms of their background characteristics is not sufficient for dismissing the results of experiments conducted on convenience samples."

[85] Druckman and Kam (2011: 50) state that an "implication is that the burden, to some extent, falls on an experiment's critic to identify the moderating factor and demonstrate it lacks variance in an experiment's sample."

Detecting Heterogeneity

Putting aside the theoretical emphasis of the prior section, one can ask how often heterogeneous effects empirically exist in experiments on a particular topic.[86] Mullnix et al. (2015) offer tentative evidence by comparing treatment effects between convenience samples and probability samples for 23 survey experiments, finding a remarkable rate of not just replication but also matched effect sizes for 18 of the studies.[87] This suggests no heterogeneity, or at least heterogeneity that does not correlate with sample composition. Coppock et al. (2018) provide a more direct test. They compare the sample average treatment effects of 27 experiments conducted on probability samples against the same experiments on MTurk. They report vast similarities that stem from homogenous effects in the samples.[88] Specifically, they explore six variables (age, education, gender, ideology, partisanship, and race) and find none systematically

[86] Another empirical question concerns whether different types of commonly used samples contain sufficient variance on most variables relevant to social inquiry. Druckman and Kam (2011) investigate this question with regard to college student samples and find that in fact, other than age, student samples do very well not only in terms of variance but even averages.

[87] That said, they find that survey experiments embedded in election exit polls tend to have distinct dynamics, perhaps suggesting something about that context or sample composition.

[88] An alternative possibility that they rule out is that the apparent effect homogeneity stems from similar heterogeneous treatment effects within samples that aggregate similarly (i.e., heterogeneous effects that do not correlate with sample composition).

moderate the effects.[89] These experiments appear to fall into case 1 above: the nature of the sample used is irrelevant to the estimation since the causal effects are ostensibly constant in the population (also see Coppock 2019b).

These striking results have implications not only for sample usage and generalizability but also for understanding opinion formation (also see Coppock et al. 2020). Yet, as Coppock et al. (2018: 12445) caution, one must not over-interpret the breadth of homogeneity for four reasons. First, their studies only include survey experiments, and thus the results do not generalize to studies conducted in laboratory or field settings; and, as a general matter, survey experiments are limited to brief interventions over short periods of time (Sniderman 2018). Second, the study topics mostly include persuasion and attitude formation (Coppock et al. 2018: 12442), thereby offering no insight into heterogeneous effects that may exist in other domains. For example, as discussed in Chapter 2, Yeager et al. (2019) report moderation across a set of psychology experiments (e.g., on conformity, decision-making) by variables similar to those examined by Coppock et al. (e.g., age, education, income, race, gender, and region). Alternatively, Green and Kern (2012) study heterogeneous reactions to a well-known question-wording experiment where respondents exhibit significantly greater support for "assistance to the poor" as opposed to "welfare." The authors find stronger treatment effects among Republicans, conservatives, younger people, and those with negative racial attitudes towards Blacks.

Third, there may be unexamined moderators; indeed, other work on persuasion and attitude formation suggests heterogeneity stems from political knowledge (Druckman and Nelson

[89] Put another way, the conditional average treatment effects are tightly clustered around the overall average treatment effect in each study version.

2003), processing style/engagement (Cacioppo et al. 1983), attitude strength (e.g., Taber and Lodge 2006), or other psychological constructs. As Coppock (2019b: 624) aptly states, "[s]ome treatments of course have different effects for different subgroups… Future disagreements about whether a convenience sample can serve as a useful database from which to draw general inferences should be adjudicated on the basis of rival theories concerning treatment effect heterogeneity…." This points again to theory as a basis to develop expectations about heterogeneity and assess the appropriateness of different samples. That said, I would be remiss if not to acknowledge, as discussed in Chapter 2, the availability of machine learning approaches to identify moderators in a post hoc manner. These methods offer crucial opportunities to systematically isolate sources of heterogeneity that can stimulate theory development (see Ratkovic 2021) and, as mentioned, play a key role in the theory-empirical analyses dialogue needed to develop clear hypotheses. This is the approach taken by Green and Kern, whose findings offer guidance to theorize about moderators and appropriate samples when it comes to framing effects on welfare/poverty.[90] The usefulness of this approach for theory building though does not remove the onus on those who critique particular samples for lacking external validity from identifying reasons why they expect the sample to be problematic.

[90] These approaches also can offer crucial applied insights; for example, Künzela et al. (2019) identify a subgroup (e.g., people who voted infrequently in recent elections) on whom voter mobilization mailers may have a backlash; this finding is important for those looking to mobilize voters. As the authors (4163) explain, "if the number of mailers is limited, one should target potential voters who voted three times during the past five elections, since this group has the highest ATE [average treatment effect] and it is a very big group of potential voters."

A final point on heterogeneity comes back to a theme of this chapter – the need to consider not just sample validity but also variation in contexts, treatments, and outcome measures. Coppock et al. (2018: 12445) explain, "this discussion of generalizability has been focused exclusively on who the subjects (or units) of the experiments are and how their responses to treatment may or may not vary… [there are] four dimensions of external validity: units, treatments, outcomes, and setting. In our study, we hold treatments, outcomes, and setting constant by design." Identifying heterogeneity on these other dimensions likely requires multi-study efforts. For example, Van Bavel et al. (2016a) show that across 100 experiments, contextual sensitivity negatively correlates with the success of a replication attempt – context-sensitive studies were less likely to replicate when conducted at a later time in a distinct setting. In this case, the authors (6455) define context based on "time (e.g., pre- vs. post-Recession), culture (e.g., individualistic vs. collectivistic culture), location (e.g., rural vs. urban setting), or population (e.g., a racially diverse population vs. a predominantly White population)" (also see Van Bavel et al. 2016b). The next step entails identifying what precisely about the contexts moderates the success of replication.[91]

Blair and McClendon (2021) offer a comprehensive framework for designing research programs – that can be conducted simultaneously or sequentially – to identify heterogeneity across contexts. This process often involves overcoming the challenge of coordinating teams of

[91] An example of a timing effect comes from the aforementioned Green and Kern (2012: 505) study; they report stronger treatment effects in their aid-for-poor versus welfare experiment during the years of the Clinton administration when welfare policy was hotly debated. Also, some crowdsourcing data seem sensitive to the time of week of the study (Arechar et al. 2017).

scholars conducting similar studies, but the potential knowledge gains are substantial. An example of such an effort comes from the Metaketa Initiative (https://egap.org/metaketa). The project facilitates collaborations of teams of researchers working on related questions and using comparable interventions and measures across contexts. Researchers conduct their individual studies in analogous manners with a goal of formally synthesizing the findings from all settings (e.g., six countries), which allows for the isolation of contextual heterogeneity (e.g., Dunning et al. 2019). The project explores thus far political accountability, natural resource governance, community policing, and women's non-electoral political participation. Of course, coordinating institutions do not exist in many research areas – but even without such efforts, individual scholars should remain attuned to the potential for heterogeneity from the sample, context, treatments, and outcome measures. As Goroff et al. (2019, 1) state, it "is time to address this context sensitivity problem in social science research. While we do not yet know *how* to solve it, we believe social scientists can make great progress by working together to build an inference engine" (italics in the original).[92]

None of these points minimize the importance of Coppock et al.'s homogeneity findings. Those results provide a stimulant, if not call, for more systematic theoretical exploration into heterogeneity of all types and consideration of implications for sample selection and external validity. The central lesson remains that the external validity of samples comes down to theory more than empirics and constitutes just one of many dimensions of generalizability.

Weighting

[92] The authors propose some intriguing meta-institutions that would extract relevant contextual information from past and ongoing studies.

The discussion of sampling thus far has ignored the question of whether one should weight experimental sample data. Weighting requires that one obtain descriptive data of the target population, typically demographics (Bethlehem and Callegaro 2014). For example, when the population includes all Americans, one can use the U.S. Census or American Community Survey for demographic population figures. One then computes weights that account for each respondent's probability of being included in the sample. For example, if the population consists of 50% men but the sample contains only 40% men, then male sample respondents will be weighted to count more in computations from the sample (and women will be counted less).[93] Survey researchers commonly use weights, even with many probability samples, to ensure the accuracy of observational inferences (e.g., the percentage of men who hold a particular attitude).

A downside of weighting is that it reduces statistical power. The relevant N becomes the "effective sample size": that is, the sample size one would have been left with if he/she created a representative sample with the given data (sans weighing). This invariably smaller N leads to larger standard errors and hence a loss of power (Piazza 2010: 97). The loss in power may partially explain why experimentalists have neglected the topic (Franco et al. 2017: 161).

[93] The process often involves iteratively adjusting for multiple demographics. The idea is to assign an adjustment weight to each respondent such that those from under-represented groups receive weights greater than 1 and those from over-represented groups receive weights less than 1. There are a host of ways to specifically weight (DeBell 2018: 520), and caps may be placed on how much a given observation can be weighted so that one respondent does not overly determine the outcomes.

How should experimentalists think about weighting? Recall the goal is to generalize a causal inference rather than a description of the population. Thus, the underrepresentation of a given group (e.g., men) may be entirely irrelevant unless that group affects/moderates the casual inference. Weighting becomes useful only if there exist heterogeneous treatment effects and the source of that heterogeneity (the moderator) would be weighted (i.e., the group is misrepresented in sample).[94] For example, weighting by sex only affects the results if men react differently to the treatment than women. Miratrix et al. (2018) investigate the impact of sample weighting with 7 survey experiments on non-probability samples. They (275) find that the "[s]ample average treatment effect estimates did not appear to differ substantially from their weighted counterparts, and they avoided the substantial loss of statistical power." In their data, heterogeneous effects do not create problems for using non-probability samples sans weights (286-287). This does not mean, however, that researchers should always avoid using sample weights. They can be essential in cases where researchers predict heterogeneity or plan to engage in exploratory work to identify moderators. One can compare weighted and unweighted samples and any difference may suggest a moderator that a researcher can then isolate (see Miratrix et al. 2018: 290).[95] That

[94] Miratrix et al. (2018: 289) explain that the population average treatment effect "can only be different from the [sample average treatment effect] when two things hold: (1) there is meaningful variation in the treatment impact, and (2) that variation is correlated with the weights."

[95] I have not touched on weighting with convenience samples such as online labor markets. Here weighting is difficult since the sampling process is unknown (Miratrix et al. 2018: 276). Another challenge is that the key moderators may be non-demographic variables (e.g., partisanship, racial

said, experimentalists only need to weight if there exists a clear expectation of heterogeneity based on an underrepresented group. Weighting then becomes a challenge (at times) since it requires relevant measures of the target population (see Hartman 2021).

The Need for Probability Samples

I have offered a defense against those who dismiss experimental work based on sample composition. First, such critiques misconstrue how one should view experiments with regard to intellectual progress – from a Popperian perspective, any test can provide useful information and consequently outright dismissal based on a given sample flies in the face of how many research agendas proceed. Second, experiments aim to generalize causal inferences and thus attending to demographic or other benchmarks of a given sample misrepresents or misconstrues the endeavor. The focus should be on theorizing, particularly about heterogeneous effects, rather than critiquing the nature of a given sample. Moreover, much of the evidence to date suggests homogeneity at least in some domains of experimental social science. A partial caveat here concerns applied experiments where isolating precise effects sizes is essential (and the theory is not sufficiently developed to identify sources of heterogeneity). Third, the sample only constitutes one dimension of external validity – one that facilitates criticism of experiments since any single study, regardless of method, typically fails to generalize across the other dimensions.

---

attitudes). It is difficult to weight on such factors since there are not clear population data available. One approach, however, is to use established probability surveys such as the General Social Survey or American National Election Study to obtain population estimates of different dispositional variables (e.g., Goldberg et al. 2020: 21).

In short, skeptics need to do better than point to samples that do not represent a given population; yet, experimentalists also need to be better in clarifying their target population and theorizing about what they are testing. These two points highlight an irony such that the rise of cheap convenience samples has generated an unproductive cycle: experimentalists often collect data before developing theory and considering design issues (as I discuss in the next chapter) and then critics dismisses these studies but for the wrong reasons (e.g., because of the data source rather than the design considerations). An experimental sample should cause concern with the expectation of underrepresented moderators; otherwise, other topics should receive more attention (e.g., measurement, contextual generalizability, etc.).

Alas, this does not mean that experimentalists should abandon probability samples, or, in the case of laboratory and field experiments, more representative samples when possible (see, e.g., Lavrakas et al. 2019). As mentioned, work to date explores only a few topics and limited moderators, and in fact, rarely, if ever, looks at the possibility of intersectional moderators (e.g., minority Democrats). Probability samples offer opportunities to investigate heterogeneous effects with more confidence, particularly if the samples can be large (see Coppock et al. 2018: 12445). This exploration can be done to isolate a priori heterogeneous predictions or post hoc to build theory. Notably, too, when it comes to non-experimental work, the evidence makes clear that probability samples offer superior accuracy on validated survey measures (e.g., sex, age, race, ethnicity, etc.) compared to all other samples, regardless of weighting approaches (MacInnis et al. 2018).[96] Further, to arrive at the conclusion that non-probability representative

[96] Probability samples also suggest a more moderate political population (in terms of attitudes towards policy issues) than non-probability samples (Bilgen et al. 2018).

or convenience samples ostensibly offer a strong inferential base, past work has engaged in comparisons with similar studies on probability samples. Put another way, probability samples serve as an irreplaceable baseline. Finally, there exist applied research domains where probability samples ensure precise causal inference. It is for these reasons that the Time-sharing Experiments in Social Sciences (TESS) program has been a valuable resource – as discussed in Chapter 1, this program has allowed researchers, on a competitive basis, to field more than 550 survey experiments on probability samples of the U.S. population (http://tessexperiments.org/). TESS served as the basis of the data for the Mullinix et al. and Coppock et al. comparisons. In the end, evaluating any sample depends on the goal of the research, nature of the population, other dimensions of external validity, and potential sources of heterogeneity.[97]

One final note concerns the grossly neglected topic of panel conditioning effects in experiments. Panel conditioning occurs when participation in a prior study affects responses in subsequent studies. The bulk of survey experiments now collect data from internet panels, including some probability sample ones, where participants partake in multiple surveys over time. While there exists clear evidence of panel conditioning in longitudinal surveys – for example, respondents in the General Social Survey become more likely to report their personal income after the first survey (e.g., Halpern-Manners and Warren 2012, Halpern-Manners et al. 2017) – scant research explores the impact on experimental studies. This area is in need of inquiry, especially given concerns about "professional respondents" who participate in hundreds

[97] In many cases the resources needed for a probability sample could be better spent on other aspects of the study (e.g., Shadish et al. 2002: 386).

of studies (e.g., Hillygus et al. 2014, Stewart et al. 2015).[98] The professional class of respondents themselves may condition experimental effects, making panel conditions a more salient threat to causal inference than the demographics of the sample.[99]

Summary

(1) Experimental samples have been the source of much critique, stemming historically from the prevalence of student samples. The rise of internet panels – particularly, opt-in non-probability and convenience labor market panels – has provided experimentalists with new opportunities for more heterogeneous samples when it comes to survey experiments.

(2) The main concern about samples, even internet panels, is whether they generalize.

    a. It is crucial to recall that the goal of an experiment is to generalize a causal inference, rather than descriptive aspects of a population.

    b. As such, generalizing depends on whether the causal effects, in the population, are homogenous, heterogeneous where the sample includes variance on the moderator, or heterogeneous where the sample includes virtually no variance on the moderator. Only the last case would produce a biased treatment estimate, as

[98] This is a particularly acute concern when it comes to crowdsourcing platforms like MTurk where there are additional issues about inauthentic responses from bots (Chmielewski and Kucker 2019).

[99] Professional respondents may anticipate experimental hypotheses, creating demand effects which occur when participants change their behavior in response to their perception of the experiment's objective (Glennerster and Takavarasha 2013: 317). Mummolo and Peterson (2019), however, show that more generally, demand effects seem unlikely, at least in political science survey experiments.

long as the heterogeneous effects are modeled. Consequently, even ostensibly wildly unrepresentative samples can provide accurate estimates.

c. Evaluating samples is more of a theoretical than empirical process since it involves identifying moderators rather than comparing population benchmarks.

(3) A growing empirical literature suggests many experimental effects are homogenous; however, this work is still in its infancy.

a. Here, more heterogeneous samples – ideally probability samples – can be useful for post hoc identification of moderators that can help stimulate theory building. Such samples also serve as an important benchmark, and, in applied domains, ensure more precise effect size estimates.

b. Weighting experimental samples may only be necessary when one expects a moderator that correlates with what one would weight (e.g., groups underrepresented in the sample).

c. Heterogeneity based on contexts, treatments, and outcome measures remains underexplored; researchers should carefully consider these dimensions when they generalize.

Conclusion

I conclude this chapter with an e-mail communication I received that raised concerns about an experimental finding: "the smaller the effect size, the easier it is to think that there might be no effect at all in a more realistic setting. To put it another way: skeptics will deny that small estimates like these prove the existence of a treatment effect in settings that we actually care about… we're probably all leaning too heavily on "existence proof" justifications of our survey experiments." The author wrote with good intention of in fact defending an experiment

and perhaps characterizing common critiques (while not necessarily agreeing them them). Nevertheless, this message exemplifies common problematic discussions of experiments. First, reference to the "realistic settings" or "settings that we actually care about" misconstrues the purpose of experiments. Matching mundane realism, which is what I imagine the author of the e-mail has in mind, is not, with the exception of clearly applied studies, an important evaluative dimension. Theoretical match is what matters. Moreover, failure to be more direct about what that "realistic setting" entails makes the critique problematic insofar as one can dismiss any study on a lack of realism. For example, one could always say even studies using behavioral data may not use behavioral data that correctly represents "typical behavior."

Second, all empirical tests evaluate a hypothesis and if the data coheres with that hypothesis then that is positive proof; if not, then one needs to re-assess, perhaps considering the possibility of heterogeneous effects, some other alteration to the theory, or rejecting the theory entirely (or, as I later discuss, identifying problems with the design/data collection). While some argue for stricter standards of what the profession should view as a meaningful statistical relationship (e.g., Benjamin et al. 2018), that is somewhat of a distinct issue. Put another way, my understanding of Popperian falsification is that "existence proofs" are virtually all one ever has, since one can never definitively accept a hypothesis. This is not to minimize the importance of identifying the conditions under which a given effect holds (as I will discuss in Chapter 5), but ultimately, no empirical finding is conclusive. Third, effect sizes from a given study or a few studies should generally be read cautiously regardless. Given that a single study never includes representative probability samples of all the dimensions of external validity, the effect size from a single experiment rarely can be generalized. As I will discuss in Chapter 5, this explains why

meta-analyses that aggregate similar studies across settings, samples, operationalizations, measurement, and time can be valuable.

I would, in fact, go so far as to argue that much more important than discussions of external validity are considerations of designing treatments that match theoretical constructs of interest and designing studies that facilitate proper comparisons across meaningful counterfactuals. As I will discuss in Chapters 4 and 6, these tasks are *much* more difficult than most realize. The main takeaways from this chapter in terms of what scholars should consider when evaluating experiments are as follows.

- If one's goal is to test a policy intervention, mundane realism – resemblance to the intervention of interest – matters. In other cases, the focus should be on constructing treatments that resemble the theoretical constructs of interest and ensuring participants take the study seriously (i.e., experimental realism). In these cases, mundane realism matters little.
- Validating a construct, as well as minimizing treatment confounds, requires empirical work via pilot testing and/or manipulation checks.
- The assessment of external validity or generalizability requires clarifying what is being generalized (e.g., the existence of a causal relationship), to what one hopes to generalize, the nature of the sample, setting, treatments, and outcome measures, and the point of comparison (e.g., relative to what "other" hypothetical study?).
- The assessment of whether one can generalize from a given sample requires much less concern about the "representativeness" of the sample, and much more concern about the existence of heterogeneous effects. Critique of a study based on its sample requires an

accompanying specification of the variable on which the sample lacks variation to test for the posited moderator.

- Representative (ideally probability) samples facilitate searching for heterogeneity, may offer more precision when effect sizes are essential (in applied studies), and serve as a useful baseline. Thus, such samples play a role but are far from necessary for most experiments.

Chapter 4: Innovations in Experimental Designs: Opportunities and Limitations

In this chapter, I turn to designing experiments. I emphasize that experimental designs need to be driven by the substantive questions being asked. While textbooks on design elements elucidate crucial design options (e.g., Schneider 2013), the "best" designs are tailored to a question and context. Experimentalists should never choose a design approach before thinking through the question being addressed, and should always be cognizant of what they want the design to accomplish. The concern that design choices drive the questions being asked likely led Smith (2020: 15), as quoted at the start of the book, to worry that "experiments will unduly constrict the questions political scientists ask." Experiments are not the right approach for many questions and, even when they are, there are no off-the-shelf designs.

I focus on three designs that have gained prominence: audit field experiments, conjoint survey experiments, and lab-in-the-field experiments.[100] These constitute examples of field, survey, and lab experiments, respectively (i.e., the main "types" of experiments). I provide readers with overviews of each design to facilitate use of the given approach. I also discuss limitations to accentuate what each design can and cannot do. I do not mean to sully the substantial advances offered by these methods, but rather to remind readers that an approach works only when it fits the question being asked, and any approach faces constraints when it

[100] The emergence of these designs can be seen from an analysis of the experimental articles in the *APSR* as displayed in Figure 1-1. Each article was coded for whether it used a "non-traditional design" which included an audit field experiment, conjoint survey experiment, lab-in-the-field experiment, or a related novel design (or a "natural experiment"). Before 2000, 4% used one of these designs. This number jumped to 13% from 2000 to 2009, and 32% after 2009.

comes to the questions it can address. I conclude the chapter with a discussion of how different experimental designs can be used in the bourgeoning area of public policy evaluations.

Audit Field Experiments

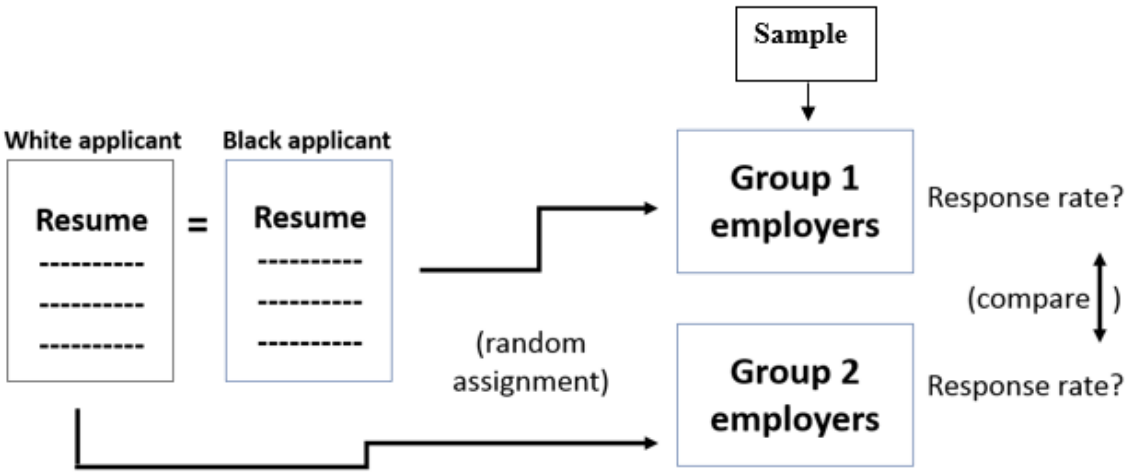
Field experiments entail the administration of the treatment in a naturalistic setting, often unobtrusively. These studies allow for the assessment of effects in naturalistic contexts, which is essential for applied studies aimed to "whisper in the ears of princes." They also can be helpful when the experimental goal involves theory building: field experiments allow researchers to "control" for some contextual confounds that may have been ignored or are too difficult to account for in a survey or laboratory experiment.

Social scientists have utilized field experiments in countless settings to study a diverse range of phenomena such as voter mobilization, election monitoring, micro-finance programs, aid programs, educational interventions, policy interventions, the impact of inter-personal contact, etc. (e.g., Bloom 2005, Gerber and Green 2012, John 2017). As I discuss further in Chapter 5, there now even exists an organization that matches researchers with organizations to collaboratively implement studies, including field experiments (i.e., research4impact). There also is a well-developed statistical literature on how to address inherent challenges in field experiments, including compliance (do those in the treatment group receive the treatment?), attrition (can those in the study be followed for measurement?), and spillover (will those in the treatment talk to those in the control?) (Gerber and Green 2012).

The particular type of field experiment on which I focus is called an audit or correspondence study. I do so, in part, because of the rapid growth in application, but also because the issues which I will regarding these design highlight basic design principles more generally. Audit studies aim to document discrimination, defined as "unequal treatment of

persons or groups on the basis of their race or ethnicity" or some other attribute (Pager and Shepherd 2008: 182). Discrimination differs from prejudice, racism, or stereotypes as it constitutes a behavioral outcome. It can stem from differential treatment – based on tastes or aggregate statistical perceptions – or biased institutional rules. Social scientists have a keen interest in documenting the presence of discrimination given the fundamental value of equality and the ideal that democratic government officials treat citizens' input equally (Dahl 1956, 1971). Moreover, there exist a host of non-discrimination legal protections based on particular attributes (e.g., laws against housing, employment discrimination).

Figure 4-1: Audit Study Logic



Audit studies entail researchers sending out fictitious or real but controlled applications / messages that are identical except for randomly varied dimensions of interest, such as the applicant's / messenger's race, religion, age, gender, disability, etc. Figure 4-1 displays the basic process. One first identifies the population of interest – in many cases in sociology and economics, this consists of a group of employers. One then acquires the sample, which in some cases can be a census if the researcher obtains contact information for all units (e.g., all employers). Next, the researcher devises, in this running example, two resumes that are identical

other than that one comes from a white applicant and the other from a Black applicant (more on how to vary race below). The researcher then sends a job application, randomly determined, to each employer and waits for a response, such as an invitation to interview for the job. If the response rates significantly differ (e.g., higher response rate to white applicants), this disparity constitutes evidence of racial discrimination in the job market, since race constitutes the only differentiating dimension across resumes. In short, the design relies on the statistical solution – via random assignment – to the Fundamental Problem of Causal Inference to conclude that race causes discrimination.[101] These studies "audit" the market place for discrimination, which can, in turn, lead to legal protections to minimize bias.[102]

Job audit market studies have a long history, stemming back to the efforts of the British Race Relations Board to identify discrimination in housing and employment in the 1960s, and fair housing audits in the U.S. in the 1970s (Gaddis 2018). One of the more influential audit studies is Pager (2003). She explores the consequences of previous incarceration on employment prospects for white and Black job seekers in Milwaukee. The addition of incarceration as a factor alongside race allows one to disentangle the two variables; this step is necessary because a disproportionate percentage of the Black population is incarcerated, confounding the two

[101] It is presumed the underlying assumptions are met, such as SUTVA (employers do not talk to one another) and the exclusion restriction (employers are not aware of the study or affected by other factors).

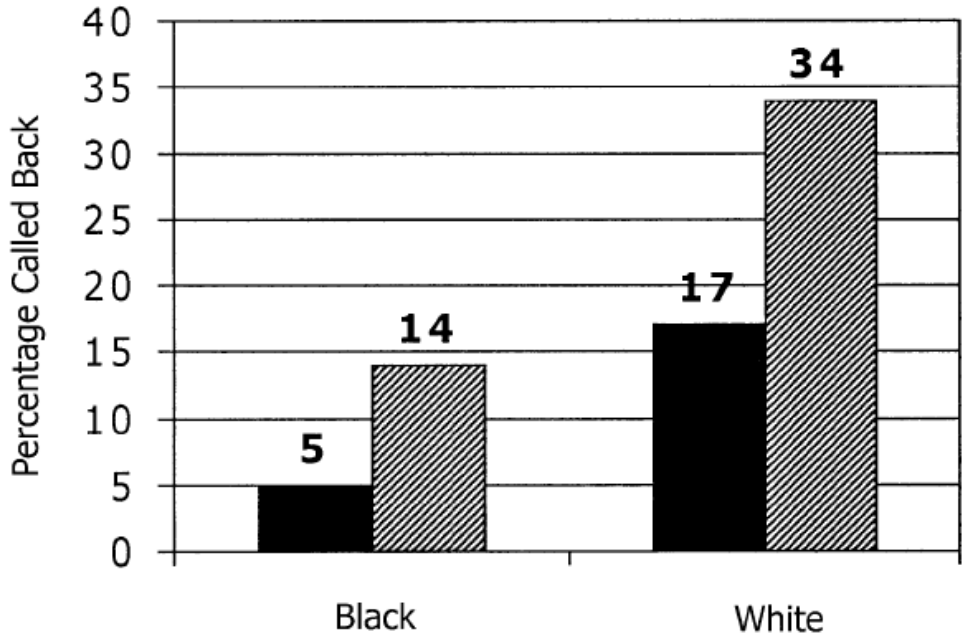
[102] The approach limits social desirability bias in measurement that could arise with survey approaches.

factors.[103] Pager assigned each employer to randomly receive (in-person) entry level applications from two similar white job applicants where one had a criminal record and the other did not, or from two similar Black job applicants where one had a criminal record and the other did not.[104] The white pair applied for 150 jobs while the Black pair applied for 200 jobs. The call back rates for different applicants (e.g., receiving an interview) appear in Figure 4-2 (taken from her paper) where the Black bars come from those with a criminal record and the striped bars come from those with no criminal record. Pager finds clear evidence of racial bias – in fact, even whites with a criminal record receive slightly more callbacks than Blacks without a criminal record. The results also reveal notable discrimination against those with a criminal record such that it decreased the likelihood from 34% to 17% for white applicants and from 14% to 5% for Black applicants. This striking evidence of discrimination played a role in the Ban the Box effort to force companies to eliminate questions on job applications about past felony records.

Figure 4-2: Pager's Results

[103] It also allows one to address the question of whether prior incarceration has a causal role in lower employment rates (as it could be a spurious relationship since incarcerated individuals may be more likely to have behavioral problems, poor interpersonal skills, substance abuse issues, etc.).

[104] The applicants were similar in terms of physical appearance and self-presentation, and were made similar in terms of objective characteristics (e.g., educational attainment and work experience).



Pager's study is one of many explorations into job market discrimination. Quillian et al. (2017) synthesize 28 hiring discrimination studies, reporting that, on average, whites receive 36% more callbacks than Blacks, and 24% more callbacks than Latinos. Moreover, the authors find no change in the level of discrimination against Blacks over the first part of the 21st century, and only modest evidence of a decline in discrimination against Latinos (also see Neumark 2018 for a review).[105] Other studies explore discrimination based on gender, age, disability, immigration status, mental health, military service, parental status, physical appearance, religious

[105] Quillian et al. (2019) analyze 97 discrimination studies in nine countries, finding significant discrimination against nonwhite natives in all countries; discrimination against white immigrants is present but low. Quillian et al. (2020) analyze 12 studies that include a job offer outcome (beyond hiring). They find additional discrimination in hiring after the callback: majority applicants receive 52% more callbacks and 128% more job offers than comparable minority applicants.

affiliation, sexual orientation, social class, and unemployment status, *inter alia* (Gaddis 2018). Audits expand beyond the job market to look at topics such as housing inquiry responses, response to roommate requests, doctor's appointment scheduling, responsiveness from professors, and the price paid for bargained goods (Gaddis 2018).

The bulk of more recent studies no longer use Pager's in-person approach (i.e., her job applicants went in-person to the employers, acting as confederates), and instead apply or send inquiries via phone, mail, online systems, and especially e-mail. Many refer to these approaches as "correspondence" studies (echoing the terminology of Internal Revenue Service tax audits and correspondence audits). Most recent studies also do not use the paired approach a la Pager but simply randomly assign a single applicant to each target (i.e., non-paired) – this is done since in correspondence studies, all aspects but the attribute of interest are kept completely constant (i.e., literal identical resumes) rather than "largely similar" as in Pager's paired in-person case.

A prominent political science application involves auditing legislative responsiveness. In their foundational study, Butler and Broockman (2011) sent (fictitious) e-mail requests to state legislators in 44 states requesting information about how to register to vote in upcoming primary elections. Each legislator received a request that randomly varied the race and partisanship of the sender. The authors, following the common approach, signaled race with the name Jake Mueller for white and DeShawn Jackson for Black, pointing to the objective correlations of those names with the given races (e.g., from census data).[106] They report that when e-mails do not signal partisanship, legislators, overall, are roughly 5% less likely to respond to the request from the

[106] They signaled party by having the requestor mention a particular party's primary.

Black sender (DeShawn).[107] In contrast, minority Democratic legislators are dramatically more likely (by 16.5%) to respond to minority inquiries. This paper stimulated an enormous literature of more than 40 studies that consistently report racial bias in responsiveness (e.g., Costa 2017). Political scientists also have used the audit design to study the responsiveness of local electoral officials (White et al. 2015), bureaucrats (Einstein and Glick 2017), welfare officers (Hemker and Rink 2017), school principals (Pfaff et al. n.d.), and more (for reviews see Butler and Crabtree 2021, Nathan and White 2021).

Audit Limitations

Audit/correspondence studies afford much insight by providing estimates of discrimination. They also follow a relatively straightforward design logic. Even so, there exist fundamental limitations of the approach that researchers need to keep in mind. I touch on three here.

The first issue concerns points of comparison – as I emphasized in Chapter 2, experimentalists need to take care when identifying the counterfactual of interest. Consider Butler and Broockman's study where their primary goal concerns whether legislators, particularly White legislators, discriminate against constituents based on their race. They find that they do since, for example, minority Democratic legislators respond to minorities 16.5% more often than to non-minorities, and white Democrats respond 6.8% less to minorities (than to non-minorities). Here the comparison is between responses to the Black versus the white inquiry.

[107] They also find legislators from both parties are significantly more responsive to co-partisans (by about 4.5%).

However, drawing further inferences become complicated. The authors state that one of their motivations involves exploring descriptive representation; in light of their results, they (472-473) state "one of the arguments often advanced for increasing the number of minority legislators through mechanisms such as majority-minority legislative districts is that elected officials better represent whom they share characteristics… our results provide direct support for the broader argument that how effectively minorities are represented does depend on the race of their representatives, regardless of their party" (472-473). The experimental results support the first point in this quote. But the second point about effective representation depending on race is less clear because one needs to think through what it means to be "effectively represented." To see why, consider that the overall response rates show that white Democrats respond to 54% of Black requests while minority Democrats respond to 46% of Black requests.[108] If one uses the rate of response rather than response across experimental conditions, then white legislators may be more effective versus minority legislators.[109]

This raises the question of what constitutes the appropriate comparison: relative response across experimental conditions or total response rates? The former suggests greater descriptive representation among minority legislators since they respond relatively more, but the latter suggests greater absolute responsiveness by white legislators.[110] Should descriptive

[108] Minority Democrats responded to about 29% of white requests.

[109] The authors group minority legislators (that include Blacks, Latinos, Arab Americans, etc.), and it may be that strictly Black legislators are more responsive – but the larger point is that one needs to think carefully about making inferences on responsiveness.

[110] One possibility is that minority Democrats disproportionally represent safe districts.

representation be based on overall response or differential response? That requires a conversation with normative theory. Put another way, the researchers' aims were met here and they used a reasonable point of comparison (and carefully drew their conclusions), but in drawing larger inferences, one needs to consider alternative points.

A similar point can be made about Pager's study: she compares an applicant with a criminal record against one without a criminal record.[111] As with Butler and Broockman, her chosen comparison makes sense given the main goal of the study. Yet, again, once one tries to draw further inferences, it becomes tricky to identify the appropriate counterfactual comparisons. For example, should the comparison be an ex-felon against someone with an ambiguous background? This question is relevant as some evidence suggests the aforementioned Ban the Box movement has ironically led to an increase in racial discrimination (Doleac and Hansen 2018): without an explicit statement of not having a criminal record, employers discriminate by assuming Black applicants are more likely to have a criminal record.

While concerns about the relevant counterfactual apply to any experimental design, it can be particularly troublesome with audits since a "pure" control group does not exist – everyone must receive a treatment inquiry of some sort to measure the response to a request. Consequently, experimentalists need to carefully attend to what comparisons support what conclusions. As is made clear by the discussion of two of the most influential audit studies in the social sciences, this is far from straightforward.

[111] Pager (2003: 951) operationalizes the latter by having the non-offender graduate from high school a year later and having a work record thereafter, thereby accounting for an ostensible complete adult life history.

A second issue concerns treatment confounds/construct validity. Most audits require signaling an attribute, such as race, to identify discrimination based on that attribute. With race, most rely on names, as in the Butler and Broockman study, often turning to objective data on the frequency of name usage by racial groups. Yet, one could question whether respondents *perceive* the race accurately. Many worry particularly about confounding race and class given the realities of the correlation between the two in the United States (Butler and Homola, 2017, Gaddis 2017). Researchers face particular challenges with audit experiments because one cannot include manipulation checks such as asking respondents for their perceptions of the requestor. This limitation then introduces a serious confound possibility (e.g., treatments differentially signal class) (Fryer and Levitt 2004).[112] The problem becomes even more acute with perceived variance in unobservable differences; for example, a legislator might assume the average college educated white constituent votes at a similar rate as an average college educated Black constituent but may suspect greater variance in voting among the latter group that, in turn, conditions response (Neumark 2012). Solutions exist such as extensive piloting, but that requires having a large enough sample pool to use for piloting. As mentioned in Chapter 3, there may even be geographic variation in name perception (Crabtree and Chykina 2018), and thus piloting might require using heterogeneous parts of the sample and then implementing the study using a host of names (Butler and Crabtree 2021). These same concerns apply to other attributes beyond names,

[112] The earlier in-person audits were particularly vulnerable to the criticism of the impact of unobservables stemming from how confederates presented themselves (Heckman 1998).

with the general point being that the inability to include manipulation checks and the challenge of large-scale piloting constrain audit designs.[113]

Third, audit designs typically aim to document the existence of racial or other types of societal biases exist. As such, their contributions remain undeniable. Yet, if not a primary goal, the designs rarely provide much insight into mechanisms. Consider a study that reveals racial discrimination. That data alone cannot reveal whether the discrimination reflects cultural racism, biological racism, racial stereotypes, or some other mechanism (Jardina and Piston 2019). In the case of the Butler and Broockman study, the mechanism could be any of those levers, or it could reflect political stereotypes about turnout or donor behavior.[114] That said, some recent innovative studies provide examples of ways to address mechanisms. One approach is to compliment an audit study with another type of experiment. For example, Quadlin (2018) uses an audit study to show that high-achieving (via grade point average) women face discrimination in the job market. She then implements a survey experiment to identify gendered standards as the mechanism (i.e., employers privilege likeability in assessing women as opposed to competence and commitment in evaluating men). Alternatively, one can design a correspondence study that manipulates

[113] One can try to control for confounds within the treatments (e.g., such as sending signals of the socioeconomic class, voting history) but this approach requires identifying potential confounds. The concern is similar to the informational equivalence issue discussed in Chapter 3.

[114] This ambiguity relates to the difficulty of distinguishing taste-based discrimination (e.g., racial animus) versus statistical discrimination (e.g., presumptions that minorities, on average, are less likely to vote even if arising from historic discrimination). Of course, regardless of whether discrimination is taste-based or statistical, the victims still lose out.

mechanisms within the treatment messages. For example, Pfaff et al. (n.d.) document discrimination by public school principals toward potential Muslim and atheist students. They further show that that discrimination increases when the requestor expresses more intense beliefs. Pfaff et al. thus identify a possible mechanism via their manipulation: the perceived costs/difficulty of accommodating those minority religions leads to discrimination. Pfaff et al. were able to do this because they had an enormous sample of more than 45,000 principals that facilitated the inclusion of more than a dozen conditions – an option often not available (and even then, they only could study one possible mechanism). Nonetheless, the idea behind the design sets an example of what research can do.

Each of these limitations accentuates the importance of aligning an appropriate design with the substantive question being asked; one must carefully consider what the comparisons reveal, what the treatments suggest, and what underlies a causal effect.[115] One also needs to remain cognizant of the major questions driving the research in the first place. For instance, audit studies in political science provide insight into representation by identifying bias in a type of responsiveness. Yet, the approach cannot speak to many other questions about democratic

[115] I have not touched on other audit design considerations, including concern about over-using the available population such that the respondents (e.g. legislators) learn of the studies and this "spoils the pool," significant ethical issues given there is no consent or debriefing, approaches to studying how to vitiate discrimination (see Butler and Crabtree 2017), the challenges of using the content of the responses as a relevant outcome measure (see Coppock 2019a), and logistical challenges (e.g., mail merges, personalizing requests). For a discussion of many of these issues, see Gaddis (2018), Butler and Crabtree (2021) and Nathan and White (2021).

representation. While some have used audits to study if legislators change their roll call voting behavior in response to constituents' communications (e.g., Bergan and Cole 2015), audits offer little insight into many factors that drive substantive representation such as over-time lobbying efforts, pressure from party leaders, and the competitive stream of communications public officials receive from citizens, groups, and the media. This type of competition defines democracy but is largely outside the purview of audits. The approach also has so far offered scant insight into another major dimension of democracy: participation. For example, one might ask whether responses or non-responses to inquiries affect efficacy and engagement among requestors. This would be a viable extension of audit experiments (e.g., exploring the reactions of real-world auditors): how much do such interactions affect participation and engagement? Further, as Butler (2014: 127) points out, the bias in political responsiveness the audits reveal raises another question of how politicians who seem to discriminate arrive in office in the first place – we need to understand "what determines who serves." Audits contribute to what we know about one dimension of democratic representation. However, they cannot themselves address many other questions about representation. Scholars need to always situate the design in the context of relevant questions, and recognize the potential and limitations. Do not conduct an audit just to do so – think about the larger question being asked and what new knowledge can be gained, particularly given the subject pools for many audits are over-taxed (i.e., there have been many experiments on elected officials).

The importance of ensuring a given design connects with the question being asked applies to any type of experiment. Moreover, the limitations highlighted above, while discussed with regard to audits, constitute widely applicable considerations. As emphasized in Chapter 2, experimentalists often fail to carefully justify their points of comparison and that can undermine

any design, particularly when a pure control / no treatment condition is not used. Also as discussed in Chapter 2, measurement and treatment design can be challenging and need particular attention with the absence of manipulation checks. Finally, testing for mechanisms entails distinct designs that should be pursued when such processes are vital for advancing knowledge.

Summary

Field experiments allow for strong causal claims in naturalistic settings. Most discussions of the limitations of field experiments revolve around internal validity challenges that arise due to non-compliance, attrition, or SUTVA violations. Yet, one must also consider the extent to which the given design fits the theory being tested; doing otherwise runs the risk of arriving at a set of empirical relationships with little understanding of the underlying processes. The opportunities and challenges can be seen with audit field experiments.

(1) Audit field experiments offer researchers a unique opportunity to study discrimination. The design entails sending ostensibly realistic requests to "receivers" and randomly varying an element(s) of the request. Researchers can then identify discrimination based on different elements.

(2) Audits have been used for decades to study job market discrimination, and more recently, to study government responsiveness. Scholars commonly worry about the ethical and logistical challenges of audits, as well as the potential for "spoiled" overused subject pools (see above footnote).

(3) Yet, there also are basic design challenges.

    a. Audits do not have a "pure" control condition, meaning researches must carefully consider the points of comparison.

b. One cannot include manipulation checks in audits, and thus, ensuring construct validity requires extensive piloting.
c. It is difficult to isolate mechanisms in audit designs, and even when such approaches can be taken, it requires large samples.
d. Evidence from audit experiments needs to be placed in the context of the questions that drive the research in the first place (e.g., questions of job market inequities, unequal political representation, etc.).

Conjoint Survey Experiments

The initial rise of survey experiments, in the 1980s-1990s, stemmed partially from technological advances. Computer-assistant telephone interviewing facilitated the implementation of experiments within surveys where one could include a host of conditions (since the interviewer could rely on the computer to generate the conditions rather than tracking paper records) (Sniderman and Grob 1996). One of the more prominent designs involves vignettes where one varies different aspects of a narrative (Mutz 2011). Take, for example, Freese and Pager's (2004) design that varies race (Black/white/not specified), past life events (laid off/fired/prison) and current circumstance (steady job/unsteady job) to gauge the impact of each factor on opinions about government assistance for the person (Mutz 2011: 55-56). The vignette reads (with experimental variations in italics):

> Michael is a twenty-six year old *[Black / white / no race specified]* male with a high school degree. About two years ago, Michael was *[laid off from work / fired from his job / sent to prison for a felony conviction]*. Prior to *[getting laid off / being fired / going to prison]*, Michael *[had held down a steady job for a few years / had trouble holding down a job for more than a few months]*. Since he *[lost his job / was released]*, Michael has been actively seeking employment, but has had great difficulty landing a job.

This design already includes 18 conditions, thereby requiring a large sample for sufficient power, and may not include all relevant criteria people use when assessing public assistance

deservingness. Conjoint survey experiments offer a solution for some multi-dimensional choice situations. These designs go back to the 1960s and 1970s (e.g., Luce and Tukey 1964, Green and Srinivasan 1978) and have been widely applied in marketing and economics. They have become even more popular throughout the social sciences in the latter part of the 2010s.

A conjoint design asks respondents to choose from or rate (e.g., on a 7-point scale) hypothetical profiles described with multiple attributes to estimate the relative impact of each attribute on choice. For example, a respondent may be shown Figure 4-3 and asked to choose which washing machine he/she would buy. Then, the respondent receives another version, such as Figure 4-4, and is asked again to make a choice. And, so on, up to 5, 6, or more times.[116] The repetition vastly enlarges the data collected, thereby allowing the experimenter to introduce a larger number of factors – for instance, here there are six attributes (e.g., capacity, load), and each one may have multiple values/levels (e.g., price may include more than what is in these figures, such as including $400, $598, $648, $700, and $755).[117]

Figure 4-3: Washing Machined Profile 1

Feature	**Washing Machine A**	**Washing Machine B**
Brand	General Electric	Samsung
Number of settings	10	12
Price	$648	$598

[116] In some cases, the design may be "single profile" meaning there is just one profile presented (e.g., one washing machine). The order of the attributes is typically randomized. Bansak et al. (2021) suggest randomizing across (rather than within) respondents to prevent confusion and cognitive overload.

[117] Another option is to offer respondents vignettes with all the variations (rather than tables).

Color	White	Beige
Capacity (cu. ft.)	4.5	4.5
Load	Front load	Top load

Figure 4-4: Washing Machined Profile 2

Feature	**Washing Machine A**	**Washing Machine B**
Brand	LG	Whirlpool
Number of settings	8	12
Price	$400	$598
Color	White	Stainless steel
Capacity (cu. ft.)	3	4.5
Load	Top load	Top load

With this set-up, it is relatively straightforward to use regression to obtain the average marginal component effect (AMCE), which is effect of a particular attribute's value against another of its values (e.g., $400 versus $598), holding the other attributes constant at their average values in the design (Hainmueller et al. 2014). Thus, one needs to carefully think of the relevant baselines for each variable/attribute when presenting the results: a given effect is relative to the specified baseline (e.g., $598 versus $400) and confined to the scenario in the experiment, since other variables stay constant at their averages in the design. One can also identify heterogeneous AMCE's where individual attributes of the respondents serve as moderators (Leeper et al. 2020).[118]

[118] As with any interaction, the moderator's causal status must be interpreted with caution as it is not randomly assigned and, in this case, also is contingent on the excluded reference category

That conjoint survey experiments allow for the study of many factors at once makes them enticing. Sniderman (2018: 265) calls their broad rise in the social sciences as "arguably the most promising design innovation in survey experiments developed over the past decade." Bansak et al. (2021) identify more than 120 political science applications from roughly 2015 to 2019. The topics include voting and public opinion (Peterson 2017), immigration (Hainmueller et al. 2015), climate change (Bechtel et al. 2013), housing (Hankinson 2018), roommate choice and partisanship (Shafranek 2019), and more. Consider Hainmueller and Hopkins' (2015) study on public views of what types of immigrants should be admitted to the U.S. The authors argue that, unlike prior work, they can look at more than a few immigrant characteristics at once. They include 9 attributes (and 50 levels overall within those attributes), asking respondents to choose which of two immigrants they would give priority for entrance, and to rate each immigrant on a 7-point "admit" scale. Figure 4-5 displays one example profile from their study, while Figure 4-6 presents their results on four of the attributes. It shows that Americans prefer well-educated immigrants who speak fluent English and do not come from Iraq.[119] The aforementioned relative nature of conjoint results mean there is a preference against Iraqi immigrants relative to immigrants from India (i.e., the baseline), and fluent, educated immigrants relative to those explicitly with no formal education and non-fluent English (i.e., the comparisons are not relative to ambiguously described immigrants).

[119] They also report a preference for skilled, experienced immigrants who plan to work upon arrival and have made no previous unauthorized trips to the U.S. (and are escaping persecution).

Figure 4-5: Immigrant Profile

Please read the descriptions of the potential immigrants carefully. Then, please indicate which of the two immigrants you would personally prefer to see admitted to the United States.

	Immigrant 1	Immigrant 2
Prior Trips to the U.S.	Entered the U.S. once before on a tourist visa	Entered the U.S. once before on a tourist visa
Reason for Application	Reunite with family members already in U.S.	Reunite with family members already in U.S.
Country of Origin	Mexico	Iraq
Language Skills	During admission interview, this applicant spoke fluent English	During admission interview, this applicant spoke fluent English
Profession	Child care provider	Teacher
Job Experience	One to two years of job training and experience	Three to five years of job training and experience
Employment Plans	Does not have a contract with a U.S. employer but has done job interviews	Will look for work after arriving in the U.S.
Education Level	Equivalent to completing two years of college in the U.S.	Equivalent to completing a college degree in the U.S.
Gender	Female	Male

	Immigrant 1	Immigrant 2
If you had to choose between them, which of these two immigrants should be given priority to come to the United States to live?	○	○

On a scale from 1 to 7, where 1 indicates that the United States should absolutely not admit the immigrant and 7 indicates that the United States should definitely admit the immigrant, how would you rate Immigrant 1?

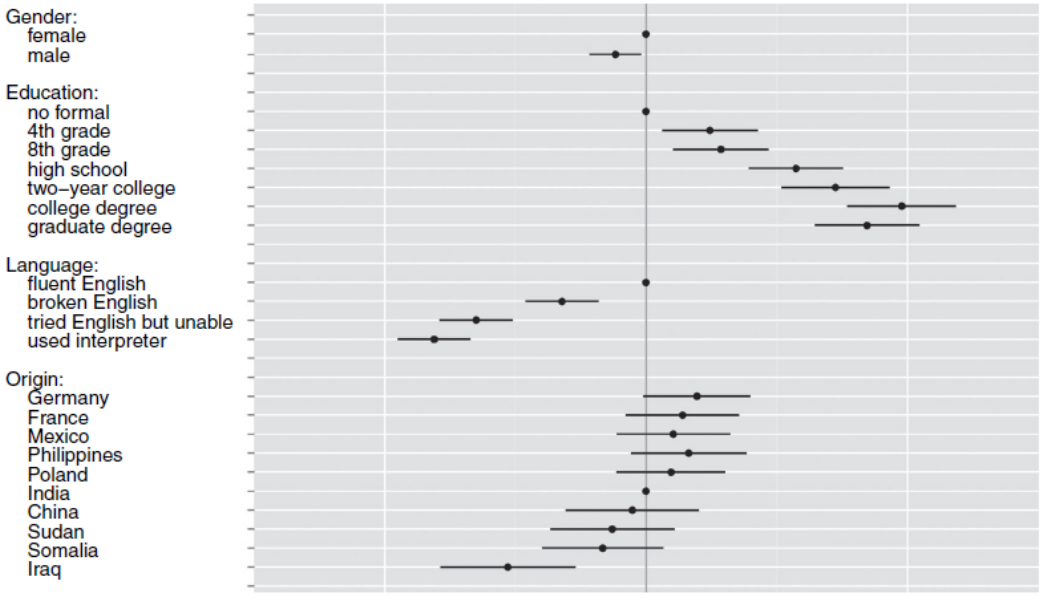
Absolutely Not Admit						Definitely Admit
1	2	3	4	5	6	7
○	○	○	○	○	○	○

Using the same scale, how would you rate Immigrant 2?

Absolutely Not Admit						Definitely Admit
1	2	3	4	5	6	7
○	○	○	○	○	○	○

Note: This figure illustrates the experimental design for the conjoint experiment.

Figure 4-6: Immigrant Conjoint Results



As prevalent as applications to immigration are studies of voting. For example, Bansak et al. (2021) explore candidate preference in the 2020 Democratic primary (i.e., preference to run against Trump), looking at age, gender, sexual orientation, race/ethnicity, previous occupation, military service experience, political experience, health care policy position, immigration policy position, and climate change policy position. Hainmueller et al. (2014: 4) look at hypothetical presidential voting based on "factors that emerged in recent campaigns," including age, religion, college education, profession, income, race/ethnicity, military service, and gender. Peterson (2017: 1194) studies congressional voting attributes, drawing from prior studies and voter guides, such as party, education, gender, family status, race, age, profession, military service, abortion policy position, and government spending policy position.[120]

[120] He shows that people rely less on partisanship as more detailed information is provided.

The power of these designs lies in their ability to test a host of factors in a fairly straightforward fashion. Moreover, it seems advantageous that survey respondents appear to not engage in satisficing behavior such that they can handle more than 10 variables/attributes, and more than 15, possibly up to 30 tasks (Bansak et al. 2021); they also seem to adjust their information processing efficiently as the amount of information increases (Jenke et al. 2020).

Conjoint Limitations

While some have raised cautionary notes about data analysis (e.g., Leeper et al. 2020), the positive aspects of conjoint experiments have swamped consideration of limitations. Yet, there exist at least three concerns that aggregate into a danger of letting the method drive the question. First, one must carefully identify the relevant available considerations in the information environment. When it comes to market products, it is relatively straightforward since the information often appears on the products. This also proved easy in one of the most cited conjoint survey experiments – the Hainmueller, Hangartner, and Yamamoto (2015) study that explores Swiss attitudes towards immigrants. They supplied information about gender, country of origin, age, years since arrival, education, integration status, and German language proficiency. They find that a paired conjoint choice design matches actual data from referenda that asked citizens to vote on whether individuals should be allowed to naturalize (e.g., country of origin matters quite a bit with penalties for those from Turkey or Yugoslavia). An often-neglected point about these results, however, is that the provided information perfectly matches what voters received in the referenda. In many applications such a match is far from straightforward. Consider the aforementioned U.S. voting conjoint studies that differ from one another in the provision of issue positions (e.g., health care, immigration, climate change or abortion and government spending) and personal characteristics (e.g., religion, sexual

orientation). Which factors should they offer? Which factors do voters typically encounter? One needs to be explicit about the environment being studied and motivate the inclusion/exclusion of different factors. That ensures sufficient construct validity of the treatment.

Second, aside from the availability of information, one needs to consider how people attend to information. Many studies show that people engage in biased searches, making it unrealistic to assume that everyone attends to the same information in the same manner (e.g., Hart et al. 2009, Iyengar and Hahn 2009, Druckman et al. 2012, Luo and Zhao 2019). Again, the Swiss example stands out given each respondent received the aforementioned information (e.g., gender, country of origin, etc.) and just that information. That worked since in the referenda being emulated, voters directly received the same information (although it remains unclear that they attended to it in the same way as the experimental respondents did). Yet, during many campaigns or similar situations, ignoring individuals' attention decisions seems problematic. Lau and Redlawsk (2001, 2006) made this point long ago; specifically, they argued that if one wants to study political campaigns or analogous environments, a static approach that exposes respondents to an information board akin to a conjoint table serves as a poor analog. A much better approach – and the one they developed – has respondents navigating and choosing information from a large amount of streaming information (i.e., a dynamic environment). I do not mean to suggest that studies must incorporate attention-seeking behavior into their designs; rather, experimentalists using conjoint designs need to consider likely availability and attention when developing stimuli to match the situation of interest.[121]

[121] As mentioned, Peterson (2017: 1194) explains that he drew information in his conjoint from voter information booklets, but how many respondents read those and processed all the

Third, in many cases, individuals lack the motivation and/or opportunity to process multiple dimensions as they are asked to do in conjoints (e.g., Fazio 1995). Ironically, discussions of conjoint experiments often focus on whether respondents engage in satisficing behavior if asked too much (e.g., Hainmueller et al. 2015: 2400). Yet, it may be that satisficing captures how people make decisions. Here again, the Swiss study is exceptional as it involved a personally high salience choice (i.e., about who would live in the country) that likely stimulated motivation. The question going forward should be whether the situation under study generates the assessment of all the dimensions offered in the conjoint design.[122]

---

dimensions he included? Notably absent from most conjoint studies of voting are cues from friends and information on personally salient issues to the respondent (although see Fowler 2020 for discussion). Another example is climate change attitude studies that exclude partisanship (e.g., Bechtel and Scheve 2013) despite partisanship, at least in the U.S., being a key driver of climate opinions (Bolsen et al. 2015).

[122] These assumptions concerning information availability, attention, and processing are implicitly assumed with the analysis approach since it identifies the relative impact of an attribute holding other factors at their means. If those other factors are not attributes that would be considered, then the impact of a given attribute may be misleading. Another issue is that the randomization of levels within attributes can significantly affect the results and thus should be carefully considered to resemble the situation (population) of interest (e.g., Hainmueller et al. 2015, Bansak et al. 2021) Finally, as Leeper et al. (2020) explain, subgroup analyses can be tricky due to respondent subgroups differing in their evaluations of the comparison points (e.g.,

The limitations I have highlighted point to design considerations that apply to any survey experiment. Experimentalists need to construct treatments that operationalize the theoretical concepts of interest (see Chapter 3's discussion of designing treatments). Survey experiments, in most cases, also presume a captive audience who receives the information being tested. That is not problematic on its face but experimentalists need to take care in specifying to what context the design generalizes. F

Summary

Few designs have so quickly captured the interest of social scientists as conjoint survey experiments. The attraction comes from the ability to introduce many variables and retain sufficient power to make inferences. Yet, the validity of these designs remains unclear, given that the main case of verification comes from a Swiss referenda study that contains qualities that differ from other applications in terms of the availability of information and the salience of the choice.[123] In fields where conjoints have a longer history, scholars have recognized concerns about introducing too many variables without theoretical motivation. For example, Ben-Akiva et al. (2019: 1) state, "The promise of stated preference experiments is that they can provide deeper and broader data on the structure of consumer preferences than is obtainable from revealed market observations, with experimental control of the choice environment that circumvents the feedback found in real market equilibria. The risk is that they give pictures of consumers that do

---

effects are relative to a baseline such as "no education" in an immigration conjoint, and respondents may differ in their evaluations of the baseline).

[123] There are other validation studies in marketing (e.g., Montgomery and Wittink 1979); however, the applicability of those to the other social sciences remains unclear.

not predict real market behavior." McFadden (2017: 161-162) identifies a set of conditions under which conjoint designs work best and these include ensuring the alternatives are "fully described" but not "overly described." In sum, one should arrive at a conjoint design when the question being asked and theory offered suggest a decision-making environment where individuals have the given information available, access it, and process it in a multi-dimensional fashion. Testing extra factors simply because the design allows for it impedes theoretical development – as it makes for a stimulus that has poor construct validity. This is the case for any survey experimental design, regardless of whether it is a conjoint.

(1) The conjoint survey design offers experimentalists the opportunity to incorporate a large number of factors by asking respondents to repeatedly rate hypothetical profiles with multiple attributes.

(2) Inferences from the design need to be interpreted in terms of the (all else constant) relative impact of a change in a given value against the designated baseline. This approach makes choosing the baseline important, since in essence it serves as the control comparison.

(3) When using the design, experimentalists should ensure the situation about which they are theorizing:

    a. includes information analogous to that in the experiment,

    b. prompts individuals to access information in a manner that echoes how they do so in the experiment, and

    c. stimulates individuals to process information in a way similar to how they do so in the experiment.

(4) Many decisions entail individuals accessing little information and processing it in a heuristic manner, and in such cases, it remains unclear that a conjoint design is appropriate.

Lab-in-the-Field Experiments

Laboratory experiments offer researchers incomparable control that facilitates addressing the Fundamental Problem of Causal Inference (e.g., using the scientific solution when applicable), and can bolster experimental realism (e.g., attention can be monitored). However, experiments in the lab also typically face substantial logistical challenges, such as coordinating with participants, monitoring treatment exposure, establishing the desired context, etc. Moreover, the nature of traditional laboratory experiments means reliance on convenience samples that can come to the lab. For this reason, researchers have sought to import the advantages of these designs to other settings by "bringing the lab" to a given population. This practice is known as a lab-in-the-field experiment: "one conducted in a naturalistic environment targeting the theoretically relevant population but using a standardized, validated lab paradigm" (Gneezy and Imas 2017: 440).

While social scientists have increasingly used lab-in-the-field experiments (e.g., Enos and Gidron 2016), their usage comes nowhere close to audits and conjoints. This scarcity likely reflects the aforementioned challenges to implementing laboratory experiments combined with even more hurdles when one moves to the field. Such difficulties include ensuring sufficient variance in the sample being studied (as discussed below), extensive piloting with convenience and targeted samples, recruiting a sufficiently large sample in the given location, accounting for safety and ethical concerns, forming and training a team to travel to the location for implementation, acquiring the needed budget, investing the time needed for data collection

which sometimes means running sessions with individuals or very small groups, confirming the literacy of the sample when necessary, and obtaining relevant permissions and involving the community when necessary (Eckel and Candelo 2021).

These hurdles mean investing in a lab-in-the-field study should not be taken lightly – it is not worth testing a proposition outside of a non-mobile (e.g., campus) laboratory just to introduce heterogeneity for the sake of doing so. As detailed in Chapter 3, when estimating an experimental treatment effect, the nature of the sample of units matters only when (1) there exist heterogeneous effects (that correlate with how the sample is drawn), and (2) either one cares about a precise effect size or there exists insufficient variance on the moderator.[124] Put another way, one reason to use the method is to exploit the control of the lab while also ensuring sample variance that would not otherwise be available. A second reason concerns contextual variation. Experiments conducted in a single location can only go so far in varying situational factors. Lab-in-the-field studies can overcome this limitation by implementing studies in distinct locations that differ in key ways. Here, one exploits context, rather than unit, variability. I next provide an

[124] An alternative reason is that a sample from the population of interest requires one bring the lab to the field. For example, Nelsen (2019) designed an experiment to explore how distinct pedagogical approaches in civic education affect political participation (across different racial groups). He implemented the study by going to high schools throughout the Chicago area where he randomly assigned students to read alternative excerpts. Here, bringing the lab to the field constituted the easiest and perhaps only way to reach the relevant population.

example of each "type" of lab-in-the-field study: sampled focused studies and context focused studies.[125]

Kim (2019) explores the question of why Americans continue to believe in the prospects of upward mobility despite growing income inequality and evidence of static or downward mobility. She theorizes that exposure to entertainment television programs that use a "rags-to-riches" narrative (e.g., America's Got Talent, Shark Tank, American Ninja Warrior, Toy Box) lead Americans to hold mobility beliefs. She offers suggestive evidence from correlational surveys, and then turns to an experiment to pin down the causal effect. In so doing, she recognizes the possibility of heterogeneous treatment effects, since those with pre-existing beliefs in the "American Dream" may be more susceptible to priming. Further, Republicans who tend to endorse individualism may exhibit stronger reactions. Kim therefore faces a design dilemma, since she seeks the control of a laboratory experiment (where she can ensure exposure to the stimuli – i.e., programs) but also needs a sample that varies in terms of prior beliefs and partisanship. On the latter point, she points out that obtaining such a sample in a liberal, metropolitan city (where she was located) is challenging. She therefore turns to a lab-in-the-field experiment.[126] She used a mobile media laboratory (with chairs, television screens, and tablet computers) in three counties where the partisan presidential vote in 2016 was split nearly 50%-

[125] Eckel and Candelo (2021) make a similar distinction calling them respectively one-context and cultural-comparison lab-in-the-field experiments.

[126] She complements her lab-in-the-field experiment with a survey experiment collected on an MTurk sample. The downside of the latter is she cannot ensure exposure, and that sample leans in a liberal direction.

50%. She recruited participants from farmer's markets, flea markets, and summer festivals, and obtained a heterogeneous partisan mix of people who seem to have faced some economic hardship but may have some standing beliefs in mobility (i.e., the American Dream). She randomly assigned respondents to watch a control broadcast (Cesar 911) or a treatment that mixed in various segments of the aforementioned reality television shows.[127]

Kim finds exposure to the rags-to-riches reality television program leads the audience to increase their beliefs in the prospect of upward mobility (by about 6%). Moreover, she finds heterogeneity such that, relative to Democrats, Republicans are much more strongly affected by the programing. Further, exposure significantly affects those who previously believed anyone who works hard can get ahead while backfiring on those who disagreed that anyone who works hard can get ahead (i.e., those who were originally pessimistic about the American Dream). Kim's study highlights the value of a lab-in-the-field experiment as she obtained a sample that allowed her to isolate differential causal effects while ensuring valid delivery of treatments. The results make clear that future work on economic mobility and entertainment television needs to attend to sample variance.

A second reason to use lab-in-the-field experiments is to exploit variation in contextual exposure. Consider, for example, Gilligan et al.'s (2014) study of civil war and social cohesion. The authors posit that, on the one hand, fatal violence from civil wars could increase collective

[127] The use of multiple segments increases the construct validity of her treatment since the treatment is not contingent on the idiosyncrasies of one show. She also engaged in extensive piloting to ensure the shows included the aforementioned features of reality television. The control show focused on life quality features that do not depend on visible financial gains.

social action via a purging mechanism (fewer social individuals leave area) and a collective coping mechanism (people band together due to a shared bad experience). They note that, alternatively, violence could depress cohesion due to an ensuing lack of trust and the destruction of civic associations. Testing the causal impact of violence requires identifying otherwise analogous people who differ only in their exposure to violence; that is, finding a way to ensure homogeneity and the scientific solution to the problem of causal inference. The authors turn to a lab-in-the-field study in Nepal (in 2009-2010). They (609) find a set of matched communities that strongly resemble one another (e.g., in terms of region, military control, timing of war exposure, ethnic and caste composition, socioeconomic development, etc.), other than exposure to low or high levels of violence.

They then went to these communities (setting up "labs" in the field) and asked participants to take part in a series of economic games that measure "sociality." They included measures of altruism (i.e., the dictator game where individuals give money to a partner), trust (i.e., where an individual gives an amount of money to a partner that is tripled and the partner decides how much to return, if any), and public goods contribution (i.e., individuals can keep an allocated amount for themselves or share it with others such that if everyone shares, everyone would be better off). Higher values in each game indicate more pro-social outcomes. The authors find substantially more pro-social outcomes in the violence-affected communities relative to those with less violence. For example, compared to those in non-violent communities, the respondents in violence-affected communities displayed about 13% more altruism (i.e., sent about 13% more in the dictator games) and 35% more trust (i.e., sent about 35% more in the trust game). These findings constitute clear evidence in favor of the first mechanism that violence increases collective social action, suggesting a foundation for building after violent experiences.

From an experimental perspective, this study shows how one can exploit the control of the laboratory to take advantage of natural variance that occurs in the field. The researchers do not control the independent variable (e.g., violence), but they still can make a casual inference about its effect due to the controlled setting – a la the scientific solution to causal inference. A fairly robust literature, particularly in economics, employs this approach. The Henrich et al. (2005) study on self-interest across cultures, discussed in Chapter 2, is a well-known example (for a detailed review, see Eckel and Candelo 2021). Researchers using this approach pay high costs for implementation (i.e., gaining access to the communities, recruiting, ensuring appropriate translations, etc.), but the payoffs can be enormous in helping to understand societal forces that ostensibly belie experimentation.

Summary

This section of the experimental design chapter differs from the former two; in those cases, I focused on the appropriate usage of popular designs, noting limitations and accentuating the importance of question and theory driven research. With lab-in-the-field designs, the challenges / limitations become apparent quickly. Caution should be taken in terms of when to make the sizeable investment in these designs – they can be useful to reach populations that contain crucial variance, or to take advantage of variance in societal conditions to test the effects of those differences.[128] In these cases, lab-in-field experiments serve as a powerful method.

[128] Eckel and Candelo (2021) explain that "a population that is outside the lab is not inherently more interesting nor more relevant than one that is in the lab… We assert that the question leads the researcher to implement this type of research because the attributes of the population or context are critical for addressing the problem."

(1) Lab-in-the-field experiments entail using a standardized experimental laboratory approach in a field setting to target a population. It allows for strong control over relevant populations.

(2) Lab-in-the-field experiments can be particularly useful to exploit needed variance in samples or contexts.

(3) The challenges inherent in conducting lab-in-the-field experiments mean they should be used when the sample or contextual variance is not otherwise available to test theoretically informed hypotheses.

By pointing to the implementation challenges, I do not mean to dissuade researchers from using lab-in-the-field experiments. More generally, there seems to be a decline in the use of laboratory experiments of any kind (Rogowski 2016). Yet, researchers would benefit from more appreciation that any type of laboratory experiment provides control vital to ensuring experimental realism, addressing the assumptions underlying casual inference, and understanding the nature of contextual factors. As detailed in Chapter 2, these advantages likely dwarf mundane realism considerations that likely play a role in the decline in laboratory experiments.

Using Experiments to Study Policy

One trend that I have thus far ignored is the growth in demand for experiments from outside organizations. While this has been the case in some disciplines, such as economics, for some time (e.g., Glennerster and Takavarasha 2013, Karlan and Appel 2016), it is a more recent trend in other disciplines. This often involves non-governmental organizations collaborating with social scientists to assess the impact of interventions (e.g., Levine 2021). This work sometimes focuses on privately motivated programs (e.g., educational programs), and other times involves

public policy evaluations. In this section, I offer a brief coda by discussing the design of policy oriented experiments that often, but not always, involve partnerships with organizations.

The idea of using experiments to study policy goes back at least to Campbell's (1969: 409) plea that "nations should be ready for an experimental approach to social reform, an approach in which we try out new programs designed to cure specific social problems, in which we learn whether or not these programs are effective, and in which we retain, imitate, modify, or discard them on the basis of apparent effectiveness on the multiple imperfect criteria available." Campbell (1969) recognizes that in many cases the ideal of random assignment is neither political feasible nor morally justifiable. That reality led him to emphasize "quasi-experimental" approaches. As explained in Chapter 2, this means using the scientific solution to causal inference by controlling for differences or assuming that non-randomly assigned units are equivalent. One famous example involves comparing the impact of Connecticut's crackdown on speeding in 1955, such that there were fewer fatalities after than before (this assumes nothing else changed).

Generally, when it comes to policy experiments, one can differentiate the solution taken to causal inference – statistical/random or scientific, which for this section I assume involves imperfect control (i.e., quasi-experimental). The outcomes of interest might be the impact of a policy a la Campbell or a policymakers' decisions. Table 4-1 displays the possibilities. In this framework, "policymaker" refers to governmental actors, *or* non-governmental actors that seek to provide services or information (e.g., non-governmental organizations (NGOs)).

Table 4-1: Using Experiments to Study Policy

	Policy Effects	**Policymaker Behavior**
Statistical Solution/Random Assignment	(1) Random allocation of government resources or policy access.	(2) Policymaker response to different forces/factors that are randomly varied.

Scientific Solution/Quasi-Experimental	(3) Nearly random allocation of government resources of policy access.	(4) Policymaker response to different forces/factors that are comparable.

In this configuration, audit studies of elected officials, as discussed earlier in this chapter, are an example for cell 2.[129] Alternatively, one can employ random assignment to study the behavior of members of an NGO or analogous policy relevant organization. For example, Han (2016) worked with a professional organization to send randomly varied versions of messages to doctors, encouraging them to sign a petition in support for the Affordable Care Act (in 2011 when the Act was being challenged on various fronts). She finds, relative to a generic message, a message emphasizing recipients' goals (by referencing their past actions), and, even more so, their personal goals (by referencing specific past statements they made) increase the likelihood of signing the petition. Han also explores how messages affect the likelihood of group members recruiting others to take policy action and attending a meeting. Overall, she shows that organizations can use relational messaging (e.g., connecting to people's personal beliefs, past history) approaches to stimulate actions that have consequences for public policy, in this case, health care.[130] Another example is Campbell and Spilker (2018) who implemented a survey experiment with individuals working for donor organizations (e.g., NGOs that offer humanitarian aid). In violence scenarios, donors tend to reduce budgetary and development aid and increase

[129] Others use survey experiments with public officials (e.g., Druckman and Valdes 2019).

[130] For a broader treatment of experiments on citizens' public policy (civic) oriented behaviors (e.g., recycling, volunteer, organ donation), see John et al. (2011). Also, see Bloom (2005) on analytic approaches for policy oriented experiments.

transitional and humanitarian aid. They do the reverse in peaceful scenarios. These studies employ random assignment to isolate factors that affect those who shape and make policy.

The other three cells have received relatively less attention. Cell 4 includes governmental responses to natural events such as disasters or more slow-moving phenomena such as climate change. For example, local governing bodies have implemented policies to respond to climate change events at distinct points in time and in different ways. This work reveals how institutional features of governments moderate policy response once climate change conditions hit a salient point (e.g., Bulkeley and Broto 2013). Alternatively, government procurement approaches seem to depend on external events such as variations in marketplace forces (e.g., Bandiera et al. 2009), and government spending can depend on near-random variations in population size and institutional rules (Egger and Koethenbuerger 2010).[131]

An example of studying non-governmental organizational responses using the scientific approach (cell 4) is Crawfurd's (2020) study of Latter-day Saints missionaries. Church officials assign missionaries in a "quasi-random" fashion using some information in a database but ultimately based on a "revelation." Crawfurd focuses on assignments to high-income European countries or low- and middle-income Asian, African, or Latin American Countries. He invokes a unit homogeneity assumption by presenting background differences between those assigned, showing only minor differences (e.g., regarding age, gender, language, prior countries visited, standardized test scores). He then reports that those assigned to African countries later report

[131] Political scientists have used experiments to explore citizens' reactions to natural events much more than policy effects (e.g., Malhotra and Kuo 2008, Healy and Malhotra 2010).

more interest in global development, and more policy actions such as donating and volunteering with international non-profits.

Cell 1 refers to the situation where the government or another policymaking body implements a policy in a literally random way. Social scientists – other than well-known Vietnam lottery studies (e.g., Erikson and Stoker 2011) – have not fully exploited these events (Grose and Wood 2020). Examples include random assignment of the implementation of federal election laws (Grose and Wood 2020), audits of lobbying disclosures (Wood and Grose 2020), and random assignment of land distribution taken from indigenous peoples (Hall et al. 2019). As Grose (2021) emphasizes, these provide unparalleled opportunities to asses policy effects and, even if not that frequent when it comes to governments, deserve more attention from scholars.

This type of random assignment policy experiment (cell 1) occurs more frequently with NGOs or non-profits who knowingly randomize to study interventions. For example, Jayachandran et al. (2017) worked with a conservation non-profit (Chimpanzee Sanctuary and Wildlife Conservation) that randomly assigned payments for ecosystem services to forest-owning households in Uganda if they conserved their forest. They find that tree cover declined less in villages that received the payment versus not. Another example is Blattman et al. (2014) who worked with the United Nations High Commission for Refugees and an NGO, the Justice and Peace Commission, to study the impact of an alternative dispute resolution educational workshop. They randomly assigned to some communities (in Liberia) to receive the workshop (i.e., 86 of 246 towns randomly received the workshop intervention). The authors find that a year later, towns where workshops took place reported a higher resolution of land dispute and less violence. On the other hand, they also find an increase in the use of illegal extrajudicial

punishments and some nonviolent disputes. A bourgeoning literature take similar approaches to look at the impact of non-governmental policy interventions (see Matanock 2021).

Finally, cell 3 is a path taken with some regularity by economists who exploit variation in policy implementation to isolate its effects. Other disciplines have ostensibly made less use of such designs. An example of the design is the Moving to Opportunity experiment implemented by the U.S. Department of Housing and Urban Development in 1994-1997. They chose five cities (Baltimore, Boston, Chicago, Los Angeles, and New York) to study the impact of housing on economic and health outcomes among more than 4,000 low-income families. They had hoped to employ a random assignment experiment (cell 1) with randomly selected families staying in low income residencies, having the option of moving to a nicer apartment in close proximity to their current neighborhoods (Section 8 treatment), or moving to a neighborhood with a poverty rate less than 10%. Since a non-trivial number of families assigned to the treatments declined to move, the design falls into cell 3 (comparing "similar" families across conditions a la the scientific approach).[132] Initial results, 4-7 years after implementation, suggested few clear effects on employment or earnings, and marginal effects on families' mental health. However, later results revealed that moving to higher income neighborhoods had dramatic positive effects on children who were young (less than 13 years old) when moving (e.g., higher income, more likely to attend college, better health on some indicators) but negative, albeit not large, effects on older children (Liebman et al. 2020).

[132] Sixty one percent of Section 8 families chose to move and 47% of new neighborhood families chose to move.

Gay (2012) takes advantage of the policy experiment to isolate political effects, showing that moving negatively affected voter turnout with it being higher among non-movers, followed by the Section 8 movers, followed by the new neighborhood movers. This effect stems from the new costs of political participation (e.g., finding one's polling place) and decreased social mobilization (e.g., socialization with one's neighbors declined which can diminish participation). Gay's work is a rare example of a political scientist using near random variation in policy implementation. Other under-explored examples would be the impact of state changes in minimum wage (e.g., does it change political beliefs?) (e.g., Card and Krueger 1994) and the impact of the over-time rollout of the food stamp program (e.g., does it affect political engagement?; e.g., Hoynes et al. 2015, 2016).

Another design that falls into cell 3 involves looking at pre- and post-intervention behaviors while also including a comparison group. For example, León et al. (2014) study an NGO intervention in rural India that provided information about family planning. They did not randomly assign the intervention, instead looking at pre- and post- intervention behavior in an area that received the intervention and a comparable albeit not equivalent area that did not (e.g., assuming causal transience, unit homogeneity). They find the intervention increased women's beliefs about their power regarding money and social relationships. This empowerment seemed to help with meeting contraception needs. Another non-governmental example is Gilligan et al. (2013) who use an exogenous shock where one of three NGOs implementing a reintegration package was not ready to start until a year later, in post-civil war Burundi. They find the program results in a 20% to 35% reduction in poverty but did not help political reintegration.

I do not offer a formal summary section. I reviewed these studies with the hope of stimulating more experiments to explore policy and those involved in policymaking. While such

studies may not fully satisfy Campbell's (1969: 409) plea for "an experimental approach to social reform," the reviewed work demonstrates enormous potential.[133] As this work proceeds, an aspirational goal is consider the full policy feedback process – from forces that drive government officials or NGOs to implement a policy to the effects of that policy on constituents and then back to how those making policy react to constituents. For instance, one could implement a dispute resolution intervention, study its effects on constituents, examine constituents' subsequent impact on policymakers, etc. (e.g., Mettler and Soss 2004). When do policies lead to backlash among constituents that then causes those making policy to retreat as opposed to generating momentum in the direction of the initial policy? Is there push back against too many dispute resolution programs in a developing country or do the programs build momentum such that demand for them increases? Another aspirational goal is to continue to address, as discussed in Chapter 3, the inevitable challenge of scaling results to inform policymakers who often face implementation challenges (Al-Ubaydli et al. 2017, 2020a,b).

Conclusion

With experiments now being a widely accepted methodology in most of the social sciences, researchers often look to new designs for innovation. This can be valuable when the designs allow for the identification of previously undocumented causal relationships and/or

[133] Some areas that may come close to meeting Campbell's vision include work in labor and development economics (e.g., Banerjee and Duflo 2009), and education evaluations (e.g., Cook 2002, Shadish and Cook 2009). As more policy oriented experiments are implemented, one consideration will be public concern about such interventions (e.g., Meyer et al. 2019a,b, Mislavsky et al. 2019, 2020, Heck et al. 2020).

theory development. Yet, the end goal remains unchanged: to accumulate knowledge to address relevant social science questions. It is crucial to not jump to designs for their novelty but only to turn to them when they offer an advantage over what could otherwise have been done.[134] There is no question the designs covered in this chapter – audit field experiments, conjoint survey experiments, and lab-in-the-field experiments – have contributed in major ways to multiple literatures. In that sense, the incorporation of these approaches highlights just how expansive experiments can be. Yet, scholars must move carefully to use these designs when appropriate and in so doing recall the basic lessons covered in the prior chapters. For example, in Chapter 2, I emphasized the importance of carefully considering points of comparison in an experiment. Audit experiments complicate identifying the relevant baselines, in part, because they lack a pure control condition. As was made clear in the discussion of Pager's and Butler and Broockman's classic studies, which baseline one uses can change what one concludes.

In Chapter 3, I argued for increased consideration of context in discussions of external validity. This applies particularly to conjoints given the possibility of incorporating unnecessary factors and/or presenting respondents with information environments that do not match those about which one theorizes. This point is not about mundane realism; rather, the concern involves putting respondents in a context where the available information and their motivations do not

[134] Of course, one cannot intuit a researcher's motivation for using a given design. However, the extent to which new designs (e.g., conjoints) become applied with such frequency following initial applications suggests an allure of prioritizing the design regardless of the substantive question (another sign is the frequency with which the names of new designs appear in titles of papers).

resemble those that underlie psychological theories about processing. I also, in Chapter 3, spent substantial time discussing the conditions under which one should worry about using particular respondent samples – the crucial question revolves around the nature of heterogeneous treatment effects. When a given sample may misestimate such effects, the lab-in-the-field approach can be extremely useful; it also offers an approach when it comes to contextual variation. Otherwise, even though lab-in-the-field experiments – just like audits and conjoints – offer a new intriguing approach, they may be unnecessary for building knowledge (although laboratory experiments more generally offer advantages that seem underappreciated). The bottom line is that novel designs offer new opportunities, but the fundamentals of what makes for a "good" experiment do not change. (I offer more details on constructing a "good" experiment in Chapter 6.) As Huber (2013: 2) explains, "[p]erhaps there's a sense that we have plenty of theories and that the main challenge we face is to figure out which variables actually have a causal effect. But this is wrong-headed – the very nature of research on causal identification requires a heightened rather diminished role for careful theorizing."[135] The usefulness of a causal test in the social sciences can only be assessed in terms of its contributions to extant knowledge and theory. The key points covered in this chapter are as follows.

[135] Huber (2013: 2) further states that some "take the strong position that social science research that cannot solve the identification problem is not worth doing, or at least is not worth publishing in leading journals. If we move towards this position, we excessively narrow the range of questions we ask, and thus unnecessarily limit our understanding of the social processes we study." This concern echoes Smith's (2020), as discussed at the start of the chapter.

- The expansion of experiments in the social sciences has led scholars to introduce novel designs; three notable examples include audit field experiments, conjoint survey experiments, and lab-in-the-field experiments. These offer new opportunities but also have limitations that recent work sometimes does not sufficiently acknowledge.
- Audit experimental designs have a long history of producing knowledge about discrimination in the job market context. Political scientists have adapted the design to study bias in democratic responsiveness. In so doing, researchers should carefully attend to making appropriate comparisons across conditions, ensuring construct validity via piloting, and using supplemental methods to pinpoint mechanisms.
- Conjoint survey experiments enable researchers to study a large number of factors; in so doing, researchers need to carefully theorize about the appropriate information environment, and respondents' attentional tendencies and information processing approaches. The concern is conjoints may over-saturate the environment and prompt attention to information that otherwise would not be accessed.
- Lab-in-the-field experiments combine the control of the lab setting with access to populations of interest. They offer a valuable approach – due to the general advantages of laboratory experiments – when one suspects heterogeneous treatment effects that require seeking out a targeted sample or variable contextual effects that cause different reactions. In such scenarios, returns on the substantial implementation costs are high.
- There are a host of experimental approaches to studying policy and contexts in which to study it that have not been fully exploited by social scientists.

- As experimentalists turn to new designs, they must be careful to recognize their limitations and use the designs when appropriate. A design is as useful only when researchers use it in a way that moves beyond extant knowledge.

Chapter 5: What to Do Before, During, and After an Experiment

Experiments constitute one part of the scientific process. As explained in Chapter 2, that process entails a series of steps that leads to the systematic production and organization of knowledge. The first step involves asking a question while the last step entails data analysis. Here I turn to practical issues involving events prior, during, and after one follows these steps (and in the context of doing an experiment). How do we arrive at questions in the first place? How do we document the formal steps we take – including the theory, hypotheses, data collection details, and analyses? What do we do when we are done – should we do it again and replicate the study?

Addressing these questions align with the open science movement. Open science initiatives accentuate the importance of making all aspects of the experiment transparent, pre-registering experiments in public repositories, and replicating prior experiments (Nosek et al. 2015: 1422, 1425, Christensen et al. 2019).[136] These are inarguably worthwhile considerations

[136] The idea of open science has a long history stemming back to when scientists first began sharing resources and communicating findings in journals. More recently, the term has been used to refer to proponents of an open research culture as encapsulated by The Transparency and Openness Promotion (TOP) Committee guidelines, spearheaded by the Center for Open Science (https://cos.io/). The TOP guidelines focus on journals' procedures and policies for publication, but in so doing, they also highlight basic principles that purportedly "translate scientific norms and values into concrete actions..." (Nosek et al. 2015: 1423; also see Dunning 2016, Christensen et al. 2019). The guidelines include standards that can be distinguished into three areas: those that reward researchers for engaging in open practices (i.e., citing data and materials,

(also see Elman et al. 2020: Chapters 6-13); however, as will become clear, I offer a cautionary perspective on pursuing open science principles too far. Social scientists need to constantly evaluate whether their beliefs about what leads to good science *actually* results in good science in practice. This is vital as part of the cultural authority of science comes from its collective, diverse nature and practices (e.g., peer review, opportunities for criticism), and an avoidance of a hidebound commitment to certain methodologies or practices (Oreskes 2019).

How Do We Ask Questions?

Experiments provide knowledge only when they offer insight into important questions. The scientific method starts with a question, but of course it also matters how one arrives at the question in the first place. How that happens goes to the heart of vigorous debates about the nature of science itself, including varying epistemic cultures (Knorr-Cetina 1999) and the reality that researchers' interests and values underlie all scientific work (whether consciously or not) (National Academies of Sciences, Engineering, and Medicine 2009). These broader issues are beyond my purview – instead, I offer a discussion of ways that social science experimentalists arrive at the questions they ask and ideas they have. It is incomplete, a la Oliver's (2004) fascinating book *The Incomplete Guide to the Art of Discovery*, which begins with: "no matter

---

replication), those that allow for reproducibility and evaluation (i.e., posting data, code, and materials in a trusted repository; having independent verification of analyses; having design transparency standards for review), and those that address values resulting from pre-registration (i.e., preregistration of the existence of the study, pre-registration of analysis plans). I discuss several of these ideas in what follows.

how well we prepare and plan, no matter how carefully we make our decisions and guide our actions, fate still seems always to play a central role in our careers and our lives" (ix).

Upon reading this quote, I reflected on how I arrived at some of my own research foci – particularly regarding partisan polarization among the public. I realized that it (roughly) sequentially involved: serving as an editor on an early article on the topic (Iyengar et al. 2012); offering comments on a book manuscript that generated a conversation (Levendusky 2013) (which led to the a two-step communication flow experiment described below); attending a meeting and ruminating on an off-handed remark about measurement from a colleague (that led to a large scale collaboration with that colleague and others); chatting with a student who happened to sit next to me at a conference dinner (that led to expanding the just mentioned collaboration to address other questions); spending a summer implementing failed efforts to pilot stimuli (to alter partisan's animus towards the other party); and taking advantage of an unexpected, albeit highly unfortunate, event (the COVID-19 pandemic) to test a theory (see Geddes 2003 on the role of world events in shaping research). That is a lot of happenstance.

At the most basic level, experimentalists constantly should be asking questions about every event they observe. An experimental approach leads one to realize the difficulty of documenting a causal relationship, and thus, it becomes almost automatic to question and contemplate relationships in the world – and their counterfactuals. Even though she is not discussing experimental work, Geddes (2003: 29) captures this dynamic in stating " good research in the field is more often motivated by curiosity about the world and intuition about cause-and-effect relationships…" Beyond this point, I provide a non-exhaustive, non-exclusive list of how to arrive at questions, captured in a three-fold categorization with the acronym *ASK*: *assessing, socializing, and kaput*.

Assessing occurs as we witness the world around us: watching people, reading the news, thinking about behavior, etc. Most social scientists enter the field for a reason or a passion based on personal assessments of the world and that can generate questions for study and experimentation. Another aspect of "assessing," beyond daily observations and personal interests, is more academic. Specifically, scholars reflect on "big questions," such as why and how people form preferences and behave the way they do, and what that implies for social, political, and economic systems. These "big" questions sometimes lead experimentalists to a particular study, but even when not, scholars should step back and consider how their work (regardless how they arrived at the topic) speaks to these larger issues. At a more mundane level, scholars read and assess the status of extant academic literatures, finding gaps and thinking of theory and experiments to fill them. In so doing, it is often useful to revisit classic works on a topic, though, do not hesitate to question such work – it is often useful to feel "indignation, annoyance, and irritation" with what has been written and these emotional reactions can generate new research (Geddes 203: 30).

More pointedly, I recommend an approach in which researchers design new experiments by building on prior designs. For example, many of the early framing studies, including the campaign finance study discussed in the earlier chapters, relied on the material used in Nelson et al.'s (1997) seminal framing experiment. These works sought to expand on Nelson et al. by replicating parts of it and then extending it by, for example, varying the source of the frame (Druckman 2001), or the number of frames offered (Sniderman and Theriault 2004, Chong and Druckman 2007). The point of replication here was not to challenge or question Nelson et al., but rather to expand on their highly credible work to isolate boundary conditions. As I will discuss

below, that Nelson et al.'s willingness to share their material shows how transparency can spark a robust literature.[137]

A second way to arrive at what questions to ask involves socializing, which can occur in a host of guises. For example, professional conferences exist, in part, to provide opportunities to discuss work with those who have similar and different perspectives. This can stimulate questions and ideas; sometimes this occurs via a formal setting such as a panel but often it occurs via informal conversations. For me, as mentioned above, an entire research agenda began due to a colleague who said in passing – literally as she was walking out of the room during a meeting break – that she has no idea what people thought about when asked to assess political parties (although she actually had ideas that were different from what prior research suggested). That comment sparked further conversations and several studies to explore what people have in their head, particularly when they express animus towards opposing partisans (e.g., Druckman et al. n.d.). This is only one of many such examples from my career. The first professional paper I ever wrote reflected a similar off-handed question that a professor of mine asked during a class break (Druckman 1996), and the main research agenda for the first part of my career came about due to ruminations at a dinner with my future collaborator (e.g., Chong and Druckman 2007). The lesson is to spend time engaging with colleagues as that often leads to questions that demand

[137] Another way to assess is to engage in exploratory research – that is, "an attempt to discover something new and interesting, by working your way through a research topic" (Swedberg 2020: 17). Here, empirical inquiry can help identify questions to pursue further.

exploration and experimental inquiry. This can occur in meetings, collegial conversations, or other contexts.[138]

This type of socialization also occurs in the classroom. Not only do students derive ideas from professors, but professors also get ideas from students. Students, graduate and undergraduate, often bring fresh perspectives to topics that can generate novel questions, and ultimately, experiments. This can occur from advising students. Graduate students can also generate ideas in collaborative group settings, such as a reoccurring lab meeting (see Druckman et al. 2018). Undergraduates often conduct research projects in small classes, which can lead to collaborations with professors (Druckman 2015). For example, multiple experiments I conducted with students stemmed from questions we discussed during class, such as how to measure drug usage among student athletes (Druckman et al. 2015) or how media affects vote choice (Druckman 2004). Further, preparing for class and observing students' reactions during class discussions can provide insight into useful research questions.[139] Undergraduate advising offers

[138] The importance of socialization as a way to generate research questions accentuates the role of equitable professional opportunities. If particular groups or individuals do not have access, they are essentially disenfranchised from a crucial way in which scholars identify questions to explore.

[139] A remarkable example of the former is Kahneman's (2002) autobiographical note about teaching early in his career: "To teach effectively I did a lot of serious thinking about valid intuitions on which I could draw and erroneous intuitions that I should teach students to overcome. I had no idea, of course, but I was laying the foundation for a program of research on

many opportunities. Students can provide novel research ideas, and an advisor can help them situate the question in an academic literature and design an experiment (when that is the appropriate method). Of course, in all these cases, advisors should be transparent about the goal of collaborating toward writing a professional article. In my experience, this regularly proves profitable for the student and for the pursuit of knowledge (and publication).[140]

Questions also frequently emerge from socializing in non-academic settings. During dinner conversations, friends and family may unknowingly raise novel questions not sufficiently addressed in the academic literature. Complete strangers have even probed me with questions that turned out to be useful. More directly, questions often emerge through relationships with individuals or organizations with whom a researcher may have a shard goal. Academic and

---

judgment under uncertainty." This program of research came to be one of the most impactful in the social sciences (e.g., Kahneman 2011).

[140] Selected examples from my experience include a framing experiment on campaign finance (Druckman and Nelson 2003), an experiment on framing versus cue taking (Druckman et al. 2010), an experiment on information search (Druckman et al. 2012), an experiment on the impact of elite polarization (Druckman et al. 2013), an experiment on irrelevant even effects (Busby et al. 2017), an experiment on pain perceptions (Druckman et al. 2018c), an experiment on incivility (Druckman et al. 2019), an experiment on climate change communication (Bayes et al. 2020), and an experiment on disability services (Druckman et al. 2020b). In each of the cases, the student generated the area of interest and general question and then we collaborated, sometimes along with a graduate student, to read the literature, derive hypotheses, design an experiment, etc.

practitioner (e.g., non-profits, campaigns) relationships are certainly not new, but researchers have increasingly leveraged them to identify pressing questions relevant to the academic's interest and the organization's mission, which can then be answered with an experiment. Such collaborations can occur informally. For instance, Gerber et al.'s (2008) well-known field experiment on how social pressure mobilizes voters came about through interactions with a campaign consultant who sought a cost-effective way to increase turnout (see Green and Gerber 2010). Collaboration can involve more formal relationships – as has been documented in various fields (see, e.g., Karlan and Appel 2016). For instance, Kalla and Broockman (2016) partnered with a liberal political organization, CREDO Action, with roughly 3.5 million members. They worked with the organization to design an experiment where members attempted to arrange meetings with members for the U.S. Congress (from one party) who had not already cosponsored a bill to ban a chemical. They randomly assigned the meeting inquires to vary whether the inquiry mentioned that those wanting to meeting were active political donors *or* concerned constituents.[141] They report some of the most striking evidence to date that donations affect access – senior policy makers made themselves available to political donors between three to four times more often than to concerned constituents. This experiment highlights the gains that can come about from an organizational-researcher partnership (also see the examples in Chapter 4).

[141] They specifically used block random assignment to ensure balance on environmental views, prior co-sponsorship of the bill in a previous Congress, years served, ideology, geography of the district office, and prior presidential vote share in the district.

Fortunately, the organization research4impact (https://www.r4impact.org/) facilitates this process by matching practitioners and social scientists (with a focus on experimentation). They also offer workshops on how to pursue research relationships; as they say on their website "Connecting with a practitioner will produce new questions, new research ideas, and possibly new formal collaborations…." (also see Levine 2021 on how to form organizational partnerships to run experiments). Even when these relationships do not go as planned and/or fail, they still lead to new ideas and lessons about what not to do next time (see Karlan and Appel 2016).

A third way to generate questions is what I refer to as "kaput," in essence, failure. In Chapter 1, I noted Campbell and Stanley's (1963: 3) statement that we "must instill in our students the expectation of tedium and disappointment and the duty of thorough persistence…" This pessimistic portrayal reflects the prevailing reality that many experimental efforts will have disappointing results and be initially interpreted as kaput. Campbell and Stanley are right that persistence is absolutely essential. Persistence includes learning from the failed results which can in fact, in the long run, lead to even better experiments (what Karlan and Appel 2016 call "learning when things go wrong"). Indeed, the first thing one should do when an experiment fails is to ask why it failed. It may be due to a problem with the design. For example, I once implemented a framing experiment focused on attitudes about urban growth initiatives. I exposed people to an argument that the issue should not be left to voters, but the frame had no effect whatsoever, and even backfired a bit. The reflection that followed the null result led me to recognize the importance of assessing the argument strength of different frames, a topic that led me to consider a new theory and several distinct experiments (e.g., Chong and Druckman 2007).

Aside from theoretical missteps/inaccuracies, an experiment can fail for countless implementation reasons, including distracted participants, poorly constructed stimuli or

measures, insufficient power, non-compliance or attrition, and related issues. Each of these design elements, as discussed in Chapter 2, require careful attention and consideration should an experiment not go as planned. An additional possibility concerns events happening outside the experimental context. One glaring example comes from an experiment I implemented regarding the way different arguments affect Americans' support for nuclear energy. The data collection took place in March, 2011, which as it turned out overlapped with the Fukushima Daiichi nuclear disaster in Japan (due to an earthquake and tsunami). This obviously undermined the data collection, as any argument for nuclear energy became impotent in the midst of the disaster. But the scenario begged the question whether I should include questions about the information environment to measure how the context moderated treatment effects (see Druckman and Leeper 2012b).

Distinct from contextual sources of failures are unexpected experimental participant reactions. In one case, I attempted to embed a survey experiment in an Election Day exit poll in Minnesota. Suffice it to say, Election Day in Minnesota means cold and snow. Given these conditions, I thought it may help to offer a $5 incentive to take the survey. Much to my surprise, however, the incentive turned people away. Apparently, the social capital is so high in Minnesota that the response rate was higher when the student pollsters did not offer the $5. I stopped offering incentives, mid-day, once it became apparent everyone was declining the money and subsequently declining to participate. The subjects seemed to feel they were contributing to the public good by taking the survey and helping students, but when money was involved, they felt insulted by my assumption they would demand pay. In another example, I sought to use experimental vignettes to assess whether athletic trainers display racial bias in pain assessments of student-athletes. In the recruitment e-mail, we referred to potential participants as "trainers."

Dozens responded that they would not participate, feeling insulted that we had not used their official titles of "Athletic Trainer" or "Certified Athletic Trainer." We apologized and adapted the language accordingly (Druckman et al. 2018c). These examples accentuate the importance of anticipating responses of participants in experiments, particularly when using targeted populations, which is an increasingly common experimental approach (e.g., Klar and Leeper 2019). It is important to talk to members of the groups about the study before you start in order to anticipate the issues that may arise. Pilots are not just for researchers to test manipulations, but also to assess response rates and unexpected responses to the study materials.

And sometimes an experiment goes kaput due to the reality that the method does not fit the question. For instance, a collaborator and I hoped to causally assess the impact of animus towards the other party on the formation of issue positions; we sought to do so by manipulating participants' levels of contempt for the other party (e.g., making Democrats dislike Republicans more) to see if increased contempt led partisans to follow party cues more on issues. We piloted nine treatments we thought could prime out-party animus and every one failed (perhaps due to a ceiling effect). We concluded an experiment would not work for this test and instead kept searching for other possibilities, one of which came about with the onset of the COVID-19 pandemic. There we used a prior survey measuring partisan animus to gauge COVID-19 attitudes, thereby allowing us to use panel data with an exogenous measure (i.e., one taken prior to the emergence of COVID-19) that linked out-party dislike to partisan cue-taking. The pilot failures prompted us to ask what method could work and we took advantage of an opportunity that naturally (albeit very unfortunately) arose (Druckman et al. 2020a).

In sum, experimental failures can be dispiriting, to say the least, but they are also inevitable. In the moment, disappointment, anxiety, and frustration constitute the natural

responses, and every failure might not straightforwardly lead to another question that works out well. Yet, in most cases, researchers can learn lessons that can advance theory, improved implementation, or new methods. In Campbell and Stanley's words, the key is to persist, even when incredibly difficult, in the face of disappointment. Success will come with diligence and innovation. Or as Oliver (2004: 34) states, the "path of science is strewn with failures of what once seemed great ideas and promising new directions... Science is built upon past failures, as well as upon successes, and a select few of those with the boldness and daring to depart from convention will always lead the way."[142]

Summary

At first glance, it may appear that arriving at a good research question should not be difficult; yet, it requires a researcher to identify an "important question" relevant to their field, and which can be realistically addressed to add to an accumulating knowledge base. It is thus not so easy. It necessitates maintaining a curiosity at all times as questions can emerge in the most

[142] That said, in all fairness, it is easier to bold and endure failure for established researchers with sizeable budgets. For early career scholars, an experimental failure may be more problematic due to a lack of resources to pursue another direction. It is here where mentoring plays a vital role to guiding earlier career scholars to make research investments in anticipation of what they would be able to do should the results not fulfill expectations (and how to assess risk in light of budgetary realities). It also accentuates the need for research funding programs targeted for early career scholars.

unexpected of settings – in essence, many experiments and findings come about due to a mix of luck, curiosity, and adherence systematic thinking (Dunbar and Fugelsang 2005).[143]

The flip side of the role of luck is that one might often feel that their diligence seems unrewarded. Diligence acts as a necessary but not sufficient ingredient in leading to a satisfying experiment. At times, experimentalists even have to let go and move on from an experiment, but do not view this as wasted time (it is part of the inevitable process). Those who maintain curiosity with at least a sliver of irreverence – they do not accept any answer as the final word – will arrive at exciting questions and success. Rather than summarizing the lessons of this section with numbered points, as with many other sections in the book, I end with the below table. The table outlines the examples of my *ASK* framework (which is not meant to be comprehensive) for arriving at questions for experimentalists.

Table 5-1: *ASK*: Examples of How to Ask Research Questions

Source of the Question	**Examples**
Assessing the world and the field.	• What questions arise from witnessing how the world works? What questions are you passionate about answering? • What are the "big" questions in the field? What questions is the existing literature addressing? What questions can be addressed by extending or generalizing other experiments?

[143] While not related to experiments, one telling (and also remarkable) story comes from Oskar Morgenstern wondering into a library and arbitrarily perusing a book which provided a crucial insight for his and von Neumann construction of game theory (Morgenstern 1976). He states (1976: 811), "I note a curious incident that shows how chance can influence the direction of scientific work."

Socializing with colleagues, teachers/students, and non-academics.	• What questions arise at professional meetings (e.g., on panels)? What questions emerge in informal conversations? What do your colleagues ask? • What do your teachers, advisors, colleagues ask/suggest? What are students' reactions? What questions do your classes ask about research findings? What questions interest students? • What questions arise in conversations with family and friends? What questions interest practitioners? Are there questions that intersect with research interests and practitioner interests?
Kaput or a failed experiment.	• Was there something wrong with the theory? • Was there something wrong with the implementation, design details, the context? Did something go wrong with the participants? • Was the best method being used?

Formalizing the Research Process

Implementation of a sound experiment requires meticulous attention to detail when it comes to identifying the best stimuli, measures, sample, and context. Yet, despite how much time researchers spend thinking through each piece of an experiment, they inevitably will forget the details, even if it is hard to imagine in the moment. For this reason, researchers need to keep careful records of all decisions (Geddes 2003: 43-88). As an example, I will discuss Druckman et al. (2018b) in detail. The experiment tests whether partisan media stories can influence people who do not watch them. This can occur via a two-step communication flow: (1) a story (e.g., about the environmental risks posed by drilling for oil) affects those who watch it (e.g., making them more opposed to drilling), (2) those who watched and had their opinions changed talk about

the issue with others who did not watch (e.g., discuss drilling), and (3) those conversations influence the people who did not watch and their opinions end up resembling the opinions of those who did watch (e.g., those who did not watch are then indirectly shaped by the media, becoming opposed to drilling) (Lazarsfeld et al. 1948). In the experiment, the authors randomly assigned people to a control group (no watching or discussing the news/issue), a group that watched partisan news but did not subsequently discuss it, a group that watched the news and then discussed it (with others who did not watch), or a group that did not watch the news but discussed the topic with those who did watch it. They demonstrate that individuals' who did not watch the news but discussed it had opinions that matched those of people who just watched it (and differed from those in the control who neither watched nor discussed). Put another way, partisan media can influence the opinions of people who do not watch it via a two-step communication flow.

A sampling of the design decisions included (1) choosing the news networks (e.g., MSNBC and Fox News), (2) deciding whether participants should be forced to watch one network or be able to choose which networks to watch (e.g., MSNBC, Fox News, PBS), (3) deciding on the medium of exposure (e.g., videos), (4) identifying the issue (e.g., drilling), (5) crafting the precise messages (e.g., focusing on environmental risks or economic gains from job creation), (6) determining the size of the discussion group (e.g., four people) and the composition of the groups (e.g., all from one party, mix from different parties), and (7) deciding on the precise outcome measures (e.g., support for drilling, partisan identity strength). The authors also made decisions regarding whether the networks should show incongruent stories (e.g. MSNBC focusing on economic gains from drilling), whether there should be conditions with news exposure and no discussion, whether there should be conditions with discussion but no news

exposure among anyone, and whether attitudes on the outcome variables should be measured pre-and post-treatment. We made other choices about the recruitment plan (e.g., from campus, community organizations), the random assignment approach (e.g., what if the plan is to have a mixed partisan group but only those from the same party arrive at the session?), and the analysis plan (e.g., which groups to compare to test the hypotheses).

While this admittedly constitutes a particularly complex experiment, the reality holds that even for the simplest of experiments, a large number of choices need to be made. It is crucial, at each stage, to explicate and write down the rationale. For example, in the case of this experiment, the authors created a document to justify the choice of drilling as the issue to study. The authors selected drilling in order to build on prior work on partisan reasoning (Levendusky 2010, Druckman et al. 2013), and to use real television segments to increase experimental realism (since they were more captivating than false stories likely could be). Drilling was also an issue on which partisans did not hold such strong priors so as to be un-moveable from media exposure.[144] While these types of details demand substantial attention in the moment, they are easily forgotten without documentation. In the case of this experiment, the "design" document filled 28 single-spaced pages, beginning with the motivation and including details about the stimuli, outcome measures, predictions, analysis plans, and more.[145]

[144] This point may seem, at first glance, to be a bias in favor of finding an effect but the rationale is that the study's goals concerned whether a two-step communication can occur and this requires initial media influence (or else there is no possible "two step").

[145] The document also included pretest analyses to ensure the segments on oil from the different networks (i.e., MSNBC and Fox) were perceived as being in the anticipated direction (i.e.,

This goal of the two-step communication flow experiment was to test a theory and identify patterns, rather than inform policy. If one is implementing an experiment with the goal to inform those who make policy directly, then a design document might also include a discussion of Al-Ubaydli et al.'s (2017, 2020a,b) aforementioned scalability considerations (see Chapter 3). For example, be clear to articulate the target policy-relevant environment such as where an intervention would be implemented (and what intervention), what actors would be involved and what are their incentives, and, as a general matter, "backwards induct to address the potential threats so scaling with the experimental design" (Al-Ubaydli et al. 2020a: 37). That is, think through how the results would translate to a precise policy situation and what challenges would arise in the process.

Documentation continues once the data are collected. Experimentalists must temper their excitement of finally analyzing the data and make sure to maintain documentation of how they conduct analyses. Specifically, this involves keeping careful notes on data collection (e.g., were the sessions run without incident?), analytical decisions such as the creation of variables (e.g., in the aforementioned experiment, it involved aggregating three items based on a factor analysis), the selection of cases (e.g., excluding pure Independents), and analysis decisions (e.g., using inverse probability weighting based on group assignment). Without documentation, by the time one decides upon the most accurate and informative result presentation, he or she may have scant memory of how to recreate it out of a mess of poorly named variables and many analysis files.

---

against and for drilling), and were equally persuasive in terms of logic (so that one segment did not exhibit a strong effect for reasons of argument strength which would have been a confound).

I advocate for writing these planning and analysis documents because they facilitate multiple goals. For one, they make it much easier to write the actual paper: the documents form part of the paper. The documents also facilitate satisfying transparency requirements that come with the publishing process. Open science's transparency expectations involve experimentalists posting their stimuli, outcome measures (e.g., surveys), data collection details, data, and analysis code in independent public registries (Miguel et al. 2014, Nosek et al. 2015: 1424, Chistensen and Miguel 2018, Aczel et al. 2020, Boudreau 2021). Transparency ensures an understanding of the scientific processes underlying empirical claims (Lupia and Elman 2014, Elman et al. 2018). The principles can be found, for example, in political science's Data Access & Research Transparency (DA-RT) initiative; Lupia and Elman 2014) that stimulated revision to the American Political Science Association's (APSA) *Guide to Professional Ethics in Political Science*. That guide (2012: 9-10) states that "researchers have an ethical obligation to facilitate the evaluation of their evidence-based knowledge claims through data access, production transparency, and analytic transparency so that their work can be tested or replicated." A large number of social science journals, including most that publish experimental research, endorse these guidelines by requiring authors to post material and data (Boudreau 2021). While scholars using other methods (e.g., qualitative approaches) debate the application of these principles to their work (e.g., Monroe 2018), experimentalists generally endorse the basic ideas. Moreover, experimentalists have gone further by generating a list of reporting expectations in papers, including participant eligibility and recruitment, number of participants, descriptive statistics, evidence of treatment delivery, and weighting procedures if used, among other considerations

(Gerber et al. 2014).[146] The life of an experimentalist will be much easier if he or she diligently keeps notes on all of these aspects through the process rather than having to go back and re-create it later. While this may sound easy and obvious, it, again requires self-discipline to slow down and document.

These transparency expectations allow other researchers to verify the results and confirm the robustness of the findings. Transparency also helps other researchers replicate and extend the experiment (e.g., using the materials), and use it for large-scale re-analyses and meta-analyses.[147] Indeed, Mullinix et al. (2015), Coppock et al. (2018), and Coppock (2019b), all discussed in Chapter 3, re-analyzed and replicated large sets of experiments because they had access to those data and materials via the TESS program (which posts all material and data). Zigerell (2018) accesses 17 previously conducted (TESS) survey experiments to perform a new analysis of racial prejudice; he reports a distinct and provocative finding that for "White participants..., pooled results did not detect a net discrimination for or against White targets, but, for Black participants…, pooled results indicated the presence of a small-to-moderate net discrimination in favor of Black targets" (1). These efforts and concomitant knowledge gains are only possible through transparency. That other scholars build on and refine prior experiments also speaks to

[146] Some of the reporting recommendations have generated debate, particularly regarding whether one should report balance tests by comparing measured covariates across experimental conditions to assess the success of randomization (c.f., Gerber et al. 2015, Mutz and Pemantle 2015).

[147] In the most extreme cases, it could allow for the identification of fabrication (Christensen and Miguel 2018: 945).

the importance of documenting the design decisions. That is, another researcher building on one's work may inquire about an aspect of the experiment and it is much better to access an old document that detailed the thinking at the time than to rely on one's memory. (Indeed, I had to access the aforementioned 28 page document for virtually all the details described earlier concerning the two-step communication flow experiment.) In essence, this helps ensure your work has more impact since helping other scholars build on work perpetuates its influence. As mentioned earlier, one way to stimulate a new experimental idea is to build on and extend prior work, to learn about the details from others, and to proactively provide details to others to ensure collective intellectual progress.

As data transparency becomes the norm, experimentalists should attend to how that could affect results. Specifically, Connors et al. (2019) point out that even small contextual changes can alter participant behavior; with transparency rules, many Institutional Review Boards now require that participants be informed of a statement such as: "if this research is published, your response will be made publicly available to other researchers. You will never be asked for your name. Only the responses you give to the questions that follow will be available for download" (Connors et al. 2019: 191). Connors et al. conduct an experiment comparing responses of individuals who receive this de-identified posting notification against those who do not.[148] They find the disclosure significantly alters responses. For example, those notified report lower levels of political knowledge (e.g., they are more likely to report they do not know answers to

[148] They also explore the effects of, in addition to the posting notification, explaining the data will be made available to reduce the risk of scientific fraud. However, the added statement has no additional effect beyond the basic disclosure.

questions), more support for abortion (i.e., more pro-choice attitudes), and increased levels of self-reported vote turnout.[149] Thus, informing respondents of data transparency can have positive effects on data quality (e.g., more accurate knowledge reports) but also exacerbate social desirability reporting (e.g., over-reporting turnout). Scholars conducing experimental work that involve measures where respondents may think about how responses reflect on them should thus account for such possible effects, perhaps by considering measurement methods that reduce social desirability bias (see Chapter 2's discussion, Rosenfeld et al. 2016).[150] This is not to say

[149] The results are conditioned by self-monitoring scores; for example, in the treatment condition, low self-monitors (i.e., those less worried about presenting themselves in a way to impress an audience) are more likely to answer "don't know" to the knowledge questions (Connors et al. 2019: 196-197), and more likely to express pro-choice attitudes (Connors et al. 2019: 199). Further, high self-monitors over-report their income when told of the disclosure (Connors et al. 2019: 200-201).

[150] In discussing transparency, two other points are worth keeping in mind. First, there are the well-acknowledged privacy concerns. Some data cannot be fully posted without comprising the anonymity of the respondents. For example, data from an audit study of state legislators that include detailed information about each legislator would make it easy to identify whether particular legislators responded. This, in turn, becomes problematic when anonymity is essential to conduct a study. Even de-identified data sometimes allow motivated third parties to identify particular respondents (Christensen and Miguel 2018: 969). In these scenarios, the solution involves not fully posting data and making it available via request and assurance from the requestor to protect anonymity (i.e., an application for restricted data access). This is the

scholars should forgo transparency with research participants, but rather that they should be cognizant of its possible effects.

In sum, conducting an experiment involves a myriad of decisions; formalizing a process to keep track of choices, ensures clear documentation, lessens the extent to which one has to rely on his or her memory, provides text for papers that eventually may be written, facilitates compliance with transparency publication guidelines, and helps in conversations with others who share similar interests. It also can provide some perspective by forcing an experimentalist to

---

approach, for example, taken by the American National Election Studies; see https://electionstudies.org/restricted-data-access/. Second, some have politicized data transparency to undermine the application of credible scientific research. For example, in 2019, the Environmental Protection Agency appealed to the need for data transparency as an explanation to ignore widely agreed upon credible scientific data that include personal health information (thereby making it un-releasable) (Friedman 2019). This action increased the difficulty of maintaining and enacting clean air and water rules that reference evidence on the effects of pollution from the confidential data. Various scientific organizations protested; for example, The National Center for Science Education said "ruling out studies that do not use open data 'would send a deeply misleading message, ignoring the thoughtful processes that scientists use to ensure that all relevant evidence is considered'" (Friedman 2019). Here a sound scientific principle that has some caveats concerning data privacy is being used to promote an ostensible political agenda (de-identifying the data would be possible but would cost hundreds of millions of dollars) (Frideman 2019). Scientists themselves need to communicate about transparency in effective ways to prevent undermining the application of science (e.g., Ioannidis 2017: 107).

think through each choice and recollect the rationale as he or she reviews prior decisions at each stage. All of that said, this entails a substantial time investment – the reality is that quality experiments typically do.

Given the current extent to which transparency expectations are formalized in the publication process, academic norms need to change. Tenure standards must evolve to recognize that quantitative productivity may be vitiated due to the increased time commitments needed to ensure transparent research. Indeed, a generation ago, one would not have had to spend time creating detailed appendices about the implementation of an experiment, or make user-friendly versions of the experimental material and data. These are non-trivial tasks for which institutions must account when assessing an individual's research. In essence, providing these materials helps provide a public good since other researches can build on them and that needs to be recognized.

Pre-Registration and Pre-Analysis Plans

Another item one should consider before implementing an experiment is to pre-register it. Pre-registration entails registering a study in an independent repository prior to data collection (Nosek et al. 2018, Christensen et al. 2019: 99-119). Authors provide enough details so that other scholars know what hypotheses have been tested on what populations (Malhotra 2021). Prominent registries include the Open Science Framework (OSF), Evidence in Governance and Politics (EGAP), and AsPredicted (Boudreau 2021). These repositories provide the scientific community with access to the universe of studies on a given topic.

If practiced by all or most scholars working on a given topic, pre-registration can help vitiate publication bias which occurs when publication decisions are based on criteria unrelated

to research quality.[151] The most notable type of publication bias, the file drawer bias, occurs if a positive test result (i.e., a statistical test that rejects the null hypothesis of no effect) is more likely to be published than a negative test result (i.e., a test that does not reject the null hypothesis) (Rosenthal 1979).[152] The published record of research skews from the true distribution of results, overstating the collective strength of positive or statistically significant findings. For example, if one out of ten studies shows that sending varying types of health-related text messages leads people to eat less fatty food, and only that one study is published, the result is a mis-portrayal of the effect of text messages.

[151] My definition of publication bias is broader than others, which focus on a bias that occurs when the publication decision depends on the result (e.g., Brown et al. 2017: 94). Implicit in these definitions, however, is the idea that the result drives decisions, all else constant. As Malhotra (2021) clarifies, publication bias occurs when "when the statistical significance of findings influences publication probability *conditional on the quality of the study design*" (italics in original). A broader definition allows for inclusion of other types of biases such as decisions based on the identity of the author or sponsor of the study. For an assessment of various types of biases, see Fanelli et al. (2017).

[152] The focus is typically on the peer-reviewed scientific literature (Brown et al. 2017: 93), rather than the so-called gray literature that includes conference papers, dissertations, etc. Rosenthal (1979: 638) states "studies published in the behavioral sciences are a biased sample of the studies that are actually carried out." The metaphor is that many studies end up in the file drawer rather than academic journals.

There exists a large literature documenting publication bias in the social sciences (e.g., Gerber et al. 2000, Gerber and Malhotra 2008, Gerber et al. 2010, Franco et al. 2014, Brown et al. 2017, Fanelli et al. 2017, Christensen and Miguel 2018: 924-931, Andrews and Kasy 2019, Christensen et al. 2019: 31-55, Berinsky et al. 2020, Malhotra 2021). Pre-registration of studies serves as an antidote to the file drawer bias problem.[153] Scholars who conduct meta-analyses to aggregate literatures can include unpublished and published experiments on a topic to assess the robustness of an effect (assuming data from unpublished studies are also made available).[154] For example, even if the aforementioned nine text messaging studies with null effects were not published, pre-registration along with data availability would ensure a more accurate synthesis of the literature. Pre-registration in that sense serves as a public good, which should not add much additional work if one writes the planning documents discussed above.[155]

[153] See Malhotra et al. (2021) for a discussion of various other solutions (also see Andrews and Kasy 2019).

[154] Pre-registration is a norm/requirement in other fields, most notably when it comes to medical clinical trials (for discussion, see Humphreys et al. 2013).

[155] Coffman and Niederle (2015: 90-91) suggest downsides of pre-registration include researchers revealing their plans before the project is completely finished – this requires consideration of the tradeoff of maintaining an up-to-date registry against ensuring privacy for a particular amount of time. They also mention that sometimes the lack of publication from a registered study may be due to budget challenges rather than null results, and more generally, a registry needs to be organized in a useful and searchable manner. These are all important

Many pre-registrations, including the aforementioned registries, also ask researchers to report the hypotheses, outcome measures, experimental conditions, sample size, plans to exclude outlier data, and a detailed *pre-analysis plan* of how they will test their hypotheses once they finish data collection. Such an analysis plan requires specific descriptions of the statistical tests that will be employed and what evidence would constitute support for a hypothesis. Pre-analysis plans supersede basic registration with the intent to vitiate data mining (p-hacking), and/or push researchers to justify why they deviate from the plan. Some argue that pre-analysis plans enhance the credibility of research (Humphreys et al. 2013, Nosek et al. 2018).

The planning documents discussed above envelope a pre-analysis plan, and serves a crucial function by forcing the experimentalist to visualize the data before collecting it. This can prevent a design error where the data collected do not allow for the testing of the hypotheses. For example, in the two-step communication flow experiment, so many decisions had to be made concerning the treatment conditions that it may have been easy to forget which condition comparisons constitute the key test. In that case, the main comparison is between those who watch the news and do not discuss it against those who do not watch but discuss it (i.e., do the conversations in essence "stand in" for the watching?). The pure control group (where individuals neither watch nor discuss) is not central. A pre-analysis plan ensured attention to identifying the appropriate analyses and the need for particular experimental conditions. Echoing Chapter 2 it forces experimentalists to carefully assess the relevant counterfactuals being tested,

practical points, as is the reality that some scholars may be unaware that a given journal requires pre-registration but that could be addressed by allowing scholars to pre-register upon submission.

collect the appropriate data and have a plan on what they will do with the data upon receiving it (also see Humphreys et al. 2013: 10)

That said, developing a pre-analysis plan and even formally registering should not hamstring discovery. This can occur when scholars insist on a hierarchical distinction between confirmatory/predictive and exploratory/postdictive hypotheses where the former appear in the pre-analysis plan and the latter do not (Nosek et al. 2015: 1423). While not always explicit, it seems that many voices in the field value *a priori* hypotheses over exploratory ones.[156] This preference follows for three reasons. First, a confirmatory hypothesis appears to be further along in the scientific process – it means a question has been asked, a theory developed, a hypothesis generated, and data collected. In contrast, exploratory work could be construed as more "theory generation," and thus earlier along in the progress of science. Second, some construe exploratory findings as requiring subsequent research (Nosek et al. 2018: 2604-2605), with more experiments/data collection needed before publication. Third, exploratory findings that are not pre-registered could be viewed as post hoc and any explanations to explain them construed as rationalizations with less credibility (Nosek et al. 2018: 2600-2601). The concern is that analysis not specified in advance could be inauthentic, with authors engaging in either questionable research practices, "p-hacking," or "fishing" to find significant results (Simmons et al. 2011,

[156] It is explicit in the aforementioned TOP guidelines that state the ideal journal "requires pre-registration of studies with analysis plans and provides link and badge in article to meeting requirements" (Nosek et al. 2015: 1424). It is worth noting that what constitutes confirmatory as opposed to exploratory hypotheses is not always clear (Devezer et al. 2020).

Humphreys et al. 2013).[157] These behaviors refer to the inappropriate analysis and re-analysis of data in order to find statistically significant results (e.g., excluding relevant conditions, outcome measures, or data points until significant results are found) (see Franco et al. 2015, 2016, Zigerell 2017).[158]

These are reasonable concerns that formal pre-analysis plans address by forcing researchers to make non-registered choices explicit, since hypotheses would be clearly delineated as confirmatory or exploratory (Coffman and Niederle 2015: 82). Yet, in assessing the value of pre-analysis plans, one must draw a distinction between the research process and the publication process. As explained, it constitutes a vital part of the research process but its role in the publication process comes with pros (e.g., addressing p-hacking) and cons. The cons occur if authors, reviewers, and editors strongly, and sometimes blindly, privilege confirmatory over exploratory results even with the just discussed arguments for why one may do so (e.g., Olken 2015, Christensen and Miguel 2018: 958-959).[159]

---

[157] Similar terms include data snooping and data butchery.

[158] A related concept is the "garden of forking paths" (Gelman and Loken 2014) where researchers, often unconsciously, make particular analytic decisions that lead to an outcome that may not sustain if other reasonable decisions had been made (and could have been made to ensure robustness).

[159] Another institution makes this norm explicit by instituting results-blind review where journals conditionally accept articles prior to data collection or without disclosure of the data collection, based on the theory and design along with the pre-analysis plan. The journal publishes the article, assuming the author(s) follows the data collection plan, regardless of the statistical

To see why, consider three downsides of strict adherence to pre-analysis plans in the publication process.[160] First, even when hypotheses are innovative and high-quality, inattention to careful data collection can lead to null results. Over-emphasis on pre-analysis plans shifts the basis of publication decisions toward the existence of *a priori* hypotheses and away from using statistical significance. If science operated ideally and research quality did not vary, this change would be an obvious improvement. Alas, as has hopefully been made clear in the prior chapters, conducting an experiment and collecting quality data involve difficult processes. A sampling of considerations includes ensuring meeting the assumptions for causal inference (e.g., excludability, SUTVA), the use of validated treatments (e.g., via manipulation checks), high levels of experimental realism, accurate and valid outcome measures, sufficient subsample sizes if moderators are to be tested, etc. Many of these considerations concern issues *during data*

significance of the results (e.g., Nyhan 2015, Findley et al. 2016). The *Journal of Experimental Political Science* offers this option. They have an initial review stage prior to the presentation of results. Then at the second review stage, the article is accepted if: 1) the research question and rationale for hypotheses did not change; 2) the experimental procedures were followed closely and any departures are noted and justified; 3) unregistered post hoc analyses are clearly labeled, justified, methodologically sound, and informative; and 4) conclusions are justified by the data (including considerations of data quality).

[160] Devezer et al. (2020) also point out that nothing prevents a pre-analysis plan from including problematic statistical procedures while giving a "veneer of rigor" (14).

collection.[161] Even the most clever and well-developed hypothesis does nothing to ensure the proper implementation of a quality experimental data collection. In contrast, the existence of a statistically significant result often indicates not only a strong theory or intuitive expectations but also sufficient statistical power, well-designed treatments and outcome variables, and carefully executed data collection (Malhotra 2021) (e.g., it is difficult to obtain significance in light of all the errors that could occur during data collection).[162] This is not to dismiss the concern of data drudging/p-hacking that sometimes to leads to statistical significance. Instead, the point is that, in the extreme, if one puts emphasis on pre-analysis plans, the result could be the publication of a lot of null results due to poor theory, designs, and/or implementation.

The answer of which publication decision criterion is preferable depends on how one defines quality. For example, Nosek et al. (2018: 2602) focus on reproducibility, whereas others might focus on the aforementioned criteria (i.e., that define quality data collection such as experimental realism, valid measures, piloted stimuli, etc.). Determining the best approach

[161] Another institution is a standard operating procedure: a statement from a lab of default practices for handling certain types of issues, such as how to handle attrition, how to define noncompliance, and whether to exclude subjects who state that they discerned the purpose of the experiment (Lin and Green 2016). This practice could, in theory, address some of these concerns if it is faithfully followed but it is not clear it can cover all the issues that could arise.

[162] That said, there are aspects of experiments that could enhance the likelihood of statistical significance, particularly demand effects. I suspect though that these are dwarfed by the various elements that make it difficult to achieve significance; moreover, the extent of demand effects remains unclear in many areas (e.g. Mummolo and Peterson 2019).

requires an empirical assessment which, to date, has not been conducted (Nosek et al. 2018: 2602). Related to this, Coffman and Niederle (2015: 84-88) suggest the benefits of pre-analysis plans in terms of limiting false positives (e.g., due to p-hacking) accrue much more in research domains where there will only be a single test of a hypothesis (e.g., an expensive, resource intensive field experiment). In such cases, there are not follow-up studies that could correct for false-positives or positive results from data drudging. These types of studies also tend to be ones that have policy goals and as such benefit from a careful consideration of the aforementioned scalability issues in the process of thinking about the analyses. In contrast, in research areas where we might expect several experimental tests that can identify the strongest theory, then protection against false positives or data drudging may be less crucial. Hence pre-analysis plans should perhaps receive less weight than statistical significance. In short, the extent of the advantage of pre-analysis plans in terms of the aggregate scientific record depends on the area of research.[163]

[163] Further, the extent of p-hacking appears to vary across disciplines (Coffman and Niederle 2015: 83-84) and method (Brodeur et al. 2020). Thus, the usefulness of pre-analysis plans as a way to limit p-hacking may be contingent on discipline (e.g., it appears to occur much more often in psychology than economics) and method (e.g., it appears to be less common with random assignment experiments). Indeed, for particular types of experiments, p-hacking may be extremely unlikely due to sheer expensive. Consider a survey experiment on a probability sample with of 2,000 and 12 items (i.e., treatments and measures) costs roughly $18,000 (https://tessexperiments.org/html/limits.html). And 12 items does not provide a lot of leeway for data drudging. Finally, there is nothing inherently problematic about publishing null results sans

A second downside with strict adherence to pre-analysis is it ostensibly assumes that any exploratory data analyses reflect post-hoc theorizing, therefore requiring further data collection. Yet, the social sciences include many approaches and theoretical frameworks, some of which may not have been considered prior to data collection. For example, Druckman and Shafranek (2020) present an audit experiment where they sent requests to undergraduate admissions offices to explore biases in responsiveness. They varied the race of the sender (i.e., African-American/White) as well as political referents in the requests (i.e., references to a civic organization, a political organization, the Young Democrats, or the Young Republicans). The authors had designed the study to assess the independent effects of racial and political bias. They find that discrimination occurs only towards African-Americans who mention politics in any way; that is, an intersectional bias, not something the authors had previously considered (i.e., they had planned to look at racial and political bias independently). The intersectional finding, though, coheres with the theory of racial threat, where members of a majority group act aversely to minorities who may threaten their political, economic, or cultural standing (e.g., Blalock 1967, Craig et al. 2018). That is, the authors' results connect with a theory that they had not initially considered. After the fact, it seems problematic if the publication process were to penalize the authors given the existence of a clear theoretical explanation that they had previously failed to

a formal pre-registration or pre-analysis plan based on an assessment of the quality of the experiment; that is, one can address publication bias by altering norms around publishing null results without requiring a pre-analysis plan for publication.

recognize.[164] One might respond by claiming the process can adapt to such situations (e.g., Humphreys et al. 2013: 10-13, 18), but that would introduce a large gray area with no clear guidelines on evaluating what constitutes more "acceptable" exploratory theorizing. In short, the complexity of social science – accounting for context and time – has resulted in disciplines with multiple approaches and theories. It is not clear that identifying applicable theories after data collection should, in turn, require the collection of more data (which, as mentioned, is a common refrain on exploratory findings). Well-grounded theorizing can take place following data collection.[165] Coffman and Niederle (2015: 88) make a similar argument and point out that we

[164] In the paper itself, the authors clarify that this was an exploratory hypothesis but attempt to argue it stands on a strong theoretical basis. How such an approach generally works in the publication process is unclear.

[165] Nosek et al. (2018) address the reality that data collections do not always go as planned, and suggest solutions including being transparent about deviations, having sequential pre-registrations, or pre-registering a decision tree depending on steps in data analyses. The *Journal of Experimental Political Science* addresses deviations by noting that at step 2 in the review process, reviewers will assess if the conclusions are justified, which includes considerations of data quality. These are reasonable points but also lead to at least two possible problems. First, if the motivation for these institutions is to prevent researchers from engaging in problematic behaviors, what would prevent them from doing so in reporting data quality issues (e.g., not exploring SUTVA violations, engagement on the part of subjects, or all manipulation checks)? Second, a null result may stem from problems in implementation that the researcher does not recognize (or does not want to recognize). For instance, as mentioned earlier, I was in the midst

can rely on the review process and robustness tests to evaluate the extent of p-hacking rather than "handcuffing" the researcher, or prioritizing research based on it being "confirmatory" or "exploratory" (also see Laitin 2013). Scholars should distinguish p-hacking from post-hoc / exploratory theorizing with the former being a questionable research practice and latter, as explained, being a part of the research process in fields with many theories (Devezer et al. 2020).[166]

---

of conducting an experiment on nuclear energy attitudes when the Fukushima Daiichi nuclear disaster occurred. This event invalidated the entire experiment as would have been obvious at the time. However, less dramatic events in the world – or the experimental context – can undermine quality and thus null results may stem from contextual changes or poor implementation rather than reflecting quality null results. Pre-analysis plans also do not directly address a host of other questionable research practices, such as running multiple experiments and only reporting one with the predicted results, or collecting more data than planned but only including a subset (Malhotra 2021).

[166] In a review of survey experiments, Sniderman (2018: 274) makes a similar point while expressing concern about pre-commitment to hypotheses: "What has become controversial is the persistence of uncertainty after presentation of results. The fashion now is to conceive of hypothesis testing as a one-off – a decisive demonstration that a claim is or is not valid. Perhaps this is useful. But a different conception of research underpins this review: learning as you go, each advance pointing to a possible new advance. If one can see ahead, one can rarely see far ahead. It is learning what you had not known that allows you to learn (something of) what you still do not know. The research process is a process – a progression of trials."

The point at which theorizing occurs connects to Roth's (1995) goals of experiments: as discussed in Chapter 2, searching facts, speaking to theorists, whispering in the ears of princes. Pre-analysis plans in the publication process may work best in testing well-developed theory or when assessing a policy intervention (especially when it is expensive to implement, as noted). Yet, much – and in some areas the clear majority – of social science experimentation involves searching for facts. In other words, data can help researchers formulate specific relationships beyond their prior intuitions and expectations. For example, Miller and Krosnick's (2000) classic article on political media effects speculated that effects "*may* be apparent only among the relatively unusual group of citizens who are both highly knowledgeable and highly trusting of the media" (304; emphasis added). This statement represents an important finding from an incredibly influential experiment, but it is not clear that the authors expected to find this relationship. Had they presented the result as exploratory, it could have hampered the study's publication and influence.

As a third downside of strict adherence to pre-analysis plans, the process may stunt innovation since scholars become incentivized to only test well-developed hypotheses. Pre-analysis plans aim to generate greater reproducibility (Nosek et al. 2015: 1423, Nosek et al. 2018: 2602), but reproducibility does not always stimulate innovation (Coffman and Nierdele 2015: 89). Relatedly, scholars may be less inclined to add downstream outcome measures to their experiments – that is, measures that may not be the subject of the main hypotheses but rather concern second-order consequences (Sondheimer 2011). For instance, Druckman et al. (2012) show that framing universal healthcare in terms of lessening inequality or raising costs not only alters healthcare opinions but also leads to changes in opinions about immigration and taxes; the healthcare frames change the way individuals think about these other issues (c.f., Hopkins and

Mummolo 2017). The concern is that scholars may limit the number of these types of post-treatment measures for fear of being accused of data mining. Alternatively, scholars may not even fully exploit the data they collect, restricting themselves to the pre-analysis plan, thereby missing potentially important dynamics that could constitute crucial innovations (Anderson 2013, Gelman 2013). In fact, it could even preclude the use of alternative statistical analyses of the same hypothesis that are essential to ensure robustness, an essential aspect of Popper's idea of consistently testing relationships in various ways.[167]

In sum, pre-registration of an experiment prior to collecting data constitutes a public good that can address publication biases, as these biases generate misleading depictions of accumulated knowledge. The inclusion of a detailed pre-analysis plan also serves a central purpose for the researcher, ensuring that the experiment will produce the data that he/she needs to test the key hypotheses. It may be that certain journals also requires pre-registration and/or pre-analysis plans and, thus, researchers should stay abreast of such developments prior to data collection. In some cases, such as an experiment testing a clear theory or an expensive one-shot experiment, the pros of adherence to a pre-analysis plan may outweigh the cons.

[167] In my own experience, I once reported the results of an experiment using an aggregated outcome measured that took the average of three items. As this was pre-registered, the reviewers found it fine but one suggested I move analyses of each separate individual item – the results of which suggested less robust relationships – to an appendix since it was not in the pre-analysis plan. This is reasonable (in the context of the paper) but the concern is pushing too far in this direction could lead to review processes that allow for rather than prevent more robust analyses.

Yet, in many cases, researchers, reviewers, and editors may do more harm than good by strictly following pre-analysis plans. They need to be conscious that doing so replaces one common criteria for publication (statistical significance) for another (clear *a priori* hypotheses). Whether reliance on *a priori* hypotheses as a criterion undermines research quality and innovation remains unclear. At the very least, when deciding whether to fully endorse these publishing institutions, one should recognize the potential downsides (e.g., publishing poorly implemented experiments, precluding the identification of relevant theories after data collection and analysis, and undermining innovation) against the possible upsides (e.g., vitiating bias against high quality null results, possibly reducing p-hacking).[168] To be clear, those who advocate for widespread and strict use of pre-analysis plans in the publication process do so with good intention. Yet, such an approach should be done with extreme caution, if at all, due to the complexity of the social sciences. The reality is that most fields consist of a multitude of theories, and that the data often clarify what theories apply.[169]

---

[168] It may be that statistical significance correlates with higher quality implementation, while pre-analysis correlates with less p-hacking (although this is speculative). Which should be privileged depends on one's perspective. Further, empirically, it remains unclear just how problematic the publication bias problem is (due to few meta-analyses) and how often inauthentic p-hacking occurs among experimentalists.

[169] Some may say that adherence to a strict pre-analysis plan can adapt to these issues as long as researchers explicitly explain themselves; however, as noted, this introduces a large gray area and the fear is scholars end up relying on the easy heuristic of whether the pre-analysis plan was strictly followed (and hence the issues I have raised).

Summary

Experiments play a role in the scientific process; and, in many ways, not the most essential role. One should always connect experiments to the reasoning that led one to use the method in the first place. This entails documenting the rationale for the approach and then every decision along the way. By so doing, one contributes to transparency and the construction of a research program where experiments build upon each another. This is a tedious but necessary part of doing experiments. The process itself makes it easy to then pre-register and submit a pre-analysis plan in a public registry. These issues – transparency, pre-registration plans, and pre-analysis plans – echo the institutional reforms put forth by the open science movement. These initiatives, if followed to the letter, have the potential to fundamentally alter how experiments are conducted, presented, and disseminated. Many of the initiatives seem enticing on their face, but there are also multiple considerations and potential downsides.[170]

Before summarizing the argument from the previous section, I make two additional points. First, I have advocated for documentation as a way to improve the care put into experiments and the insights coming out of them. The open science motivations are broader, hoping to improve the practice of science writ large. Whether it does so remains an open question. Christensen and Miguel (2018: 970) explain, there "is also the question about the impact that the adoption of these new practices will ultimately have on the reliability of empirical research. Will the use of study registries, PAPs [pre-analysis plans], disclosure

[170] A tangentially related point is that open science practices present a social dilemma since they are individually costly but (advocates say) they would benefit the collectivity (see Kraft-Todd and Rand n.d.).

statements, and open data and materials lead to improved research quality in a way that can be credibly measured and assessed? To this point, the presumption among advocates… is that these changes will indeed lead to improvements, but rigorous evidence on these effects… will be important in determining which practices are in fact most effective."[171] As mentioned, open science initiatives strongly endorse the use of formal pre-analysis plans based on the belief that it generates better science, but ultimately it remains unclear that the practice in fact does so.

This leads to my second point: whether one fully endorses some of the reforms depends on how one evaluates research quality. One metric that receives considerable attention is replicability, the topic to which I next turn. As I will discuss, whether replicability should be a central metric remains unclear. The following summarizes the main points of the above section.

(1) Experimentalists benefit from documenting every design decision in detail, reviewing the rationale and the implications. They also should develop a plan for how the data will be analyzed.
    a. This ensures a long-standing record, facilitates writing a paper later, and helps to satisfy transparency and publication expectations such as providing access to stimuli and outcome measures.
    b. This type of record is important for future work that may expand on a given design.
    c. A data analysis plan pushes the researcher to confirm that data collected will provide a suitable test of the hypotheses of interest.

[171] Disclosure statements refer to reporting standards as well as funding and conflict of interest statements.

(2) Pre-registration refers to authors, prior to data collection, formally registering their study in an independent registry. In the aggregate, pre-registration of studies on a given topic serve as an antidote to the file drawer publication bias problem where statistically significant results are, all else constant, favored in the publication process.

(3) Formal pre-analysis plans refer to a part of a pre-registration that details the exact plan for data analysis. Many draw a distinction between confirmatory hypotheses that are part of a pre-analysis plan and exploratory hypotheses that are not.

    a. Advocates argue that following pre-analysis plans help minimize the inappropriate analysis of data and thus puts scientific discovery on firmer footing.

    b. Yet, many downsides should give one pause when it comes to strict adherence to a formal pre-analysis plan: it privileges *a priori* theory as a basis for publication over statistical significance (which may correlate with quality given the number of mistakes one can make in an experiment that would prevent finding significant results), ignores that many theories exist including theories an experimentalist may not have initially considered, and constrains innovation and the study of downstream consequences.

    c. Pre-analysis plans when testing clear and well-developed theory and/or practical interventions that involve large investments are sensible. In the latter case, the expense of the study may preclude further work that can build on unexpected findings. In other situations, however, the downsides of strict pre-analysis plans may outweigh their benefits.

Doing It Again?: Replication

Once one finishes an experiment, the next step is to analyze the data and, typically, to write a paper. The documentation process just discussed provides the author with part of the paper in advance; otherwise, it often helps to emulate the writing from papers that the experimentalist admires (also see Gerber and Green 2012: 424-446). Beyond writing, another issue that invariably arises is whether to "replicate" the study. Even more, some suggest research agendas focused on the replication of a prior experiment or experiments (e.g., OSC 2015: 943, aac4716-1, Nosek et al. 2018: 2602; c.f., Devezer et al. 2020). Most would agree that replication plays a meaningful role in science by "correcting" extant findings that do not hold or by tempering over-generalizations (e.g., Nosek et al. 2015: 1423).[172] That said, in thinking about

[172] In recent years, few issues have garnered as much attention as replication. Many believe that social science finds itself in a midst of a replication crisis (e.g., Nature 2014, Finkel et al. 2015, Baker 2016, Smaldino and McElreath 2016, Motyl et al. 2017, National Academies of Sciences, Engineering, and Medicine 2019). This stems partially from the widely discussed Open Science Collaboration's (2015) project that involved more than 250 scholars attempting to replicate 100 experiments in three highly ranked psychology journals (from 2008). They (2015: 943) report that "39% of effects were subjectively rated to have replicated the original result" (also see Doyen et. al. 2012, Lynott et al. 2014, Hagger et al. 2016, Wagenmakers et al. 2016). That said, other replication efforts have been more successful (e.g., Klein et al. 2014, Mullinix et al. 2015, Camerer et al. 2016, Camerer et al. 2018, Coppock et al. 2018, Stark et al. 2018, Yeager et al. 2019, Ruggeri et al. 2020). The extent of the crisis is thus debatable (see Fanelli 2018: 2628), and, for the reasons I next discuss, interpreting the results of any failed replication is tricky. Finally, one rarely discussed point is that an individual replication study can only serve as a

whether replication should be a part of a research agenda – either as a starting point or as a follow-up – one must consider what replications entail exactly, how they should be interpreted, why they can be ambiguous, and the opportunity cost of investing in replications.

Defining Replication

The first question concerns how to define replication. Freese and Peterson (2017) offer a useful approach by differentiating "new" studies along two dimensions: (1) are the data new? (yes or no), and (2) is the analysis the same as in the original study? (yes or no). They characterize the old-data/same analysis cell as being "verifiability," since in essence one seeks to verify what was done (i.e., re-run the same analysis on the same data). The old-data/different analysis cell refers to "robustness," as that means checking that the same results hold (e.g., using a different but still appropriate statistical/analytical technique). Freese and Peterson call the new data/similar analysis category "repeatability," since, for many, the expectation is that when one collects data with a new sample using a similar design/procedure/analysis the results should repeat. The final cell involves "generalization" – when a researcher collects new data using distinct methods or *in different settings*. Table 5-2 provides a summary of this framework.

Table 5-2: Freese and Peterson's (2017) Forms of "Replication"

	Similar Analysis	**Different Analysis**
Old Data	Verifiability	Robustness
New Data	Repeatability (aka Replication)	Generalization

"correction" to the extent that its finding enter the literature (often the published literature) and there may be publication biases unique to replication studies (see Berinsky et al. 2020).

I focus here, following many recent discussions, on Freese and Peterson's "repeatability." Should an experimentalist devote time to repeating an experiment? Many refer to "repeatability" as "replication," as will I (e.g., Bollen et al. 2015, National Academies of Sciences, Engineering, and Medicine 2019).[173]

In thinking about replication, recall that experiments use or exploit an intervention to address the Fundamental Problem of Causal Inference. That problem, as explained in Chapter 2, involves comparing *Outcome(treatment, unit) versus Outcome(control, unit)*, or determining the difference, D_O, between: (*Outcome(treatment, unit)* - *Outcome(control, unit*))$_O$. Is there a significant D_O? The "O" subscript indicates an *original* data collection. One cannot perfectly compute D_O, and thus, experimentalists address the problem by employing a design using the statistical (random assignment) or scientific (control) approach. This includes all of the underlying assumptions discussed in Chapter 2, as well as dealing with the potential problems that could arise in conducting of an experiment. Put another way, one must meet the relevant assumptions required for strong causal inference.

Replication, in essence, involves addressing the same Fundamental Problem of Causal Inference but doing it another time with a new data source. Here "R" refers to *replication* and the goal is to estimate: D_R = (*Outcome(treatment, unit)* - *Outcome(control, unit*))$_R$. For many, replication succeeds if $D_O \approx D_R$ (i.e., if the process is replicated, the outcome should replicate as

[173] The term "reproducibility" is often synonymous with Freese and Peterson's "verifiability" (National Academies of Sciences, Engineering, and Medicine 2019).

well).[174] Consequently, there exists a compounded problem of causal inference since it entails addressing the problem twice, thereby making and hopefully ensuring the satisfaction of the underlying assumptions of the design. However, it becomes even more difficult since the original experimental result also is a function of the elements of external validity – the sample (S), the treatment (T), the outcome measure (M), and the context (C). O is a $f(S, T, O, C)_O$. R also is a function of its sample, treatment, measure, and context – R is a $f(S, T, O, C)_R$. For a replication to be "successful," then, one not only has to address the Fundamental Problem of Causal Inference twice, but also assume/ensure fairly invariant samples, treatments, outcome measures, and contexts, or at least ones that do not moderate the experimental effect. That is, $f(S, T, O, C)_O \approx f(S, T, O, C)_R$. Clearly, these are difficult conditions to meet, and often very challenging even to assess. I refer to this as the fundamental problem of replication: addressing the Fundamental Problem of Causal Inference twice *and* addressing approximate invariance across the original and replication studies in samples, treatments, measures, and contexts.[175] Experimentalists considering replication must assess not only (the statistical or scientific) solutions to the Fundamental Problem of Causal Inference but also systematically the dimensions of external validity. Failure to do so could result in a very misleading set of results. If one pursues a follow-

[174] Or put another way, this compares: $(Outcome(treatment, unit) - Outcome(control, unit))_O$ versus $(Outcome(treatment, unit) - Outcome(control, unit))_R$, with the expectation that a successful replication will mean the near equivalence of the functions being compared.

[175] Brandt et al. (2014) provide a useful guide to replication, revealing the large number of considerations in play.

up data collection, then one must think through the elements of the fundamental problem of replication.

Interpreting Replications

How does one assess if something replicated ($D_O \approx D_R$) (e.g., National Academies of Sciences, Engineering, and Medicine 2019)? There exist a host of distinct metrics scholars use to assess whether a single or multiple replication attempts repeat prior work, and it remains unclear how to identify the most relevant metric (Camerer et al. 2016: 1434).[176] An even more

[176] Examples include replication/reproducibility rates, which are the proportion of studies that produce significant effects in the same direction (e.g., Mullinix et al. 2015,OSC 2015, Camerer et al. 2016, Camerer et al. 2018, Coppock et al. 2018), prediction intervals where one can estimate, say, the 95% prediction interval for the original estimate and test how many replications fall in that interval (or alternatively, confidence intervals) (e.g., Cumming 2008, OSC 2015, Patil et al. 2016, Camerer et al. 2016, Camerer et al. 2018), peer beliefs about replication such as prediction markets (e.g., Dreber et al. 2015, OSC 2015, Camerer et al. 2016, Camerer et al. 2018), effect size differences between original and replication studies (e.g., OSC 2015, Camerer et al. 2016), meta-analytic approaches (e.g., OSC 2015, Camerer et al. 2016, Camerer et al. 2018), the small telescopes approach that estimates whether a replication effect size is significantly smaller than a "small effect" in the original study with a one-sided test at the 5% level (e.g., Simonsohn 2015, Camerer et al. 2018), a Bayesian approach that assigns a probability to a truth claim and then views replications as way to increase evidence for or against the original claim (e.g., Goodman et al. 2016), and the positive predicted value, which is the proportion of positive results in statistics (e.g., Goodman and Greenland 2007).

perplexing methodological issue concerns the presumed null hypothesis with a replication,: studies replicate (e.g., $D_O = D_R$). The question then typically is whether we should reject the null hypothesis, which would suggest a failed replication. This approach means documenting replication entails *accepting* a null hypothesis (Hedges 2019). As explained in Chapter 2, experimental analyses generally aim to *reject* a null of no difference (e.g., treatment mean = control mean), with the goal to offer evidence of a causal effect. That is, one wants to ultimately reject, not accept the null hypothesis. Recognizing this contrary logic between experimental tests and replication approaches, Hedges and Schauer (2019: 11-12) suggest that "if the goal of conducting a replication is to determine that study results are similar, then this [i.e., H_o: $D_O = D_R$] is the wrong inferential structure. Instead, the burden of proof should be on replication, rather than non-replication. In this setup, the null hypothesis is that the studies failed to replicate… and rejecting the null hypothesis would mean that we conclude that the studies replicate." We would be looking to see if we can reject the null that $D_O \neq D_R$. This changes the way to think about replication as it requires conclusive evidence of replication rather than evidence of non-replication (to reject the null). It shifts the burden of proof to replication (also see Hedges 2019).[177] Replication efforts, with this logic, should not assume that the original H_0 (from the study being replicated) has been irrevocably rejected.

This highlights why many tests are needed to determine whether a causal proposition has or has not withstood falsification. Practically, this means a research agenda focused on

[177] Hedges (2019) and Hedges and Schauer (2019) further explain that due to statistical uncertainty in the original and replication studies, it is often impossible to adequately assess the success of a replication with a single replication attempt.

replication should consider a distinct inferential structure. If instead one pursues replication to assess if a theoretical proposition holds in multiple tests – that is, to conduct follow-up tests after an initial experiment – one should discuss the follow-up (replication) results in terms of finding or not finding evidence consistent with the initial hypothesis, rather than in terms of "replication." This highlights the problem of comparing effect sizes between an original study and a replication since particular effect sizes likely vary due to unidentified sources. This segues into my next point which concerns the ambiguity of replications.

The Ambiguity of Replication

The fundamental problem of replication is inherently difficult (also see Wong and Steiner 2018). On the external validity dimensions, using identical treatments and outcome measures does not ensure the replication of the context (or timing). Moreover, even a similarly drawn sample from the same population can introduce uncertainty as many sample pools evolve over time due to a changing world (see, e.g., Arechar and Rand 2020, Aronow et al. 2020).[178] When one ostensibly fails to find results consistent with a prior experiment (i.e., $D_O \neq D_R$ in the typical treatment), it provides an opportunity for learning a la my above discussion of "kaput."

A failure to find the same results again could stem from an inability to satisfy internal validity assumptions in addressing the Fundamental Problem of Causal Inference (e.g., low experimental realism). Or it, may reflect external validity issues such that generalization may be

[178] Peyton et al. (2020) explore this issue with regard to respondent attention levels and suggest approaches to correct for shifting levels of attention (to address what they refer to as temporal validity).

constrained to particular conditions.[179] In political science – a discipline focused on context and time (e.g., Druckman and Lupia 2006) – a vast number of findings undoubtedly apply only to particular contexts and times. The same surely holds in other fields where social and economic settings alter behaviors. Indeed, as discussed in Chapter 3, studies independently deemed less sensitive to contextual considerations (e.g., time, rural/urban) were more likely to replicate in a massive replication effort of 100 experiments (Van Bavel et al. 2016a).

To take this argument a step further, consider replication from a Popperian perspective. One attempts to test a hypothesis that already has evidence in its favor. However, if the replication does not produce consistent results then one either abandons the theory, or, assuming no methodological errors, amends the theory based on its conditionality (on replication and theory development, see Klein 2014).[180] This process then can lead to a more focused test to try to directly isolate conditions/moderators. Put another way, it often is difficult to know if an attempted replication constitutes a replication study or a generalization study, due to the challenge of discerning what qualifies as "different settings" (Bollen et al. 2015: 7). Freese and Peterson (2017: 158) acknowledge this slippery slope in pointing out that "what empirical work tests is not so much propositions as such but ideas about the scope in which the propositions apply… The chronic ambiguity of generalizability is whether explanations revised to

[179] Another possibility, when replicating other's work, is it stems from questionable research practices in the original experiment. This explanation seems to be a common implicit presumption in the aforementioned replication crisis narrative, yet there is little evidence along these lines (see Fanelli 2018: 2628).

[180] Realistically, one should not rely on a single replication (Cook and Campbell 1979: 31).

accommodate new results constitute a legitimate advance or are simply ad hoc explanations…" This strikes me as the vital question with any experiment that aims to repeat what had previously been done. In fact, in some cases it would be odd for an experiment implemented at a later time to find a result consistent with prior studies, such as studies on topics where societal attitudes (e.g., gender, race) have clearly changed over time (e.g., Burden et al. 2017, Valenzuela and Reny 2021, Krupnikov et al. n.d.). Devezer et al. (2020) make the point that replication need not and perhaps should not be seen as cornerstone of science, as it depends on the field/area of study. Among other dynamics, context and time dependency means that "science does – rather often, in fact – make claims about non-reproducible phenomena and deems such claims to be true in spite of the non-reproducibility. In these instances what scientists do is to define and implement appropriate criteria for assessing the rigor and the validity of the results…without making a reference to replication or reproduction of an experimental result" (Devezer et al. 2020: 2-3).

Given the fundamental problem of replication, it seems in many, if not, most cases, it would be most productive for researchers to treat inconsistent results as a generalization issue. The challenge, then, for the researcher is to identify the scope conditions (also see Redish et al. 2018: 5043).[181] In short, I recommend a forward-looking view of "replication" – if one finds

[181] In particular, there needs to be considerably more attention to theorizing about context. This point is unfortunately often missed. Take, for instance, this statement about replications: "studies differ at minimum in their method factors… [including] seemingly minor factors such as the social context, the subject pool, and the time of day" (McShane et al. 2019: 99). Social context, samples, and timing are far from "minor factors." Similarly, I had an experience of being part of a large scale replication project where my task was to replicate an experiment that looked at the

inconsistent results, they might consider it an opportunity to build in a new direction rather than a time to look back dismissively at prior experiments (Anderson et al. 2016).

I am far from the first to point out the challenges and ambiguity of replication. Most notably, Collins (1985, 2016) accentuates the interpretative nature of experiments and replications in the sciences. In essence, every time an experiment is done again, it involves new explanatory factors – even if unmeasured – due to unavoidable changes between experiments (e.g., in timing, context, sample). One should think about such changes rather than assuming invariance. The end goal should be to accumulate knowledge on a given topic. Testing a proposition with repeated experiments and/or expanding on prior designs is a route to knowledge, but only when experimentalists carefully consider potential differences and use inconsistent results (e.g., kaput) as a way to generate new work.

One example of carefully considering differences comes from Westwood et al. (2019). In prior work, Lelkes and Westwood (2017) revealed limits to the extent partisans discriminate against each another in the United States (e.g., they do not endorse direct harm to their political

---

role of opinions about government services (e.g., education, health care) on candidate choice in the U.S. (between a clear versus ambiguous candidate). The original data were collected in 2007 whereas the replication took place in 2020. Much time and expense were put into ensuring the same stimuli and type of sample were used in the replication, but at no point in the process was there discussion about the changed context which included the 2020 data being collected in the midst of a global pandemic and a contentious presidential campaign. It is possible that changed context would not affect the relationship under study but it certainly should have been carefully considered.

opponents). Those data came from 2014 and the authors realized that the partisan context had substantially changed by 2017, with the election of President Trump. They re-ran the experiment in 2017, speculating but not finding an increase in prejudice. Here the authors sought to assess contextual change and replicated as a way to test for the effects (although they find scant evidence of change). Busby and Druckman (2018) do the same thing in their replication study of how irrelevant effects can impact political opinion, replicating their prior experiment from a distinct context. Both these examples highlight the usefulness of what Janz and Freese (n.d.) call replicating yourself.

Another example is Stark et al. (2018), who replicate survey question order effects across contexts; they show, for example, that priming evenhandedness via question order occurs more strongly in countries with more individualistic as opposed to collectivistic cultures (priming is not needed in collectivist cultures). They conclude that the norm of evenhandedness does not depend on individual attributes as much as the variations in culture. Here the contextual effects were theorized in advance and replications tested them. These examples reveal that it can be beneficial to think about context before conducting a replication.

Practically, experimentalists should neither assume a prior result will replicate nor launch a research agenda for the sole purpose of replication. Rather, use prior designs and systematically extend them to explore new explanatory factors, particularly those related to context, and seek to generalize results.

Opportunity Costs to Replication

I just argued that a replication ultimately should be done in the service of testing an extant theory, empirical relationship, or policy intervention. This understanding appears often lost on those doing replications and creates a bit of a dilemma. On the one hand, a good

replication tries to emulate all that was done in the original study. On the other hand, a good experiment uses the most valid constructs, etc. For me, the latter goal should take precedence. If that means the replication becomes less of a match to the prior study, the tradeoff is acceptable if it makes for a stronger test of the proposition under study.[182] The ultimate end goal is not replication but assessing the causal relationship under consideration so as to inform a theory (Stroebe and Strack 2014, Nosek and Errington 2019). If a replication fails, a Popperian perspective should lead a researcher to ask what to do with the underlying theory, not what it means for the "rate of replication" *per se*. As Smaldino (2019: 9) states, "we don't just want science to be reproducibility. We want it to help us to make better sense of the world…. we need good theory."

The implication is that a research agenda that replicates with the focus on mechanically repeating what was done in the past comes with an opportunity cost of doing something that can more directly advance theory. At times, doing the same exact thing makes sense, to be sure, but even then experimentalists need to carefully consider the aforementioned inevitable differences over time and context. This also applies to operationalizations – recall the project described in Chapter 3 where different teams of researchers sought to test the same hypotheses from the areas of moral judgment, negotiation, and implicit cognition. In so doing, they opted for different stimuli that, in turn, led to very distinct results (Landy et al. 2020). This project of "conceptual replication" provided crucial insight into the robustness of different theoretical constructs concerning their operationalizations, and would have been missed if the teams instead all strictly

[182] Of course if the aim of a replication is to explicitly assess over-time changes, then one should keep elements the same; see Chapter 2's discussion of measurement tradeoffs.

replicated one approach. The point is that resources and times are best spent in the service of addressing the questions that motivate our research agendas and not strictly replicating past work for the sake of doing so. This perspective – that replications inherently will vary – questions the worth of programs that purportedly aim to serve as "checks on science" via replication.

That said, this is not meant as an argument against repeating past study designs. As discussed earlier, a fruitful way to design an experiment entails building on past efforts. Given the challenges of creating sound stimuli, valid measures, and useful designs, researchers can benefit from extending designs that have already overcome such challenges. This often involves replicating parts of the design and building in additional conditions/stimuli. For example, recall the Druckman and Nelson (2003) campaign finance framing experiment detailed in prior chapters. Thus far, it has been described as a study that randomly exposed some respondents to a special interests frame (for campaign finance regulation) and others to a free speech frame (again campaign finance regulation). This experiment revealed a framing effect (e.g., the special interests frame led to more support for regulation) which confirms much prior work on the topic in line with the classic Nelson et al. (1997) study. However, Druckman and Nelson (2003) also sought to identify conditions under which framing effects occur, theorizing that they may not when exposure is followed by inter-personal discussions among people with varying opinions. They thus added experimental conditions that included conversations and show that these interactions condition framing effects. In essence, they conceptually replicated canonical framing studies in route to adding to them and identifying generalizability conditions. They sought to generalize and innovate via replication rather than to replicate for replication's sake. The very factors that make experiments so difficult to conduct and interpret apply even more so to replications. This is why a fruitful strategy is to build on well-established experimental protocols

when designing new studies. Validation and verification have a role to play (Dunning 2016), but achieving those goals need not prevent simultaneous innovation. Time is better spent using replication as a route to novelty, as, in essence that is being done whether the experimentalist knows it or not (i.e., every experiment is a new test under new conditions and thus providing something novel).

Accumulating Studies to Identify Effect Sizes

This line of argument leads me back to a point brought up in Chapter 2 about effect sizes. There I suggested that unless a policy goal drives an experimental agenda (see the scalability discussion in Chapter 3), the effect size of a treatment from a given experiment pales in significance to whether one falsified the hypothesis. Here I offer a caveat, arguing that the importance of effect sizes lies in the accumulation of evidence, via pure replications or not. It is through this process that one might isolate an effect size which I acknowledge, even for non-policy oriented experiments, can be a quantity of interest for substantive reasons that do and, as I have argued, should drive the any experimental design. For example, whether the effect is small or large informs debates such as: how much negative advertisement influences voting turnout?; how much does being a minority undermine legislative responsiveness?; and how much does being a female candidate affect electability?.

However, to see why identifying an effect size is best done (or can only be reliably done) via an accumulation of evidence, consider experiments on party cue effects. One could argue the effect size from a given experiment on party cues (e.g., where the experimenter asks respondents about a policy after randomly receiving an endorsement from a party or no endorsement) matters by providing knowledge about the extent to which parties shape opinions. The impact of party endorsements informs what we know about political attitude formation and normative theories of

democratic representation (e.g., Druckman 2014). Yet, studies suggest the effect of party cues on policy opinions range from 3% to 43% of the policy opinion scale (Bullock 2011; also see Bullock 2020). Tappin (2020) explains that such variations reflect huge differences across designs, including the type of baseline policy information provided, the source of the cue (e.g., a party leader, the party in general, etc.), the format of the cue (e.g., varying percentages of partisans support a given policy), whether cues come from one or more parties, distinct outcome measures (e.g., binary versus scaled outcome and in the case of the latter the extremity of the scale endpoints), and more (also see Slothuus 2016). Tappin further points out – and then focuses upon – how party cue experiments differ dramatically in the policy issues used – e.g., there is no sampling from a population of policy issues (Druckman and Leeper 2012a, Clifford et al. 2019; also see Chapter 2 discussion). This astute observation makes clear that inferring and generalizing about a specific effect size from one or a few experiments is challenging, if not misleading due to invariable design differences that one may not even imagine until a mature literature develops.

This is why I have minimized the relevance of effect sizes from single experiments. That said, experimentalists can pursue a few approaches to identify the size of a given treatment. First, naturally occurring data, particularly when an intervention occurs either randomly or in a way where one can be confident about unit homogeneity (see Chapter 2) can offer insight. For instance, Slothuus and Bisgaard (2020) track Danish citizens' policy opinions when their political parties, without warning, reversed their positions on two salient welfare issues (i.e., satisfying the unit homogeneity assumption). They find partisans changed their opinions in line with their parties by roughly 15 percentage points. Their study provides a naturally occurring baseline of party cue effects against which experiments can compare themselves and isolate

design features to explain deviations from the baseline. Second, insight can be garnered by comparing effect sizes across types of experiments so as to see if there is a consistency in effect sizes (e.g., fields versus laboratory; Coppock and Green 2015). Strong correspondence leads to confidence in the size of a treatment.

Third, the ideal approach entails aggregating multiple experiments – including if not especially replications – since doing so envelopes unknowable dimensions that vary across experiments (e.g., timing, context, sample, etc.). This ostensibly contradicts my prior hesitation about replications. To clarify, my point is that any replication unavoidably differs from prior work and this is exactly why aggregating similar studies provides progress by, in essence, controlling for unidentified confounds. This reality accentuates the advantages of replicating parts of prior experiments but at the same time extending them, or, as stated above, "innovate via replication."

Meta-analyses synthesize treatment effect sizes to arrive at a single effect size estimate. For example, one may be interested in the effect of negative political advertising on voting turnout: how much does it change turnout? As discussed in Chapter 2, an initial experiment reported that those exposed to the negative advertisements are 2.5% less likely to vote than those exposed to no political advertisement (Ansolabehere et al. 1994: 833).[183] But that single study had idiosyncratic features of focusing on particular campaigns (in California) at a particular time (1990s) with particular types of negative ads (e.g., focusing on salient campaign themes). Once many more studies were conducted across various samples, contexts, times, and stimuli, Lau et

[183] They find that positive advertisements mobilize turnout by about 2.5%; thus, exposure to a negative as opposed to a positive advertisement reduces vote intention by 5%.

al. (2007) meta-analyzed them (111 studies) to isolate the likely effect size (also see Lau et al. 1999). They find an aggregate effect size of a fairly meaningless -.07. Here then the accumulation of studies led to a conclusion of no significant effect size but one can easily imagine a situation where an aggregation of results leads to a positive or negative and substantively meaningful effect. For instance, Costa (2017) meta-analyzes 41 experiments on legislative responsiveness – a la Butler and Broockman (2011) – isolating an effect size such that requests from minority constituents are almost 10% less likely to receive a response than non-minority constituents (249). Oschatz and Marker (2020) meta-analyze 14 experiments on the persuasiveness of narrative messages (i.e., those with a beginning, middle, and end that provides information about the scene, characters and conflict), finding that it changes attitudes by about .14 standard deviations more than analogues non-narrative messages.

And, Schwarz and Coppock (2020) meta-analyze 67 candidate choice experiments on gender; these studies had produced wide ranging effects ranging from women candidates receiving nearly 11% fewer votes than comparable men candidates to women garnering greater than 9% more votes. The authors aggregate the studies to identify an effect estimate of being a woman candidate (relative to a male candidate) leads to a roughly 2 percentage point gain in vote margin. This example highlights not only the danger of over-interpreting a single treatment effect, given the 20% range in effect sizes in the literature, but also how taking stock of a variation in treatments can advance knowledge by pushing researchers to consider the source of the variation (Borenstein et al. 2009: Part 4). Schwarz and Coppock (2020) exploit the fact that the studies explored a host of other factors aside from gender, finding, for example, that the

gender bonus is marginally larger for white candidates than Black candidates (although short of significance).[184]

Indeed, meta-analyses provide the most insightful when they identify relationships between effects sizes and aspects of the intervention, context, sample, etc. (e.g., Lepper et al. 1999). In that vein, it is essential that any aggregation endeavor includes only studies that employ commensurate treatments and outcome measures. As meta-analyses proliferate, ensuring comparability and inclusion of high quality studies becomes vital; ultimately, a meta-analysis constitutes an observational study that combines data based on the quantity and not the quality of observations and thus the author(s) must serve as quality gatekeepers (Packer 2017).

In sum, a given experimental treatment effect is limited by the conditions of the test, conditions that one typically cannot fully appreciate until considering a range of studies that introduce plausible variations in the tests (a la the party cue examples). Thus, a single effect should be cautiously interpreted but can become essential, informing substantive debates when one synthesizes a literature. Replications play a vital role in that process, particularly when they explicitly extend upon what is already known.[185]

Summary

When one completes an experiment, an obvious question concerns what to do next (beyond writing a paper). Whether one should invest in conducting another experiment on the

[184] They also look at respondent characteristics such as gender (women respondents give a larger bonus) and partisanship (Republican respondents give a smaller bonus).

[185] This suggests that replication research programs are better off attempting to replicate fewer phenomena many times than replicating many phenomena a few times (McShane et al. 2019).

same topic – that is replicate prior work – is not an easy question to answer, particularly in light of its emphasis via the open science movement. Replications play a vital role in the scientific progress, but often less so for replication's sake, and more as a route to generalization, innovation, and aggregation. A summary of my main points concerning how to think about and pursue replications is as follows.

(1) Replication typically refers to repeating a prior experimental design with new data. Doing so entails addressing "the fundamental problem of replication." This refers to the reality that any replication compounds the Fundamental Problem of Causal Inference since it must be done twice, *and* one must attend to the dimensions of external validity in the process (e.g., are the samples, treatments, measures, and contexts comparable?). Experimentalists interested in repeating a study need to systematically assess these issues (instead of assuming they are met).

(2) If an experimentalist's research agenda is to replicate prior studies, the inferential structure needs to treat the null hypothesis as a failed replication. When the goal is to conduct another study as a follow-up, an experimentalist should focus on finding evidence consistent with the initial hypothesis (rather than in terms of replication of the effect found in the earlier experiment).

(3) Replication efforts are inherently ambiguous given variations in the dimensions of external validity. It can be more productive to think of repeated studies (replications) as attempts to generalize. This can be done by systematically assessing changes prior to the replication or doing it afterwards. Regardless, the goal should be to consider how another (replication) experiment informs the original hypothesis or hypotheses.

(4) Replication for replication's sake comes with an opportunity cost. Instead, a repeated study should be done to advance theory. This often means repeating a prior design, but extending it in innovative ways. In that sense, replication is a crucial route to progress in experimental social science because it provides the basis for innovation in theory.

(5) While the effect sizes from a single experiment and/or a replication should be interpreted with great caution, a series of similar experiments, including replications, can be aggregated via a meta-analysis to isolate an average effect size and to explore sources of variations in effects.

Conclusion

The goal of experimentation is to arrive at generalizable causal inferences. This is done in the context of the larger scientific process that, realistically, does not operate mechanically from question to theory to hypothesis to test to result. There exists an inevitable back and forth between data and theory in much of the social sciences, where multiple theories exist to explain behavior in varying contexts. That reality should be a meaningful consideration as people think about experiments. It also affects how to approach steps prior, during, and after an experiment. Questions emerge from various sources, including failures from other experiments. Conducting an experiment is much more complex than it appears at first glance, which is why it is important to maintain clear records throughout the process. That said, experimentalists should not be constrained by formal pre-analysis plans, or mechanical efforts to replicate prior work. As Sniderman (1995: 465) states, "'*to replicate an existing finding – to follow the precise path taken by a previous researcher*, and then improve on the data or methodology in one way or another.' Believe that, and you miss what science – including social science – is really about. Accuracy, meticulousness, exactness – all are virtues – but they are minor virtues and should not be worshiped as the soul of science. Imagination, originality, creativity, seeing what others not only

failed to see but did not even suspect – that is the heart of science of the first order" (italics in original). To be clear, replication and innovation are not incompatible. In fact replication, when thought of as a way to generalize and aggregate, offers is a springboard to innovation. The key is to maintain curiosity, do so rigorously, and remember the ultimate goal is to answer questions about the world. In that spirit, the highlights of this chapter are as follows.

- The questions that can lead to an experiment come from a host of sources as captured by *ASK: assess* the world and the field, *socialize* with colleagues, teachers/students, and non-academics, and build on experiments that failed or were *kaput*.
- Document every design decision in detail, starting with the rationale for the study through the plan to analyze the data. This ensures a careful record and makes writing a paper and meeting transparency requirements easier.
- Experimentalists should pre-register the existence of their study as one approach for helping with publication bias.
- Part of the documentation should include a careful plan on how to analyze the data. However, if an experimentalist formally adds that plan to the pre-registration, do not be hamstrung by it – as an author or a reviewer – as that privileges certain types of work, ignores the reality of there being many theories in the social sciences, and can limit innovation. (The exception is in cases of very well developed theory and/or large scale policy experimentation.)
- Once an experiment is done, it is worth considering doing it again. But this involves addressing the "the fundamental problem of replication" by accounting for common, if not inevitable, variability in dimensions of external validity.

- Repeat experiments or replications should typically be used to test the same hypotheses as the original (rather than to assess the replication rate per se), and as a way to generalize, innovate, and aggregate in a given area.

Chapter 6: Designing "Good" Experiments

Most texts on experiments focus on design specifics, types of experiments (e.g., laboratory, survey, field), how to conduct analyses, and/or particular applications. I have taken a distinct approach by focusing on how to think about experiments. I have done so in response to the massive changes that have occurred, over the past decades, with regards to social science experimentation. It was just twenty years ago when I received journal reviews on multiple occasions dismissing experimental work as "inappropriate" and thus rejecting the submissions on their face. Now, it is not atypical to receive reviews questioning whether non-experimental work can be useful if one wants to make a causal claim. Much has changed and this evolution has occurred concomitant with new technological opportunities and altered sociological norms about science (e.g., open science).

One might think that during this transformative time, the fundamentals of experimentation have changed as well. Yet, they have not; if anything, scholars may be less apt to attend to some important details, since data collection costs have become so much lower and thus less is at stake with each data collection. With that in mind, in this chapter I highlight the steps involved in designing a "good" experiment. Consistent with the theme of the book, these steps place experiments in the larger social scientific process in which experiments constitute one step. The idea is not to offer a checklist; rather, it involves a research process in which one situates the study.[186] I now turn to these steps, after which I offer brief concluding thoughts.

Steps to Designing "Good" Experiments

[186] Gerber et al. (2014) and Boudreau (2021) offer superb "checklists" for specific aspects of experiments (also see Christensen et al. 2019: 143-157).

What it means to be a "good" experiment is unclear – and increasingly so given recognition that null results can be of interest. I have obviously touched on a host of criteria, including satisfying the causal inference assumptions, experimental realism, attending to multiple dimensions of external validity, carefully articulating the relevant points of comparison, and so on. Ultimately, a "good" experiment is one that offers a contribution to extant knowledge. Here I provide a list of steps that, all else constant, should lead to improved experiments.[187] In some sense, they are straightforward steps that echo the scientific process and – perhaps ironically given my critiques – formalize the open science practices for the sake of producing better quality research (aside from ensuring it is transparent, etc.). To be clear, following these steps involves the investment of substantial time to maximize the worth of experimental data collection. For me, the process takes on average about nine months, although it varies depending on the complexity and novelty of the experiment (i.e., novelty in the sense of my own prior knowledge on the topic). The time also will of course differ among individuals depending on their work style, career stage, and life circumstance. I list out the steps in broad strokes (for more details, including examples, see Druckman et al. 2018a). With each step, I provide a brief example, but obviously that barely touches on the relevant issues.

- *Big picture idea:* a short (i.e., few pages) document on the general topic and why it is relevant to understanding social, political, and/or economic phenomena. This document ideally iterates multiple times for feedback. It serves as an essential starting point by including a "big think" question and explaining why it is worth addressing. Even when one builds on a well-established experimental paradigm, this step serves as a check to

[187] This list originally appeared in Druckman et al. (2018a).

ensure existing work has developed in a way that has not lost sight of what scholars ultimately want to understand.

    - For example, instead of immediately designing an audit study of democratic responsiveness, a scholar should step back and consider why one would want to study responsiveness (e.g., relative to other aspects of representation or democratic processes).

- *Detailed literature review*: an exhaustive search of research on the topic and detailed descriptions of specific studies. Ideally, it leads to the identification of multiple potential research directions, some of which are tabled for the future. It is crucial here to look for literature across fields and methods to assess the state of knowledge and the strengths and weaknesses of different approaches. It is at this stage that the researcher should identify specific gaps in existing knowledge that he/she hopes to address. The literature covered should be determined by substance and not method.

    - For example, one should not only review audit studies on democratic responsiveness; also included should be work on what we know about democratic responsiveness more generally, including theoretical work on representation (e.g., Mansbridge 2003, Disch 2011), qualitative work on elected officials (e.g., Fenno 1978), quantitative work on the public opinion-policy connection (e.g., Erikson et al. 2002), and so on. Most of this document will not ultimately become part of a publication but it serves as a foundation for the project and its potential contribution.

- *Research question(s) and outcomes*: given the identification of a gap in existing work, the next step is to put forth a specific question (or questions) to be addressed. This includes identifying the precise outcome variable(s) of interest.
    - For example, say one wants to study bias in responsiveness – one specific gap might be whether there is discrimination in response to requests based on one's "personal position" relative to the elected officials. Fenno (1978) identifies four "constituents," including the geographic (i.e., those in the district), re-election (i.e., supporters), primary (i.e., strong supporters), and personal (i.e. intimates). The question of interest may be how representatives respond to members of these distinct constituents – do they favor intimates over those in the district?[188] The outcome may be response and also the content of the response to a constituency service request. One may further want to compare the effect of any bias based on constituencies against racial bias given the extant literature, as discussed, offers substantial evidence of racial discrimination.
- *Theory and hypotheses:* development of a theory and hypotheses to be tested. This often involves accessing distinct literatures, sometimes from other disciplines. Researchers should take their time to derive concrete and specific predictions. As part of this step, potential mediators and/or moderators should be specified. Also, in putting forth predictions, one must be careful to isolate the comparisons to be used (see Chapters 2 and 4).

[188] As discussed in Chapter 5, Kalla and Broockman (2016) present a related experiment on the impact of donor versus non-donor requests for access.

- For example, in developing a theory about whether an elected representative may be more or less likely to respond to inquiries from distinct constituencies, one may turn to work on campaign communications, fundraising, interest groups, voting, corporate strategy (in economics/business), interpersonal interactions and trust (in psychology), etc. Further, one may posit a mechanism such as trust – e.g., does the elected official trust that response to a given inquiry will generate meaningful support? – and moderators such as the electoral competitiveness of a given district. The researcher must think through the relevant comparisons. It might be most interesting to compare all other types of responsiveness relative to the geographic constituency condition, as normatively the official represents those individuals. Alternatively, perhaps comparing supporters to intimates is more meaningful to see the impact of resources on responsiveness. Or, comparisons against racial bias may be most intriguing to assess the relative sizes of inequities in representation.

- *Research design*: the scholar puts forth a design (see Leeper 2011). This includes at least six components, as follows.
    1) Discussion of the designs used by others who have addressed similar questions, and how the proposed design connects with previous work. As emphasized in Chapter 5, the ideal strategy is to utilize and extend prior designs.
    2) Discussion of how such a design will provide data relevant to the larger question(s).
    3) Identifying where the data will come from, which includes:

i. Consideration of the sample and any potential biases. Importantly, as highlighted in Chapter 2, it is essential to explicitly state the target populations to which one hopes to make an inference. This includes the units, contexts, measures, and outcomes.

ii. Detailed measures and where the measures were obtained – for example, were they used in prior studies? The measures need to clearly connect to the hypotheses, including the mediators/moderators. One should discuss the validity and accuracy of the measures (see Chapter 2).

4) In many cases, the design may be too practically complex (e.g., number of experimental conditions relative to a realistic sample size), and decisions must be made on what can be trimmed without interfering with the goal of the study. This process will typically require a power analysis (see Chapter 2).
5) A discussion of necessary pre-tests of stimuli, question wordings, etc.
6) Issues related to internal and external validity should be discussed (see Chapter 3).
    - For example, one should consider distinct approaches and designs used to address questions about legislative responsiveness. If an audit approach is chosen, past work should be reviewed with reference to the initial "big picture" question. One might use a sample of state legislators, as is common, to conduct an audit (assuming the level of constituency theory put forth applies). It would be important to justify operationalization decisions – for instance, the geographic constituency could be operationalized with a zip code reference while the strong supporter condition could be operationalized with

reference to donations. It is important to justify that those operationalizations would be perceived as such by officials. This requires piloting and consideration of the information equivalence problem discussed in Chapter 3 (e.g., do certain zip codes also signal wealth?). Then one must turn to the content of the request and how responses might be coded, etc. Finally, as mentioned, if the projected sample size is too small, some experimental conditions may have to be cut. If the aforementioned design planned to cross the four types of constituents with their race, then 16 conditions may be too many and difficult decisions must be made.

- *Data collection document:* a step-by-step plan of how data will be collected so as not to later forget such details as recruitment, implementation, etc.
    - For example, how will one obtain contact information for elected officials, what e-mail accounts will be used to send inquiries on what dates, how are bounce back e-mails treated, etc.?
- *Data analysis plan:* there needs to be a clear data analysis plan—how *exactly* will the data be used to test hypotheses? The researcher should directly connect the design and measures to the hypotheses. This often involves making a table with each measure and how it maps onto specific hypotheses. What techniques will be used if the data collection is imperfect (e.g., smaller sample size than expected)? The plan ensures that the right data will be collected, and it provides a blueprint of what to do once one receives the data.
    - For example, a la the prior discussion, what conditions will be compared to what and how might outcomes be coded? One might initially compare percentages of

responses; however, analyses might also include coding the content in which case precise coding rules need to be laid out, as do approaches to analysis.

    - This step is akin to an internal pre-analysis plan – it is why in Chapter 5 I emphasized the importance of such a plan for research quality. The point is to ensure that the researcher collects the right data and knows what to do with the data once it is collected – rather than to differentiate confirmatory and exploratory hypotheses. It is common to deviate from this plan, but having a plan is essential to avoid mis-collecting data.

- *Institutional Review Board:* the researcher should complete the Institutional Review Board (IRB) process. If the project involves non-trivial deception or other aspects that could prevent/delay IRB approval, the researcher should directly contact an IRB representative early in the development of the project. In many cases, the researcher will work with the IRB to figure out what is feasible. The process can take weeks or months and so starting early is important (perhaps even before the prior steps). Researchers should attend to ethical, legal, and political considerations where applicable (King and Sands 2015). An additional benefit of following the steps outlined above is that the researcher will have already prepared nearly all information and documents required for an IRB application.
    - For example, for an audit study with no consent or debriefing, an IRB that is not familiar with such studies may generate many iterations that could slow down implementation if not addressed sufficiently early.
- *Merging the pieces:* at this point, the researcher merges the aforementioned pieces into a single document and, ideally, others review it. This step serves as a check that in the

process of moving from the larger abstract research question to the specific analysis plan a disconnect did not emerge. In some instances, the particular design no longer speaks to the initial motivating question or no longer clearly fills a gap in the literature.

    - For example, one may have started out by asking about if elected officials equally represent their constituents but ended with a design that had to be cut due to power considerations that explores the relative bias in responsiveness towards "intimates" and "strong supporters" as conditioned by race (e.g., varied race in the design). This remains an interesting question but no longer is about the treatment of the geographic constituency writ large.

- *Implementation:* from here, data are collected and analyzed, and the planning document serves as a guide to writing up the results and, potentially, identifying reasons why one may not have found what was expected. A part of this implementation involves careful record-keeping of data analytic choices. This also requires careful attention and documentation of challenges that arise during data collection such as participants not arriving or complying, programming errors, naturally occurring events during data collection that could alter the data, and other unanticipated problems (see Karlan and Appel 2016).

The entire approach means that writing much of a paper entails cutting out superfluous parts as one has an exhaustive experimental design document before data are even collected. There is nothing magical about these steps. The framework simply breaks down the research process into manageable pieces and forces researchers to think through each specific decision, which reduces the likelihood of a design error. In an age where data are so easily obtainable, the risk is that researchers jump right to the design stage (the 5th or 10 steps), or the data analysis

plan (the 7th of 10 steps), without considering the construction of a theory. Open science initiatives that focus on transparency, pre-registration, and replication rarely mention the first four steps (prior to the research design). Yet, one can only arrive at a meaningful pre-analysis plan by spending time thinking about the question being asked, reviewing the relevant literature, specifying a specific question and outcome, and developing a theory and hypotheses. These early steps are as vital to conducting meaningful experiments as are ensuring transparency or conducting replications. Put another way, much of what makes a good experiment occurs well before one even starts to think about the experimental particulars. When one gets to the step of experimental specifics, the theory should guide the choice of conditions, comparisons, and analytic approaches. The experimentalist should be able to easily explain what the test will add to extant knowledge and how the data will be analyzed to do so. Only then can we start to worry about many of the topics covered in this book and other discussions about the experimental method.

Conclusion

I began this book with two quotes from presidents of the American Political Science Association. The first quote from more than 100 years ago suggested experiments were an impossibility in political science (Lowell 1910). Methodological and technological developments, along with scholarly ingenuity, have shown this is not the case – experiments play a crucial role not just in political science but the social sciences more generally. The second quote from 2020 expressed concern that an overreliance on experiments would constrict the questions social scientists address (Smith 2020). This concern is reasonable given the remarkable rise of experiments and their central place in recent epistemological debates about open science. I share the concern and my hope is that I have addressed it with a two-pronged answer. First,

thinking experimentally introduces a host of considerations that are difficult to address. Consequently, in many cases the hurdles to experimentation will hopefully cause scholars to recognize the relevance of other methods to answering particular questions. Second, the steps that make for a good experiment are exactly those which hopefully prevent experiments from constricting the questions asked – experiments are worthwhile when they constitute an appropriate method to address the question embedded in a larger research enterprise. Put another way, if the constriction concern proves accurate, not only might social scientists narrow the questions they address but they may do so in less than ideal ways (e.g., the experiments will not themselves be compelling).

Alas, I think there is reason for hope that experiments have not yet constricted questions generally or in a way that has substantially undermined quality. The last two decades have seen a host of foundational contributions from experiments and most are used in sound, thoughtful ways that advance knowledge. Social science experimentation can evolve in even more exciting ways given the new opportunities that come with widespread acceptance of the method and technological and sociological advances. But, this will only happen if scholars "think" about experiments carefully, recognizing not only their possibilities but also their limitations: any experiment requires making a host of assumptions, inferential leaps, and much tedious work. As I stated at the outset, experiments seem often to be designed and implemented quickly and not connected to the full scientific process. This is a problem; experiments need to be thought of as one part of a scientific process and not the first part. They need to be used when appropriate and build on / have an interplay with questions, observations, and theory. Moreover, conducting a quality experiment requires thinking through a litany of decisions discussed through the book. A

good experiment is slow moving (given the host of considerations) which is counter to the current fast moving temptations available in the social sciences.

Campbell and Stanley (1963: 3) worried about the poverty of experiments stemming from a lack of data and null results. Those concerns increasingly have become less relevant; however, a new poverty of poor designs, inappropriate analyses, limited use of data, and/or flawed interpretation is always a possibility. Preventing that poverty requires engaging in careful experimental thinking that will ensure the sustained contributions of experiments to the accumulation of social science knowledge.

References

Aczel, Balazs, Barnabas Szaszi, Alexandra Sarafoglou, Zoltan Kekecs, Šimon Kucharský, Daniel Benjamin, Christopher D. Chambers, Agneta Fisher, Andrew Gelman, Morton A. Gernsbacher, John P. Ioannidis, Eric Johnson, Kai Jonas, Stavroula Kousta, Scott O. Lilienfeld, D. Stephen Lindsay, Candice C. Morey, Marcus Monafò, Benjamin R. Newell, Harold Pashler, David R. Shanks, Daniel J. Simons, Jelte M. Wicherts, Dolores Albarracin, Nicole D. Anderson, John Antonakis, Hal R. Arkes, Mitja D. Back, George C. Banks, Christopher Beevers, Andrew A. Bennett, Wiebke Bleidorn, Ty W. Boyer, Cristina Cacciari, Alice S. Carter, Joseph Cesario, Charles Clifton, Ronán M. Conroy, Mike Cortese, Fiammetta Cosci, Nelson Cowan, Jarret Crawford, Eveline A. Crone, John Curtin, Randall Engle, Simon Farrell, Pasco Fearon, Mark Fichman, Willem Frankenhuis, Alexandra M. Freund, M. Gareth Gaskell, Roger Giner-Sorolla, Don P. Green, Robert L. Greene, Lisa L. Harlow, Fernando Hoces de la Guardia, Derek Isaacowitz, Janet Kolodner, Debra Lieberman, Gordon D. Logan, Wendy B. Mendes, Lea Moersdorf, Brendan Nyhan, Jeffrey Pollack, Christopher Sullivan, Simine Vazire and Eric-Jan Wagenmakers. 2020. "A Consensus-Based Transparency Checklist." *Nature Human Behaviour* 4: 4-6.

Alferes, Valentim R. 2012. *Methods of Randomization in Experimental Design*. Thousand Oaks, CA: Sage Publications.

Allport, Gordon W. 1954. *The Nature of Prejudice.* New York, NY: Basic Books.

Al-Ubaydli, Omar, Min Sok Lee, John A. List, Claire L. Mackevicius, and Dana Suskind. 2020a. "How Can Experiments Play a Greater Role in Public Policy?: Twelve Proposals from an Economic Model of Scaling." *Behavioural Public Policy*, doi:10.1017/bpp.2020.17.

Al-Ubaydli, Omar, John A. List and Dana L. Suskind. 2017. “What Can We Learn from Experiments?: Understanding the Threats to the Scalability of Experimental Results.” *American Economic Review: Papers & Proceedings* 107: 282-286. https://doi.org/10.1257/aer.p20171115

Al-Ubaydli, Omar, John A. List, and Dana Suskind. 2020b. “The Science of Using Science: Towards an Understanding of the Threats to Scalability.” *International Economic Review* doi: 10.1111/iere.12476.

American Political Science Association (APSA). 2012. *A Guide to Professional Ethics in Political Science*. Washington DC: American Political Science Association.

Anderson, Christopher J., Štěpán Bahník, Michael Barnett-Cowan, Frank A. Bosco, Jesse Chandler, Christopher R. Chartier, Felix Cheung, Cody D. Christopherson, Andreas Cordes, Edward J. Cremata, Nicolas Della Penna, Vivien Estel, Anna Fedor, Stanka A. Fitneva, Michael C. Frank, James A. Grange, Joshua K. Hartshorne, Fred Hasselman, Felix Henninger, Marije van der Hulst, Kai J. Jonas, Calvin K. Lai, Carmel A. Levitan, Jeremy K. Miller, Katherine S. Moore, Johannes M. Meixner, Marcus R. Munafò, Koen I. Neijenhuijs, Gustav Nilsonne, Brian A. Nosek, Franziska Plessow, Jason M. Prenoveau, Ashley A. Ricker, Kathleen Schmidt, Jeffrey R. Spies, Stefan Stieger, Nina Strohminger, Gavin B. Sullivan, Robbie C. M. van Aert, Marcel A. L. M. van Assen, Wolf Vanpaemel, Michelangelo Vianello, Martin Voracek, Kellylynn Zuni. 2016. “Response to Comment on ‘Estimating the Reproducibility of Psychological Science’.” *Science* 351: 1037-c.

Anderson, Richard G. 2013. “Registration and Replication: A Comment.” *Political Analysis* 21: 38-39.

Andrews, Isaiah, and Maximilian Kasy. 2019. "Identification of and Correction for Publication Bias." *American Economic Review* 109: 2766-2794.

Angrist, Joshua D. 1990. "Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records." *The American Economic Review* 80: 313-336.

Angrist, Joshua D., Guido W. Imbens, and Donald B. Rubin. 1996. "Identification of Causal Effects Using Instrumental Variables." *Journal of the American Statistical Association* 91: 444-455.

Ansolabehere, Stephen, Shanto Iyengar, Adam Simon, and Nicholas Valentino. 1994. "Does Attack Advertising Demobilize the Electorate?" *American Political Science Review* 88: 829-838.

Ansolabehere, Stephen, Jonathan Rodden, and James M. Snyder Jr. 2008. "The Strength of Issues: Using Multiple Measures to Gauge Preference Stability, Ideological Constraint, and Issue Voting." *American Political Science Review* 102: 215-232

Arceneaux, Kevin, Alan S. Gerber, and Donald P. Green. 2006. "Comparing Experimental and Matching Methods Using a Large-Scale Voter Mobilization Experiment." *Political Analysis* 14: 37–62.

Arceneaux, Kevin, and Martin Johnson. 2013. *Changing Minds or Changing Channels?: Partisan News in Age of Choice*. Chicago: University Chicago Press.

Arechar, Antonio A., Gordon T. Kraft-Todd, and David G. Rand. 2017. "Turking Overtime: How Participant Characteristics and Behavior Vary over Time and Day on Amazon Mechanical Turk." *Journal of the Economic Science Association* 3: 1-11.

Arechar, Antonio A., and David G. Rand. 2020. "Turking in the Time of COVID." Working Paper, Massachusetts Institute of Technology, https://psyarxiv.com/vktqu.

Aronow, Peter M., Josh Kalla, Lilla Orr, and John Ternovski. 2020. “Evidence of Rising Rates of Inattentiveness on Lucid in 2020.” Working Paper, Yale University.

Aronson, Elliot, Marilynn B. Brewer, and J. Merill Carlsmith. 1985. “Experimentation in Social Psychology.” In Gardner Lindzey and Elliot Aronson, eds., *Handbook of Social Psychology* 3rd Edition. New York: Random House.

Aronson, Elliot, and J. Merill Carlsmith. 1968. “Experimentation in Social Psychology.” In Gardner Lindzey and Elliot Aronson, eds., *Handbook of Social Psychology* 2nd Edition. Reading, MA: Addison-Wesley.

Aronson, Elliot, Timothy D. Wilson, Marilynn B. Brewer. 1998. “Experimentation in Social Psychology.” In Daniel T. Gilbert, Susan T. Fiske, and Gardner Lindzey, eds., *The Handbook of Social Psychology*. 4th Edition. Boston: The McGraw-Hill Companies, Inc.

Asch, Solomon E. 1956. “Studies of Independence and Conformity: I. A Minority of One Against a Unanimous Majority.” *Psychological Monographs: General and Applied* 70: 1-70.

Baker, Monya, 2016. “Is There a Reproducibility Crisis?” *Nature* 533: 452-455.

Bandiera, Oriana, Andrea Prat, and Tommaso Valletti. 2009. “Active and Passive Waste in Government Spending: Evidence from a Policy Experiment.” *American Economic Revi*ew 99: 1278-1308.

Banerjee, Abhijit V., and Esther Duflo. 2009. “The Experimental Approach to Development Economics.” *Annual Review of Economics* 1: 151-178.

Bansak, Kirk, Jens Hainmueller, Daniel J. Hopkins, Teppei Yamamoto. 2021. “Conjoint Survey Experiments" In James N. Druckman, and Donald P. Green, eds., *Advances in Experimental Political Science*. New York: Cambridge University Press.

Barabas, Jason, and Jennifer Jerit. 2010. “Are Survey Experiments Externally Valid?” *American Political Science Review* 104: 226-242.

Baron, Reuben M., and David A. Kenny. 1986. “The Moderator–Mediator Variable Distinction in Social Psychological Research: Conceptual, Strategic, and Statistical Considerations.” *Journal of Personality and Social Psychology* 51: 1173-1182.

Bassi, Anna. 2020. “Experiments.” In Dirk Berg-Schlosser, Bertrand Badie, Leonardo Morlino, eds., *The SAGE Handbook of Political Science*. Thousand Oaks, CA: Sage Publications.

Bayes, Robin, James N. Druckman, Avery Goods, Daniel C. Molden. 2020. “When and How Different Motives Can Drive Motivated Political Reasoning.” *Political Psychology* 41: 1031-1052.

Beath, Andrew, Fotini Christia, and Ruben Enikolopov. 2013. “Empowering Women through Development Aid: Evidence from a Field Experiment in Afghanistan.” *American Political Science Review* 107: 540-557.

Beatty, Paul C., and Gordon B. Willis. 2007. “Research Synthesis: The Practice of Cognitive Interviewing.” *Public Opinion Quarterly* 71: 287–311.

Bechtel, Michael M., and Kenneth F. Scheve. 2013. “Mass Support for Global Climate Agreements Depends on Institutional Design.” *Proceedings of the National Academy of Sciences* 110: 13763-13768.

Belli, Robert F., Sean E. Moore, John VanHoewyk. 2006. “An Experimental Comparison of Question Forms Used to Reduce Vote Overreporting.” *Electoral Studies* 25: 751-759.

Ben-Akiva, Moshe, Daniel McFadden, and Kenneth Train. 2019. “Foundations of Stated Preference Elicitation: Consumer Behavior and Choice-Based Conjoint Analysis.” *Foundations and Trends® in Econometrics* 10: 1-144.

Benjamin, Daniel J., James O. Berger, Magnus Johannesson, Brian A.Nosek, E.-J. Wagenmakers, Richard Berk, Kenneth A. Bollen, Björn Brembs, Lawrence Brown, Colin Camerer, David Cesarini, Christopher D. Chambers, Merlise Clyde, Thomas D. Cook, Paul De Boeck, Zoltan Dienes, Anna Dreber, Kenny Easwaran, Charles Efferson, Ernst Fehr, Fiona Fidler, Andy P. Field, Malcolm Forster, Edward I. George, Richard Gonzalez, Steven Goodman, Edwin Green, Donald P. Green, Anthony Greenwald, Jarrod D. Hadfield, Larry V. Hedges, Leonhard Held, Teck Hua Ho, Herbert Hoijtink, James Holland Jones, Daniel J. Hruschka, Kosuke Imai, Guido Imbens, John P.A. Ioannidis, Minjeong Jeon, Michael Kirchler, David Laibson, John List, Roderick Little, Arthur Lupia, Edouard Machery, Scott E. Maxwell, Michael McCarthy, Don Moore, Stephen L. Morgan, Marcus Munafó, Shinichi Nakagawa, Brendan Nyhan, Timothy H. Parker, Luis Pericchi, Marco Perugini, Jeff Rouder, Judith Rousseau, Victoria Savalei, Felix D. Schönbrodt, Thomas Sellke, Betsy Sinclair, Dustin Tingley, Trisha Van Zandt, Simine Vazire, Duncan J. Watts, Christopher Winship, Robert L. Wolpert, Yu Xie, Cristobal Young, Jonathan Zinman, and Valen E. Johnson. 2018. “Redefine Statistical Significance.” *Nature Human Behavior* 2: 6-10.

Berent, Matthew K., Jon A. Krosnick., and Arthur Lupia. 2016. “Measuring Voter Registration and Turnout in Surveys.” *Public Opinion Quarterly* 80: 597–621.

Bergan, Daniel E., and Richard T. Cole. 2015. “Call Your Legislator: A Field Experimental Study of the Impact of a Constituency Mobilization Campaign on Legislative Voting.” *Political Behavior* 37: 27-42.

Berinsky, Adam J., James N. Druckman, and Teppei Yamamoto. 2020. “Publication Biases in Replication Studies.” *Political Analysis* 10.1017/pan.2020.34.

Berinsky, Adam J., Gregory A. Huber, and Gabriel S. Lenz. 2012. “Evaluating Online Labor Markets for Experimental Research: Amazon.com's Mechanical Turk.” *Political Analysis* 20: 351-68.

Berinsky, Adam J., Michele F. Margolis, and Michael W. Sances. 2014. “Separating the Shirkers from the Workers? Making Sure Respondents Pay Attention on Self-Administered Surveys.” *American Journal of Political Science* 58: 739-753.

Berkowitz, Leonard, and Edward Donnerstein. 1982. “External Validity is More Than Skin Deep: Some Answers to Criticisms of Laboratory Experiments.” *American Psychologist* 37: 245-257.

Bethlehem, Jelke, and Mario Callegaro. 2014. “Introduction to Part IV: Weighting Adjustments.” In Mario Callegaro, Reg Baker, Jelke Bethlehem, Anja S. Göritz, Jon A. Krosnick, and Paul J. Lavrakas. *Online Panel Research: A Data Quality Perspective*. West Sussex, UK: John Wiley & Sons, Ltd.

Biemer, Paul P. 2010. “Total Survey Error: Design, Implementation, and Evaluation.” *Public Opinion Quarterly* 74: 817-848.

Bigler, Rebecca S., and Julie Milligan Hughes. 2010. “Reasons for Skepticism about the Efficacy of Simulated Social Contact Interventions.” *American Psychologist* 65: 132-133.

Bilgen, Ipek, J. Michael Dennis, Nadarajasundaram Ganesh. 2018. “Measuring the Undercounted in Policy Attitude Surveys: Probability-based Panel Recruitment Nonresponse Follow-up Impact on Sample Composition and Outcome Measures.” Working paper, NORC at the University of Chicago.

Bisgaard, Martin. 2019. “How Getting the Facts Right Can Fuel Partisan-Motivated Reasoning.” *American Journal of Political Science* 63: 824–839.

Blair, Edward, and Johnny Blair. 2015. *Applied Survey Sampling*. Los Angeles, CA: Sage Publications.

Blair, Graeme, Jasper Cooper, Alexander Coppock, Macartan Humphreys. 2019. "Declaring and Diagnosing Research Designs." *American Political Science Review* 113:838-859.

Blair, Graeme, Alexander Coppock, and Margaret Moor. 2020. "When to Worry About Sensitivity Bias: Evidence from 30 Years of List Experiments." *American Political Science Review* 114: 1297-1315.

Blair, Graeme, and Gwyneth McClendon. 2021. "Conducting Experiments in Multiple Contexts." In James N. Druckman, and Donald P. Green, eds., *Advances in Experimental Political Science*. New York: Cambridge University Press.

Blalock, Hubert M. 1967. *Toward a Theory of Minority-Group Relations*. John Wiley & Sons.

Blatt, Jessica. 2018. *Race and the Making of American Political Science*. Philadelphia: University of Pennsylvania Press.

Blattman, Christopher, Alexandria C. Hartman, and Robert A. Blair. 2014. "How to Promote Order and Property Rights under Weak Rule of Law?: An Experiment in Changing Dispute Resolution Behavior through Community Education." *American Political Science Review* 108: 100-120.

Bloom, Howard S., ed. 2005. *Learning More from Social Experiments: Evolving Analytic Approaches*. New York: Russell Sage Foundation.

Bollen, Kenneth, John T. Cacioppo, Robert M. Kaplan, Jon A. Krosnick, James L. Olds, and Heather Dean, 2015. "Social, Behavioral, and Economic Sciences Perspectives on Robust and Reliable Science." *Report of the Subcommittee on Replicability in Science Advisory Committee to the National Science Foundation Directorate for Social, Behavioral, and*

Economic Sciences. https://www.nsf.gov/sbe/AC_Materials/SBE_Robust_and_Reliable_Research_Report.pdf

Bolsen, Toby, James N. Druckman, and Fay Lomax Cook. 2015. "Citizens', Scientists', and Legislators' Beliefs about Global Climate Change." *The Annals of the American Academy of Political and Social Science* 658: 271-295.

Bond, Robert M. Christopher J. Fariss, Jason J. Jones, Adam D. I. Kramer, Cameron Marlow, Jaime E. Settle, and James H. Fowler. 2012. "A 61-Million-Person Experiment in Social Influence and Political Mobilization." *Nature* 489: 295-298.

Bond, Rod, and Peter B. Smith. 1996. "Culture and Conformity: A Meta-Analysis of Studies Using Asch's (1952b, 1956) Line Judgment Task." *Psychological Bulletin* 119: 111–137.

Borenstein, Michael, Larry V. Hedges, Julian P. T. Higgins, and Hannah R. Rothstein. 2009. *Introduction to Meta-Analysis*. Chichester, UK: John Wiley & Sons, Ltd.

Boudreau, Cheryl. 2021. "Transparency in Experimental Research." In James N. Druckman, and Donald P. Green, eds., *Advances in Experimental Political Science*. New York: Cambridge University Press.

Brandt, Mark J., Hans IJzerman, Ap Dijksterhuis, Frank J. Farach, Jason Geller, Roger Giner-Sorolla, James A. Grange, Marco Perugini, Jeffrey R. Spies, Anna van 't Veer. 2014. "The Replication Recipe: What Makes for a Convincing Replication?" *Journal of Experimental Social Psychology* 50: 217-224.

Brodeur, Abel, Nikolai Cook, and Anthony Heyes. 2020. "Methods Matter: p-Hacking and Publication Bias in Causal Analysis in Economics." *American Economic Review* 110: 3634-3660.

Brown, Andrew W., Tapan S. Mehta, and David B. Allison. 2017. "Publication Bias in Science." In Kathleen Hall Jamieson, Dan M. Kahan, and Dietram A. Scheufele, eds., *The Oxford Handbook of the Science of Science Communication*. New York: Oxford University Press.

Buhrmester, Michael, Tracy Kwang, and Samuel D. Gosling. 2011. "Amazon's Mechanical Turk: A New Source of Inexpensive, Yet High-Quality, Data?" *Perspectives on Psychological Science* 6: 3-5.

Bulkeley, Harriet, and Vanesa Castán Broto. 2013. "Government by Experiment?: Global Cities and the Governing of Climate Change." *Transactions of the Institute of British Geographers* 38: 361-375.

Bullock, John G. 2011. "Elite Influence on Public Opinion in an Informed Electorate." *American Political Science Review* 105: 496-515.

Bullock, John G. 2020. "Party Cues." In Elizabeth Suhay, Bernard Grofman, and Alexander H. Trechsel, eds., *Oxford Handbook of Electoral Persuasion.* New York: Oxford University Press.

Bullock, John G., Alan S. Gerber, Seth J. Hill, and Gregory A. Huber. 2015. "Partisan Bias in Factual Beliefs about Politics." *Quarterly Journal of Political Science* 10: 519-578.

Bullock, John G., and Donald P. Green. 2020. "The Failings of Conventional Mediation Analysis and a Design-Based Alternative." Working Paper, Northwestern University.

Bullock, John G., and Shang E. Ha. 2011. "Mediational Analysis Is Harder Than It Looks." In James N. Druckman, Donald P. Green, James H. Kuklinski, and Arthur Lupia, eds., *Cambridge Handbook of Experimental Political Science*. New York, NY: Cambridge University Press**.**

Burden, Barry C. 2000. “Voter Turnout and the National Election Studies.” *Political Analysis* 8: 389-398.

Burden, Barry C., Yoshikuni Ono, and Masahiro Yamada. 2017. “Reassessing Public Support for a Female President.” *The Journal of Politics* 79: 1073-1078.

Burnham, Terence C., and Robert Kurzban. 2005. “On the Limitations of Quasi-Experiments.” *Behavioral and Brain Sciences* 28: 818-819.

Busby, Ethan C., and James N. Druckman. 2018. “Football and Public Opinion: A Partial Replication and Extension.” *Journal of Experimental Political Science* 5: 4-10.

Busby, Ethan C., James N. Druckman, and Alexandria Fredendall. 2017. “The Political Relevance of Irrelevant Events.” *The Journal of Politics* 79: 346-350.

Butler, Daniel M. 2014. *Representing the Advantaged: How Politicians Reinforce Inequality*. New York: Cambridge University Press.

Butler, Daniel M., and David E. Broockman. 2011. “Do Politicians Racially Discriminate Against Constituents?: A Field Experiment on State Legislators.” *American Journal of Political Science* 55: 463-477.

Butler, Daniel M., and Charles Crabtree. 2017. “Moving Beyond Measurement: Adapting Audit Studies to Test Bias-Reducing Interventions.” *Journal of Experimental Political Science* 4: 57-67.

Butler, Daniel M., and Charles Crabtree. 2021. “Audit Studies in Political Science.” In James N. Druckman, and Donald P. Green, eds., *Advances in Experimental Political Science*. New York: Cambridge University Press.

Butler, Daniel M., and Jonathan Homola. 2017. “An Empirical Justification for the Use of Racially Distinctive Names to Signal Race in Experiments.” *Political Analysis* 25: 122-130.

Buzin, Andrei, Kevin Brondum, and Graeme Robertson. 2016. “Election Observer Effects: A Field Experiment in the Russian Duma Election of 2011.” *Electoral Studies* 44: 184-191.

Cacioppo, John T., Richard E. Petty, and Katherine J. Morris. 1983. “Effects of Need for Cognition on Message Evaluation, Recall, and Persuasion.” *Journal of Personality and Social Psychology 45*: 805-818.

Camerer, Colin G. Anna Dreber, Rskil Forsell, Teck-Hua Ho, Jürgen Huber, Magnus Johannesson, Michael Kirchler, Johan Almenberg, Adam Altmejd, Taizan Chan, Emma Heikensten, Felix Holzmeister, Taisuke Imai, Siri Isaksson, Gideon Nave, Thomas Pfeiffer, Michael Razen, Hang Wu. 2016. “Evaluating Replicability of Laboratory Experiments in Economics.” *Science* 351: 1433-1436.

Camerer, Colin F, Anna Dreber, Felix Holzmeister, Teck-Hua Ho, Jürgen Huber, Magnus Johannesson, Michael Kirchler, Gideon Nave, Brian A. Nosek, Thomas Pfeiffer, Adam Altmejd, Nick Buttrick, Taizan Chan, Yiling Chen, Eskil Forsell, Anup Gampa, Emma Heikensten, Lily Hummer, Taisuke Imai, Siri Isaksson, Dylan Manfredi, Julia Rose, Eric-Jan Wagenmakers and Hang Wu. 2018. “Evaluating the Replicability of Social Science Experiments in *Nature* and *Science* between 2010 and 2015.” *Nature Human Behaviour* 2: 637-644.

Campbell, Donald T. 1969. “Prospective: Artifact and Control.” In Robert Rosenthal and Robert Rosnow, eds., *Artifact in Behavioral Research*. New York: Academic Press.

Campbell Donald. T. Donald W. Fiske. 1959. “Convergent and Discriminant Validation by the Multitrait-Multimethod Matrix.” *Psychological Bulletin* 56: 81-105.

Campbell, Donald T., and Julian C. Stanley. 1963. *Experimental and Quasi-Experimental Designs for Research*. Chicago: Rand McNally & Company.

Campbell, Susanna P., and Gabriele Spilker. 2020. “Aiding War or Peace? The Insiders’ View on Aid to Post-Conflict Transitions.” Working paper, https://ssrn.com/abstract=3576116.

Card, David, and Alan Krueger. 1994. “Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania.” *American Economic Review* 84: 772-793.

Cárdenas, Juan Camilo, and Jeffrey Carpenter. 2008. Behavioural Development Economics: Lessons from Field Labs in the Developing World.” *Journal of Development Studies* 44: 337-364.

Chmielewski, Michael, and Sarah C. Kucker. 2019. “An MTurk Crisis?: Shifts in Data Quality and the Impact on Study Results.” *Social Psychological and Personality Science*. https://doi.org/10.1177/1948550619875149.

Chong, Dennis, and James N Druckman. 2007. “Framing Public Opinion in Competitive Democracies.” *American Political Science Review* 101: 637-655.

Christensen, Garret, Jeremy Freese, and Edward Miguel 2019. T*ransparent and Reproducible Social Science Research: How to Do Open Science*. Oakland, CA: University of California Press.

Christensen, Garret, and Edward Miguel. 2018. “Transparency, Reproducibility, and the Credibility of Economics Research.” *Journal of Economic Literature* 56: 920-980.

Cirone, Alexandra, and Brenda Van Coppenolle. 2018. "Cabinets, Committees, and Careers: The Causal Effect of Committee Service" *The Journal of Politics* 80: 948-963.

Clifford, Scott, Thomas J. Leeper, and Carlisle Rainey. 2019. "Increasing the Generalizability of Survey Experiments Using Randomized Topics: An Application to Party Cues." Paper presented at the annual meeting of the American Political Science Association, Washington, D.C.

Coffman, Lucas C., and Muriel Niederle. 2015. "Pre-Analysis Plans Have Limited Upside, Especially Where Replications Are Feasible." *The Journal of Economic Perspectives* 29: 81-97.

Cohn, Nate and Kevin Quealy. 2019. "The Democratic Electorate on Twitter Is Not the Democratic Electorate." *The New York Times,* April 9th.

Collins H.M. 1985. *Changing Order: Replication and Induction in Scientific Practice*. Chicago: University of Chicago Press.

Collins Harry. 2016. "Reproducibility of Experiments: Experimenters' Regress, Statistical Uncertainty Principle, and the Replication Imperative." In Harald Atmanspacher, and Sabine Maasen, eds., *Reproducibility: Principles, Problems, and Prospects*. New York: Wiley.

Connors, Elizabeth C., Yanna Krupnikov, and John Barry Ryan. 2019. "How Transparency Affects Survey Responses." *Public Opinion Quarterly* 83: 185-209.

Cook, Thomas D. 2002. "Randomized Experiments in Educational Policy Research: A Critical Examination of the Reasons the Educational Evaluation Community Has Offered for Not Doing Them." *Educational Evaluation and Policy Analysis* 24: 175-199.

Cook, Thomas D., and Donald T. Campbell. 1979. *Quasi-Experimentation: Design and Analysis Issues for Field Settings*. Chicago: Rand-McNally.

Cooper, Russell, Douglas V. DeJong, Robert Forsythe, and Thomas W. Ross. 1993. "Forward Induction in the Battle-of-the-Sexes Games." *American Economic Review* 83: 1303-1316.

Coppock, Alexander. 2019a. "Avoiding Post-Treatment Bias in Audit Experiments." *Journal of Experimental Political Science* 6: 1-4.

Coppock, Alexander. 2019b. "Generalizing from Survey Experiments Conducted on Mechanical Turk: A Replication Approach." *Political Science Research and Methods* 7: 613-628.

Coppock, Alexander, and Donald P. Green. 2015. "Assessing the Correspondence between Experimental Results Obtained in the Lab and Field: A Review of Recent Social Science Research." *Political Science Research Methods* 3: 113-131.

Coppock, Alexander, Seth J. Hill, and Lynn Vavreck. 2020. "The Small Effects of Political Advertising Are Small Regardless of Context, Message, Sender, or Receiver: Evidence from 59 Real-time Randomized Experiments." *Science Advances* 6: eabc4046.

Coppock, Alexander, Thomas J. Leeper, and Kevin J. Mullinix. 2018. "Generalizability of Heterogeneous Treatment Effect Estimates across Samples." *Proceedings of the National Academy of Sciences* 115: 12441-12446.

Costa, Mia. 2017. "How Responsive Are Political Elites?" *Journal of Experimental Political Science* 4: 241-254.

Cox, David R. 1958. *Planning of Experiments*. New York: Wiley.

Crabtree, Charles, and Volha Chykina. 2018. "Last Name Selection in Audit Studies." *Sociological Science* 5: 21-28.

Craig, Maureen A., Julian M. Rucker, and Jennifer A. Richeson. 2018. "Racial and Political Dynamics of an Approaching 'Majority-Minority' United States." *The ANNALS of the American Academy of Political and Social Science* 677: 204-214.

Crawfurd, Lee. 2020. "Contact and Commitment to Development: Evidence from Quasi-Random Missionary Assignments." *Kyklos*. https://doi.org/10.1111/kykl.12255.

Crisp, Richard J., Sofia Stathi, Rhiannon N. Turner, and Senel Husnu. 2009. "Imagined Intergroup Contact: Theory, Paradigm and Practice." *Social and Personality Psychology Compass* 3: 1–18.

Crisp, Richard J., and Rhiannon N. Turner. 2009. "Can Imagined Interactions Produce Positive Perceptions?: Reducing Prejudice Through Simulated Contact. *American Psychologist* 64 231-240.

Cumming, Geoff. 2008. "Replication and p Intervals: P Values Predict the Future Only Vaguely, but Confidence Intervals Do Much Better." *Perspectives on Psychological Science* 3: 286-300.

Dafoe, Allan, Baobao Zhang, and Devin Caughey. 2018. "Information Equivalence in Survey Experiments." *Political Analysis* 26: 399-416.

Dahl, Robert A. 1956. *A Preface to Democratic Theory*. Chicago: University of Chicago Press.

Dahl, Robert A. 1971. *Polyarchy: Participation and Opposition*. New Haven: Yale University Press.

Davis, Taylor, Erin P. Hennes, and Leigh Raymond. 2018. "Cultural Evolution of Normative Motivations for Sustainable Behaviour." *Nature Sustainability* 1: 218-224.

de Rooij, Eline A., Donald P. Green, and Alan S. Gerber. 2009. "Field Experiments on Political Behavior and Collective Action." *Annual Review of Political Science* 12 (1):389–95.

DeBell, Matthew. 2018. "Best Practices for Creating Survey Weights." In David Vannette and Jon Krosnick, eds., *The Palgrave Handbook of Survey Research*. Cham, Switzerland: Palgrave Macmillan.

DeBell, Matthew, Jon A. Krosnick, Katie Gera, David S. Yeager, and Michael P. McDonald. 2018. "The Turnout Gap in Surveys: Explanations and Solutions." *Sociological Methods & Research*. https://doi.org/10.1177/0049124118769085.

Delli Carpini, Michael X., and Scott Keeter. 1996. *What Americans Know About Politics and Why it Matters*. New Haven: Yale University Press.

Desposato, Scott, ed. 2016. *Ethics and Experiments: Problems and Solutions for Social Scientists and Policy Professionals*. New York: Routledge.

Devezer, Berna, Danielle J. Navarro, Joachim Vandekerckhove, Erkan Ozge Buzbas. 2020. "The Case for Formal Methodology in Scientific Reform." Working Paper, University of Idaho. bioRxiv preprint, https://doi.org/10.1101/2020.04.26.048306.

Dickhaut, John W., J. Leslie Livingstone, and David J. H. Watson. 1972. "On the Use of Surrogates in Behavioral Experimentation." *The Accounting Review* 47, Supplement: 455-471.

Disch, Lisa. 2011. "Toward a Mobilization Conception of Democratic Representation." *American Political Science Review* 105: 100-114.

Doherty, Daniel, Alan S. Gerber, and Donald P. Green. 2006. "Personal Income and Attitudes toward Redistribution: A Study of Lottery Winners." *Political Psychology* 27: 441-458.

Doleac, Jennifer L, and Benjamin Hansen. 2019. "The Unintended Consequences of 'Ban the Box': Statistical Discrimination and Employment Outcomes When Criminal Histories

Are Hidden.”. *Journal of Labor Economics*. https://www.journals.uchicago.edu/doi/10.1086/705880.

Doyen, Stéphane , Olivier Klein, Cora-Lise Pichon, and Axel Cleeremans. 2012. “Behavioral Priming: It's All in the Mind, but Whose Mind? *PLoS One* 7: e29081.

Dreber, Anna, Thomas Pfeiffer, Johan Almenberg, Siri Isaksson, Brad Wilson, Yiling Chen, Brian A. Nosek, and Magnus Johannesson. 2015. “Using Prediction Markets to Estimate the Reproducibility of Scientific Research.” *Proceedings of the National Academy of Sciences* 112: 15343-15347.

Druckman, James N. 1996. “Party Factionalism and Cabinet Durability.” *Party Politics* 2: 397-407.

Druckman, James N. 2001. “On The Limits Of Framing Effects: Who Can Frame?” *The Journal of Politics* 63: 1041-1066.

Druckman, James N. 2004. “Priming the Vote: Campaign Effects in a U.S. Senate Election.” *Political Psychology* 25: 577-594.

Druckman, James N. 2014. “Pathologies of Studying Public Opinion, Political Communication, and Democratic Responsiveness.” *Political Communication* 31: 467-492.

Druckman, James N. 2015. “Merging Research and Undergraduate Teaching in Political Behavior Research.” *PS: Political Science & Politics* 48: 53-57.

Druckman, James N., Jordan Fein, and Thomas J. Leeper. 2012. “A Source of Bias in Public Opinion Stability.” *American Political Science Review* 106: 430-454.

Druckman, James N., Mauro Gilli, Samara Klar, and Joshua Robison. 2015. “Measuring Drug and Alcohol Use Among College Student-Athletes.” *Social Science Quarterly* 96: 369-380.

Druckman, James N., Cari Lynn Hennessy, Kristi St. Charles, and Jonathan Weber. 2010. "Competing Rhetoric Over Time: Frames Versus Cues." *The Journal of Politics* 72: 136-148.

Druckman, James N., and Donald P. Green, eds. 2021. *Advances in Experimental Political Science*. New York: Cambridge University Press.

Druckman, James N., Donald P. Green, James H. Kuklinski, and Arthur Lupia. 2006. The Growth and Development of Experimental Research Political Science. *American Political Science Review* 100: 627-636.

Druckman, James N., Donald P. Green, James H. Kuklinski, and Arthur Lupia, eds. 2011. *Cambridge Handbook of Experimental Political Science*. New York: Cambridge University Press.

Druckman, James N., S.R. Gubitz, Matthew S. Levendusky, and Ashley Lloyd. 2019. "How Incivility On Partisan Media (De-)Polarizes the Electorate." *The Journal of Politics* 81: 291-295.

Druckman, James N., Adam J. Howat, and Kevin J. Mullinix. 2018a. "Graduate Advising in Experimental Research Groups." *PS: Political Science & Politics* 51: 620-624.

Druckman, James N., and Cindy D. Kam. 2011. "Students as Experimental Participants: A Defense of the 'Narrow Data Base'." In James N. Druckman, Donald P. Green, James H. Kuklinski, and Arthur Lupia, eds., *Cambridge Handbook of Experimental Political Science*, New York: Cambridge University Press.

Druckman, James N., Samara Klar, Yanna Krupnikov, Matthew Levendusky and John Barry Ryan. 2020a. "Affective Polarization, Local Contexts, and Public Opinion in America." *Nature Human Behavior* https://doi.org/10.1038/s41562-020-01012-5.

Druckman, James N., Samara Klar, Yanna Krupnikov, Matthew Levendusky and John Barry Ryan. N.d. "(Mis-)Estimating Affective Polarization." *The Journal of Politics*, Forthcoming.

Druckman, James N., and Thomas J. Leeper. 2012a. "Is Public Opinion Stable?: Resolving the Micro/Macro Disconnect in Studies of Public Opinion." *Daedalus* 141: 50-68.

Druckman, James N., and Thomas J. Leeper. 2012b. "Learning More from Political Communication Experiments: Pretreatment and Its Effects." *American Journal of Political Science* 56: 875-896.

Druckman, James N., and Matthew S. Levendusky. 2019. "What Do We Measure When We Measure Affective Polarization?" *Public Opinion Quarterly* 83: 114-122.

Druckman, James N., Matthew S. Levendusky, and Audrey McLain. 2018b. "No Need to Watch: How the Effects of Partisan Media Can Spread via Inter-Personal Discussions." *American Journal of Political Science* 62: 99-112.

Druckman, James N., Jeremy Levy, and Natalie Sands. 2020b. "Minimizing Bias in Higher Education Disability Accommodation Services." Working Paper, Northwestern University.

Druckman, James N., and Arthur Lupia. 2006. "Mind, Will, and Choice: Lessons From Experiments in Contextual Variation." In Robert E. Goodin and Charles Tilly, eds., *The Oxford Handbook of Contextual Political Analysis*. Oxford: Oxford University Press.

Druckman, James N., and Mary C. McGrath. 2019. "The Evidence for Motivated Reasoning in Climate Change Preference Formation." *Nature Climate Change* 9: 111-119.

Druckman, James N., and Kjersten R. Nelson. 2003. "Framing and Deliberation: How Citizens' Conversations Limit Elite Influence." *American Journal of Political Science* 47: 729-745.

Druckman, James N., Erik Peterson, and Rune Slothuus. 2013. "How Elite Partisan Polarization Affects Public Opinion Formation." *American Political Science Review* 170: 57-79.

Druckman, James N., and Richard M. Shafranek. 2020. "The Intersection of Racial and Partisan Discrimination: Evidence from a Correspondence Study of Four-Year Colleges." *The Journal of Politics* 82: 1602-1606.

Druckman, James N., Sophie Trawalter, Ivonne Montes, Alexandria Fredendall, Noah Kanter, and Allison Paige Rubenstein. 2018c. "Racial Bias in Sport Medical Staff's Perceptions of Others' Pain." *The Journal of Social Psychology* 158: 721-729.

Druckman, James N., and Julia Valdes. 2019. "How Private Politics Alters Legislative Responsiveness." *Quarterly Journal of Political Science* 14: 115-130.

Dunbar, Kevin, and Jonathan Fugelsang. 2005. "Scientific Thinking and Reasoning." In Keith J. Holyoak, and Roberg G. Morrison, eds., *The Cambridge Handbook of Thinking and Reasoning*. Cambridge: Cambridge University Press.

Dunning, Thad. 2012. *Natural Experiments in the Social Sciences: A Design-Based Approach*. Strategies for Social Inquiry. New York: Cambridge University Press.

Dunning, Thad. 2016. "Transparency, Replication, and Cumulative Learning: What Experiments Alone Cannot Achieve." *Annual Review of Political Science* 19: S1–S23.

Dunning, Thad, Guy Grossman, Macartan Humphreys, Susan Hyde, Craig McIntosh, and Gareth Nellis. 2019. *Information, Accountability, and Cumulative Learning: Lessons from Metaketa I*. Cambridge: Cambridge University Press.

Dwyer, Patrick C., Alexander Maki, and Alexander J. Rothman. 2015. "Promoting Energy Conservation Behavior in Public Settings." *Journal of Environmental Psychology* 41: 30-34.

Eckel, Catherine, and Natalia Candelo. 2021. “How to Tame Lab-in-the-Field Experiments.” In James N. Druckman, and Donald P. Green, eds., *Advances in Experimental Political Science*. New York: Cambridge University Press.

Egami, Naoki, and Erin Hartman. 2020. “Elements of External Validity: Framework, Design, and Analysis.” Working Paper, Columbia University and University of California, Los Angeles.

Egami, Naoki, and Kosuke Imai. 2015. “Causal Interaction in High Dimension.” Working Paper, Princeton University.

Egger, Peter, and Marko Koethenbuerger. 2010. “Government Spending and Legislative Organization: Quasi-Experimental Evidence from Germany.” *American Economic Journal: Applied Economics* 2: 200-212.

Einstein, Katherine Levine and David M. Glick. 2017. “Does Race Affect Access to Government Services? An Experiment Exploring Street-Level Bureaucrats and Access to Public Housing.” *American Journal of Political Science* 61: 100-116.

Eldersveld, Samuel J. 1956. “Experimental Propaganda Techniques and Voting Behavior.” *American Political Science Review* 50: 154-165.

Elman, Colin, John Gerring, and James Mahoney, eds. 2020. *The Production of Knowledge: Enhancing Progress in Social Science*. Cambridge: Cambridge University Press.

Elman, Colin, Diana Kapiszewski, and Arthur Lupia. 2018. “Transparent Social Inquiry: Implications for Political Science.” *Annual Review of Political Science* 21: 29-47.

Enos, Ryan D., and Noam Gidron. 2016. “Intergroup Behavioral Strategies as Contextually Determined: Experimental Evidence from Israel.” *The Journal of Politics* 78: 851-867.

Erikson, Robert S, Michael B. Mackuen, and James A. Stimson. 2002. *The Macro Polity*. Cambridge: Cambridge University Press.

Erikson, Robert S., and Laura Stoker. 2011. "Caught in the Draft: The Effects of Vietnam Draft Lottery Status on Political Attitudes." *American Political Science Review* 105: 221-237.

Fanelli, Daniele. 2018. "Is Science Really Facing a Reproducibility Crisis, and Do We Need It To?" *Proceedings of the National Academy of Sciences* 115: 2628-2631.

Fanelli, Daniele, Rodrigo Costas, and John P. A. Ioannidis. 2017. "Meta-Assessment of Bias in Science." *Proceedings of the National Academy of Sciences* 114: 3714-3719.

Fazio, Russell H. 1995. "Attitudes as Object-Evaluation Associations: Determinants, Consequences, and Correlates of Attitude Accessibility." In Richard E. Petty, and Jon A. Krosnick (eds.), *Attitude Strength: Antecedents and Consequences*. Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

Fenno, Richard F. 1978. *Home Style: House Members in their Districts*. Boston: Little, Brown

Findley, Michael G., Nathan M. Jensen, Edmund J. Malesky, and Thomas B. Pepinsky. 2016. "Can Results-Free Review Reduce Publication Bias?: The Results and Implications of a Pilot Study." *Comparative Political Studies* 49: 1667–1703.

Findley, Michael G., Daniel L. Nielson, and J. C. Sharman. 2014. *Global Shell Games: Experiments in Transnational Relations, Crime, and Terrorism*. New York: Cambridge University Press.

Finkel, Eli. J., Paul W. Eastwick, and Harrty T. Reis. 2015. "Best Research Practices in Psychology: Illustrating Epistemological and Pragmatic Considerations with the Case of Relationship Science." *Journal of Personality and Social Psychology 108:* 275-297.

Fiorina, Morris P., and Charles R. Plott. 1978. “Committee Decisions under Majority Rule.” *American Political Science Review* 72: 575-598.

Fowler, Anthony. 2020. “Partisan Intoxication or Policy Voting?” *Quarterly Journal of Political Science* 15: 141-179.

Franco, Annie, Neil Malhotra, and Gabor Simonovits. 2014. “Publication Bias in the Social Sciences: Unlocking the File Drawer.” *Science* 345: 1502-1505.

Franco, Annie, Neil Malhotra, and Gabor Simonovits. 2015. “Underreporting in Political Science Survey Experiments: Comparing Questionnaires to Published Results.” *Political Analysis* 23: 306-312.

Franco, Annie, Neil Malhotra, and Gabor Simonovits. 2016. “Underreporting in Psychology Experiments: Evidence from a Study Registry.” *Social Psychological and Personality Science* 7: 8-12.

Franco, Annie, Neil Malhotra, Gabor Simonovits, and L. J. Zigerell. 2017. “Developing Standards for Post-Hoc Weighting in Population-Based Survey Experiments.” *Journal of Experimental Political Science* 4: 161-172.

Freese, Jeremy, and Devah Pager. 2004. “Who Deserves a Helping Hand?: Attitudes about Government Assistance for the Unemployed by Race, Incarceration Status, and Worker History.” Paper Presented at the Annual Meetings of the American Sociological Association, San Francisco, CA.

Freese, Jeremy, and David Peterson. 2017. “Replication in Social Science.” *Annual Review of Sociology* 43: 147-165.

Friedman, Daniel, and Shyam Sunder. 1994. *Experimental Economics: A Primer for Economists*. New York: Cambridge University Press.

Friedman, Lisa. 2019. “E.P.A. to Limit Science Used to Write Public Health Rules.” *The New York Times*, November 11. https://www.nytimes.com/2019/11/11/climate/epa-science-trump.html.

Friedman, Milton. 1953. *Essays in Positive Economics.* Chicago: University of Chicago Press.

Fryer, Jr., Roland G., and Steven D. Levitt. 2004. “The Causes and Consequences of Distinctively Black Names.” *The Quarterly Journal of Economics* 119: 767-805.

Gaddis, S. Michael. 2017. “How Black Are Lakisha and Jamal?: Racial Perceptions from Names Used in Correspondence Audit Studies.” *Sociological Science* 4: 469-489.

Gaddis, S. Michael, ed. 2018. *Audit Studies: Behind the Scenes with Theory, Method, and Nuance*. Cham, Switzerland: Springer International Publishing

Gaines, Brian J., James H. Kuklinski, and Paul J. Quirk. 2007. “The Logic of the Survey Experiment Reexamined.” *Political Analysis* 15: 1-20.

Gay, Claudine. 2012. “Moving to Opportunity: The Political Effects of a Housing Mobility Experiment.” *Urban Affairs Review* 48: 147-179.

Geddes, Barbara. 2003. *Paradigms and Sand Castles: Theory Building and Research Design in Comparative Politics*. Ann Arbor, MI: The University of Michigan Press.

Geering, John. 2001. *Social Science Methodology: A Critical Framework*. Cambridge: Cambridge University Press.

Gelman, Andrew. 2013. “Preregistration of Studies and Mock Reports.” *Political Analysis* 21: 40-41

Gelman, Andrew, and Eric Loken. 2014. “The Statistical Crisis in Science: Data-dependent Analysis–a ‘Garden of Forking Paths’–Explains Why Many Statistically Significant Comparisons Don’t Hold Up.” *American Scientist* 102: 460-465.

Gerber, Alan S. Kevin Arceneaux, Cheryl Boudreau, Conor Dowling, Sunshine Hillygus. 2015. "Reporting Balance Tables, Response Rates and Manipulation Checks in Experimental Research: A Reply from the Committee That Prepared the Reporting Guidelines." *Journal of Experimental Political Science* 2: 216-229.

Gerber, Alan. Kevin Arceneaux, Cheryl Boudreau, Conor Dowling, Sunshine Hillygus, Thomas Palfrey, Daniel R. Biggers and David J. Hendry. 2014. "Reporting Guidelines for Experimental Research: A Report from the Experimental Research Section Standards Committee." *Journal of Experimental Political Science* 1: 81-98.

Gerber, Alan S., and Donald P. Green. 2000. "The Effects of Canvassing, Telephone Calls, and Direct Mail on Voter Turnout." *American Political Science Review* 94: 653-663.

Gerber, Alan S., and Donald P. Green. 2008. "Field Experiments and Natural Experiments." In Janet M. Box-Steffensmeier, Henry E. Brady, and David Collier, eds., *The Oxford Handbook of Political Methodology*. Oxford: Oxford University Press.

Gerber, Alan S., and Donald P. Green. 2012. *Field Experiments: Design, Analysis, and Interpretation*. New York: W. W. Norton.

Gerber, Alan S., Donald P. Green, and Christopher W. Larimer. 2008. "Social Pressure and Voter Turnout: Evidence from a Large-Scale Field Experiment." *American Political Science Review* 102: 33-48.

Gerber, Alan S., Donald P. Green, and David Nickerson. 2000. "Testing for Publication Bias in Political Science." *Political Analysis* 9: 385-392.

Gerber, Alan S., and Neil Malhotra. 2008. "Publication Bias in Empirical Sociological Research: Do Arbitrary Significance Levels Distort Published Results?" *Sociological Methods & Research* 37: 3-30.

Gerber, Alan S., Neil Malhotra, Connor M. Dowling, and David Doherty. 2010. "Publication Bias in Two Political Behavior Literature." *American Political Research* 38(4): 591-613.

Gilligan, Michael J., Eric N. Mvukiyehe, and Cyrus Samii. 2013. "Reintegrating Rebels into Civilian Life: Quasi-Experimental Evidence from Burundi." *Journal of Conflict Resolution* 57: 598-626.

Gilligan, Michael J., Benjamin J. Pasquale, and Cyrus Samii. 2014. "Civil War and Social Cohesion: Lab-in-the-Field Evidence from Nepal." *American Journal of Political Science* 58: 604-619.

Glennerster, Rachel, and Kudzai Takavarasha. 2013. *Running Randomized Evaluations: A Practical Guide*. Princeton University Press.

Glynn, Adam N. 2021. "Advances in Experimental Mediation Analysis." In James N. Druckman, and Donald P. Green, eds., *Advances in Experimental Political Science*. New York: Cambridge University Press.

Gneezy, Uri, and Alex Imas. 2017. "Lab in the Field." In Abhijit Banerjee, and Esther Duflo, eds., *Handbook of Economic Field Experiments*, Amsterdam: Elsevier.

Goldberg, Matthew H., Sander van der Linden, Matthew Ballew, Seth A. Rosenthal, and Anthony Leiserowitz. 2020. "Convenient but Biased? The Reliability of Convenience Samples in Research about Attitudes Toward Climate Change." Retrieved from https://osf.io/2h7as/.

Goodman, Steven N., Daniele Fanelli, and John P. A. Ioannidis. 2016. "What Does Research Reproducibility Mean?" *Science Translational Medicine* 8: 341ps12341ps12.

Goodman, Steven, and Sander Greenland. 2007. “Why Most Published Research Findings Are False: Problems in the Analysis.” *PLoS Medicine* 4: e168.

Goroff, Daniel L., Neil A. Lewis, Jr., Anne M. Scheel, Laura Scherer, and Joshua A. Tucker. 2019. “The Inference Engine: A Grand Challenge to Address the Context Sensitivity Problem in Social Science Research.” Working Paper, Cornell University.

Gosnell, Harold F. 1926. “An Experiment in the Stimulation of Voting.” *American Political Science Review* 20: 869-874.

Green, Donald P., and Alan S. Gerber. 2010. “Introduction to Social Pressure and Voting: New Experimental Evidence.” *Political Behavior* 32: 331-336.

Green, Donald P., and Holger L. Kern. 2012. “Modeling Heterogeneous Treatment Effects in Survey Experiments with Bayesian Additive Regression Trees.” *Public Opinion Quarterly* 76: 491-511.

Green, Donald P., Mary C. McGrath, and Peter M. Aronow. 2013. “Field Experiments and the Study of Voter Turnout.” *Journal of Elections, Public Opinion and Parties* 23: 27-48.

Green, Donald P., and Andrej Tusicisny. 2012. “Statistical Analysis of Results from Laboratory Studies in Experimental Economics: A Critique of Current Practice.” Paper presented at the North American Economic Science Association (ESA) Conference, Tucson, AZ, US, November 16-17.

Green, Paul E., and V. Srinivasan. 1978. “Conjoint Analysis in Consumer Research: Issues and Outlook.” *Journal of Consumer Research* 5: 103-123.

Grimmer, Justin, Sean J. Westwood, and Solomon Messing. 2015. *The Impression of Influence: Legislator Communication, Representation, and Democratic Accountability*. Princeton: Princeton University Press.

Grose, Christian R. 2014. “Field Experimental Work on Political Institutions.” *Annual Review of Political Science* 17: 355-370.

Grose, Christian R. 2021. “Experiments, Political Elites, and Political Institutions.” In James N. Druckman, and Donald P. Green, eds., *Advances in Experimental Political Science*. New York: Cambridge University Press.

Grose, Christian R., and Abby K. Wood. 2020. “Randomized Experiments by Government Institutions and American Political Development. *Public Choice* 185: 401-413.

Groves, Robert M. 2011. “Three Eras of Survey Research.” *Public Opinion Quarterly* 75: 861-871.

Groves, Robert M. Floyd J. Fowler Jr., Mick P. Couper, James M. Lepkowski, Eleanor Singer, and Roger Tourangeau. 2009. *Survey Methodology*. Hoboken, New Jersey: John Wiley & Sons.

Guala, Francesco. 2005. *The Methodology of Experimental Economics*. New York: Cambridge University Press.

Guala, Francesco. 2009. “Methodological Issues in Experimental Design and Interpretation.” In Harold Kincaid, and Don Ross, eds., *The Oxford Handbook of Philosophy of Economics*. New York: Oxford University Press.

Guess, Andrew M. 2021. “Experiments Using Social Media Data.” In James N. Druckman, and Donald P. Green, eds., *Advances in Experimental Political Science*. New York: Cambridge University Press.

Guetzkow, Harold, and Joseph J. Valadez, eds. 1981. *Simulated International Processes*. Beverly Hills, CA: Sage.

Habyarimana, James, Macartan Humphreys, Daniel N. Posner, and Jeremy M. Weinstein. 2007. "Why Does Ethnic Diversity Undermine Public Goods Provision?" *American Political Science Review* 101: 709-725.

Hagger, M. S., N. L. D. Chatzisarantis, H. Alberts, C. O. Anggono, C. Batailler, A. R. Birt, R. Brand, M. J. Brandt, G. Brewer, S. Bruyneel, D. P. Calvillo, W. K. Campbell, P. R. Cannon, M. Carlucci, N. P. Carruth, T. Cheung, A. Crowell, D. T. D. De Ridder, S. Dewitte, M. Elson, J. R. Evans, B. A. Fay, B. M. Fennis, A. Finley, Z. Francis, E. Heise, H. Hoemann, M. Inzlicht, S. L. Koole, L. Koppel, F. Kroese, F. Lange, K. Lau, B. P. Lynch, C. Martijn, H. Merckelbach, N. V. Mills, A. Michirev, A. Miyake, A. E. Mosser, M. Muise, D. Muller, M. Muzi, D. Nalis, R. Nurwanti, H. Otgaar, M. C. Philipp, P. Primoceri, K. Rentzsch, L. Ringos, C. Schlinkert, B. J. Schmeichel, S. F. Schoch, M. Schrama, A. Schütz, A. Stamos, G. Tinghög, J. Ullrich, M. vanDellen, S. Wimbarti, W. Wolff, C. Yusainy, O. Zerhouni, and M. Zwienenberg. 2016. "A Multilab Preregistered Replication of the Ego-Depletion Effect." *Perspectives on Psychological Science* 11: 546-573.

Hainmueller, Jens, Dominik Hangartner, and Teppei Yamamoto. 2015. "Validating Vignette and Conjoint Survey Experiments against Real-World Behavior." *Proceedings of the National Academy of Sciences* 112: 2395-2400.

Hainmueller, Jens, and Daniel J. Hopkins. 2015 "The Hidden American Immigration Consensus: A Conjoint Analysis of Attitudes toward Immigrants." *American Journal of Political Science* 59: 529-548.

Hainmueller Jens. Daniel J. Hopkins and Teppei Yamamoto. 2014. “Causal Inference in Conjoint Analysis: Understanding Multidimensional Choices via Stated Preference Experiments.” *Political Analysis* 22: 1-30.

Hall, Andrew B., Connor Huff, and Shiro Kuriwaki. 2019. “Wealth, Slaveownership, and Fighting for the Confederacy: An Empirical Study of the American Civil War.” *American Political Science Review* 113:658-673.

Halpern-Manners, Andrew, and John Robert Warren. 2012. “Panel Conditioning in Longitudinal Studies: Evidence From Labor Force Items in the Current Population Survey.” *Demography* 49: 1499-1519.

Halpern-Manners, Andrew, John Robert Warren, and Florencia Torche. 2017. “Panel Conditioning in the General Social Survey.” *Sociological Methods & Research* 46: 103-124.

Han, Hahrie. 2016. “The Organizational Roots of Political Activism: Field Experiments on Creating a Relational Context.” *American Political Science Review* 110: 296-307.

Hankinson, Michael. 2018. “When Do Renters Behave Like Homeowners?: High Rent, Price Anxiety, and NIMBYism.” *American Political Science Review* 112: 473-493.

Hart, William, Dolores Albarracín, Alice H. Eagly, Inge Brechan, Matthew J. Lindberg, and Lisa Merill. 2009. “Feeling Validated Versus Being Correct.” *Psychological Bulletin* 135: 555-588.

Hartman, Erin. 2021. “Generalizing Experimental Results.” In James N. Druckman, and Donald P. Green, eds., *Advances in Experimental Political Science*. New York: Cambridge University Press.

Hauser David J., Phoebe C. Ellsworth, and Richard Gonzalez. 2018. “Are Manipulation Checks Necessary?” *Frontiers in Psychology* 9: 998.

Hauser, David J., and Norbert Schwarz. 2015. “It’s a Trap!: Instructional Manipulation Checks Prompt Systematic Thinking on ‘Tricky’ Tasks.” *SAGE Open*. https://doi.org/10.1177/2158244015584617

Healy, Andrew J., and Neil Malhotra. 2010. “Random Events, Economic Losses, and Retrospective Voting: Implications for Democratic Competence.” *International Quarterly Journal of Political Science* 5: 193-208.

Heck, Patrick R., Christopher F. Chabris, Duncan J. Watts, and Michelle N. Meyer. 2020. “Objecting to Experiments Even While Approving of the Policies or Treatments They Compare.” *Proceedings of the National Academy of Sciences* 117: 18948-18950.

Heckman, James J. 1998. “Detecting Discrimination.” *Journal of Economic Perspectives* 12: 101-116.

Hedges, Larry V. 2019. “The Statistics of Replication.” Working Paper, Northwestern University.

Hedges, Larry V., and Jacob Schauer. 2019. “More than One Replication Study is Needed for Unambiguous Tests of Replication.” Working Paper, Northwestern University.

Hemker, Johannes, and Anselm Rink. 2017. “Multiple Dimensions of Bureaucratic Discrimination: Evidence from German Welfare Offices.” *American Journal of Political Science* 61: 786-803.

Henrich, Joseph, Robert Boyd, Samuel Bowles, Colin Camerer, Ernst Fehr, Herbert Gintis, Richard McElreath, Michael Alvard, Abigail Barr, Jean Ensminger, Natalie Smith Henrich, Kim Hill, Francisco Gil-White, Michael Gurven, Frank W. Marlowe, John Q.

Patton, and David Tracer. 2005. "'Economic Man' in Cross-Cultural Perspective: Behavioral Experiments in 15 Small-Scale Societies." *Behavioral and Brain Sciences* 28: 795-815.

Henrich, Joseph, Steven J. Heine, and Ara Norenzayan. 2010. "The Weirdest People in the World?" *Behavioral and Brain Sciences* 33: 61-83.

Hermann, Charles F., and Margaret G. Hermann. 1967. "An Attempt to Simulate the Outbreak of World War I ." *American Political Science Review* 61: 400-416.

Hillygus, D. Sushine, Natalie Jackson, and McKenzie Young. 2014. "Professional Respondents in Non-probability Online Panels." In Mario Callegaro, Reg Baker, Jelke Bethlehem, Anja S. Göritz, Jon A. Krosnick, and Paul J. Lavrakas. *Online Panel Research: A Data Quality Perspective*. West Sussex, UK: John Wiley & Sons, Ltd.

Ho, Daniel, E., and Kosuke Imai. 2008. "Estimating Causal Effects of Ballot Order from a Randomized Natural Experiment: The California Alphabet Lottery, 1978-2002." *Public Opinion Quarterly* 72: 216-240.

Hoenig, John M., and Dennis M. Heisey. 2001. "The Abuse of Power." *The American Statistician* 55: 19-24.

Holbrook, Allyson L., and Jon A. Krosnick. 2010. "Measuring Voter Turnout By Using The Randomized Response Technique: Evidence Calling Into Question The Method's Validity." *Public Opinion Quarterly* 74: 328-343.

Holland, Paul W. 1986. "Statistics and Causal Inference." *Journal of the American Statistical Association* 81: 945-960.

Hopkins, Daniel J., and Jonathan Mummolo. 2017. "Assessing the Breadth of Framing Effects." *Quarterly Journal of Political Science* 12: 37-57.

Horowitz, Michael C., and Matthew S. Levendusky. 2011. “Drafting Support for War: Conscription and Mass Support for Warfare.” *The Journal of Politics* 73: 524-534.

Hoynes, Hilary, Leslie McGranahan and Diane Schanzenbach, 2015. “SNAP and Food Consumption”. In Judith Bartfeld, Craig Gundersen, Timothy Smeeding, and James P. Ziliak, eds., *SNAP Matters: How Food Stamps Affect Health and Well-Being*. Stanford: Stanford University Press.

Hoynes, Hilary, Diane Whitmore Schanzenbach and Douglas Almond, 2016. “Long-Run Impacts of Childhood Access to the Safety Net,” *American Economic Review* 106: 903-934.

Huber, John. 2013. “Is Theory Getting Lost in the ‘Identification Revolution’?” *The Political Economist* Summer: 1-3.

Huff, Connor, and Dustin Tingley. 2015. “Who Are These People? Evaluating the Demographic Characteristics and Political Preferences of MTurk Survey Respondents.” *Research & Politics* 2: 1-12.

Humphreys, Macartan, Raul Sanchez de la Sierra, Peter van der Windt. 2013. “Fishing, Commitment, and Communication: A Proposal for Comprehensive Nonbinding Research Registration.” *Political Analysis* 21: 1–20.

Humphreys, Macartan, and Jeremy M. Weinstein. 2009. “Field Experiments and the Political Economy of Development.” *Annual Review of Political Science* 12: 367-378.

Husnu, Senel, and Richard J. Crisp. 2010. “Elaboration Enhances the Imagined Contact Effect.” *Journal of Experimental Social Psychology* 46: 943-950.

Husnu, Senel, and Richard J. Crisp. 2011 “Enhancing the Imagined Contact Effect.” *The Journal of Social Psychology* 151: 113-116.

Hyde, Susan D. 2007. “The Observer Effect in International Politics: Evidence from a Natural Experiment.” *World Politics* 60: 37-63.

Hyde, Susan D. 2015. “Experiments in International Relations: Lab, Survey, and Field.” *Annual Review of Political Science* 18: 403-424

Hyde, Susan D., and Nikolay Marinov. 2014. “Information and Self-Enforcing Democracy: The Role of International Election Observation.” *International Organization* 68: 329-359.

Ichino, Nahomi, and Matthias Schündeln. 2012. “Deterring or Displacing Electoral Irregularities? Spillover Effects of Observers in a Randomized Field Experiment in Ghana.” *The Journal of Politics* 74: 292-307.

Imai, Kosuke, Luke Keele, Dustin Tingley and Teppei Yamamoto. 2011. “Unpacking the Black Box of Causality: Learning about Causal Mechanisms from Experimental and Observational Studies.” *American Political Science Review* 105: 765-789.

Imai, Kosuke, Dustin Tingley and Teppei Yamamoto. 2013. “Experimental Designs for Identifying Causal Mechanisms.” *Journal of the Royal Statistical Society* 176: 5-51.

Imai, Kosuke and Teppei Yamamoto. 2013. “Identification and Sensitivity Analysis for Multiple Causal Mechanisms: Revisiting Evidence from Framing Experiments.” *Political Analysis* 21: 141-171.

Imbens, Guido W., and Donald B. Rubin. 2015. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. New York: Cambridge University Press.

Ioannidis, John P.A. 2005. “Why Most Published Research Findings Are False.” *PLoS Medicine* 2: e124.

Ioannidis, John P.A. 2017. “Statistical Biases in Science Communication: What We Know About Them and How They Can Be Addressed.” In Kathleen Hall Jamieson, Dan M. Kahan,

and Dietram A. Scheufele, eds., *Oxford Handbook of the Science of Science Communication*, New York: Oxford University Press.

Iyengar, Shanto, and Kyu S. Hahn. 2009. "Red Media, Blue Media: Evidence of Ideological Selectivity in Media Use." *Journal of Communication* 59: 19-39.

Iyengar, Shanto, and Donald R. Kinder. 1987. *News That Matters: Television and American Opinion*. Chicago: The University of Chicago Press.

Iyengar, Shanto, Mark D. Peters, and Donald R. Kinder. 1982. "Experimental Demonstrations of the 'Not-So-Minimal' Consequences of Television News Programs." *American Political Science Review* 76: 848-858.

Iyengar, Shanto, Gaurav Sood, and Yphtach Lelkes. 2012. "Affect, Not Ideology: A Social Identity Perspective on Polarization." *Public Opinion Quarterly* 76: 405-431.

Jamieson, Thomas, and Nicholas Weller. 2020. "The Effects of Certain and Uncertain Incentives on Effort and Knowledge Accuracy." *Journal of Experimental Political Science* 7: 218-231.

Janz, Nicole, and Jeremy Freese. N.d. "Replicate Others as You Would Like to Be Replicated Yourself." *PS: Political Science and Politics*, Forthcoming.

Jardina, Ashley, and Spencer Piston. 2019. "Racial Prejudice, Racial Identity, and Attitudes in Political Decision Making." *Oxford Research Encyclopedia, Politics*. DOI: 10.1093/acrefore/9780190228637.013.966

Jayachandran, Seema, Joost de Laat, Eric F. Lambin, Charlotte Y. Stanton, Robin Audy, and Nancy E. Thomas. 2017. "Cash for Carbon: A Randomized Trial of Payments for Ecosystem Services to Reduce Deforestation." *Science* 357: 267-273.

Jenke, Libby, Kirk Bansak, Jens Hainmueller, and Dominik Hangartner. 2020. "Using Eye-Tracking to Understand Decision-Making in Conjoint Experiments." *Political Analysis* 10.1017/pan.2020.11/.

John, Peter. 2011. *Making Policy Work*. New York, NY: Routledge.

John, Peter. 2017. *Field Experiments in Political Science and Public Policy: Practical Lessons in Design and Delivery*. New York: Routledge.

Kahan, Dan M. 2017. "Misconceptions, Misinformation, and the Logic of Identity-Protective Cognition." Cultural Cognition Project Working Paper Series No. 164. Available at SSRN: https://ssrn.com/abstract=2973067 or http://dx.doi.org/10.2139/ssrn.2973067.

Kahneman, Daniel. 2002. "Daniel Kahneman: Biographical." NobelPrize.org. Nobel Media AB. https://www.nobelprize.org/prizes/economic-sciences/2002/kahneman/biographical/

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

Kalla, Joshua L., and David E. Broockman. 2016. "Campaign Contributions Facilitate Access to Congressional Officials: A Randomized Field Experiment." *American Journal of Political Science* 60: 545-558.

Kam, Cindy D., Jennifer R. Wilking, and Elizabeth J. Zechmeister. 2007. "Beyond the 'Narrow Data Base': Another Convenience Sample for Experimental Research." *Political Behavior* 29: 415-440.

Kane, John V., and Jason Barabas. 2019. "No Harm in Checking: Using Factual Manipulation Checks to Assess Attentiveness in Experiments." *American Journal of Political Science* 63: 234-249.

Kane, John V., Yamil R. Velez, and Jason Barabas. 2020. “Analyze the Attentive & Bypass Bias: Mock Vignette Checks in Survey Experiments.” Working Paper, New York University.

Karlan, Dean, and Jacob Appel. 2016. *Failing in the Field: What We Can Learn When Field Research Goes Wrong*. Princeton: Princeton University Press.

Kenny, David A., and Charles M. Judd. 2019. “The Unappreciated Heterogeneity of Effect Sizes: Implications for Power, Precision, Planning of Research, and Replication.” *Psychological Methods* 24: 578-589.

Kertzer, Joshua D., and Ryan Brutger. 2016. “Decomposing Audience Costs: Bringing the Audience Back into Audience Cost Theory.” *American Journal of Political Science* 60: 234-249.

Kim, Eunji. 2019. “Entertaining Beliefs in Economic Mobility.” Ph.D. Dissertation, University of Pennsylvania.

Kinder, Donald R., and Thomas R. Palfrey, eds. 1993. *Experimental Foundations of Political Science*. Ann Arbor: The University of Michigan Press.

King, Gary. 1991. “‘Truth’ Is Stranger than Prediction, More Questionable than Causal Inference.” *American Journal of Political Science* 35: 1047-1053.

King, Gary, Robert O. Keohane, and Sidney Verba. 1994. *Designing Social Inquiry: Scientific Inference in Qualitative Research*. Princeton: Princeton University Press.

King, Gary, and Melissa Sands. 2015. “How Human Subjects Research Rules Mislead You and Your University, and What to Do About It.” Working Paper, Harvard University.

Klar, Samara, and Thomas J. Leeper. 2019. “Identities and Intersectionality: A Case for Purposive Sampling in Survey Experimental Research.” In Paul J. Lavrakas, Michael W.

Traugott, Courtney Kennedy, Allyson L. Holbrook, Edith D. de Leeuw, and Brady T. West, eds., *Experimental Methods in Survey Research: Techniques that Combine Random Sampling with Random Assignment*. Hoboken, NJ: John Wiley & Sons, Inc.

Klein, Richard A., Kate A. Ratliff, Michelangelo Vianello, Reginald B. Adams Jr., Štepán Bahník, Michael J. Bernstein, Konrad Bocian, Mark J. Brandt, Beach Brooks, Claudia Chloe Brumbaugh, Zeynep Cemalcilar, Jesse Chandler, Winnee Cheong, William E. Davis, Thierry Devos, Matthew Eisner, Natalia Frankowska, David Furrow, Elisa Maria Galliani, Fred Hasselman,Joshua A. Hicks, James F. Hovermale, S. Jane Hunt, Jeffrey R. Huntsinger, Hans IJzerman, Melissa-Sue John, Jennifer A. Joy-Gaba, Heather Barry Kappes, Lacy E. Krueger, Jaime Kurtz, Carmel A. Levitan, Robyn K. Mallett, Wendy L. Morris, Anthony J. Nelson, Jason A. Nier, Grant Packard, Ronaldo Pilati, Abraham M. Rutchick, Kathleen Schmidt, Jeanine L. Skorinko, Robert Smith, Troy G. Steiner, Justin Storbeck, Lyn M. Van Swol, Donna Thompson, A. E. van 't Veer, Leigh Ann Vaughn, Marek Vranka, Aaron L. Wichman, Julie A. Woodzicka, and Brian A. Nosek. 2014. "Investigating Variation in Replicability: A ''Many Labs'' Replication Project." *Social Psychology* 45:142-152.

Klein, Stanley B. 2014. "What Can Recent Replication Failures Tell Us about the Theoretical Commitments of Psychology?" *Theory & Psychology* 24: 326-338.

Knorr-Cetina, Karin. 1999. *Epistemic Cultures: How the Sciences Make Knowledge*. Cambridge, MA: Harvard University Press.

Kraft-Todd, Gordon T., and David G. Rand. N.d. "Practice What You Preach: Credibility-Enhancing Displays and the Growth of Open Science." *Organizational Behavior and Human Decision Processes*, Forthcoming.

Kramer, Adam D.I., Jamie E. Guillory, and Jeffrey T. Hancock. 2014. “Experimental Evidence of Massive-Scale Emotional Contagion through Social Networks.” *Proceedings of the National Academy of Sciences* 111: 8788-8790.

Krosnick, Jon A., and Stanley Presser. 2010. “Question and Questionnaire Design.” In Peter V. Marsden, and James D. Wright. *Handbook of Survey Research*. Bingley: Emerald.

Kruglanski, Arie W. 1975. “The Human Subject in the Psychology Experiment: Fact and Artifact.” In Leonard Berkowitz, ed., *Advances in Experimental Social Psychology*. New York: Academic Press.

Krupnikov, Yanna. 2011. “When Does Negativity Demobilize? Tracing the Conditional Effect of Negative Campaigning on Voter Turnout.” *American Journal of Political Science* 55: 797-813.

Krupnikov, Yanna, and Adam Seth Levine. 2014. “Cross-Sample Comparisons and External Validity.” *Journal of Experimental Political Science* 1: 59-80.

Krupnikov, Yanna, Hannah Nam, Hillary Style. 2021. “Convenience Samples in Political Science Experiments.” In James N. Druckman, and Donald P. Green, eds., *Advances in Experimental Political Science*. New York: Cambridge University Press.

Krupnikov, Yanna, Hillary Style, and Michael Yontz. N.d. “Does Measurement Affect the Gender Gap in Political Partisanship?” *Public Opinion Quarterly*, Forthcoming.

Kuhn, Thomas S. 1962. *The Structure of Scientific Revolutions*. Chicago, IL: University of Chicago Press.

Künzel, Sören R., Jasjeet S. Sekhon, Peter J. Bickel, and Bin Yu. 2019. "Metalearners for Estimating Heterogeneous Treatment Effects Using Machine Learning." *Proceedings of the National Academy of Sciences* 116: 4156-4165.

Laitin, David D. 2013. "Fisheries Management." *Political Analysis* 21: 42-47.

Lakatos, Imre. 1970. "Falsification and the Methodology of Scientific Research Programmes." In Imre Lakatos, and Alan Musgrave, eds., *Criticism and the Growth of Knowledge*. Cambridge: Cambridge University Press.

Landy, Justin F., Miaolei Liam Jia, Isabel L. Ding, Domenico Viganola, Warren Tierney, Anna Dreber, Magnus Johannesson, Thomas Pfeiffer, Charles R. Ebersole, Quentin F. Gronau, Alexander Ly, Don van den Bergh, Maarten Marsman, Koen Derks, Eric-Jan Wagenmakers, Andrew Proctor, Daniel M. Bartels, Christopher W. Bauman, William J. Brady, Felix Cheung, Andrei Cimpian, Simone Dohle, M. Brent Donnellan, Adam Hahn, Michael P. Hal, William Jiménez-Leal, David J. Johnson, Richard E. Lucas, Benoît Monin, Andres Montealegre, Elizabeth Mullen, Jun Pang, Jennifer Ray, Diego A. Reinero, Jesse Reynolds, Walter Sowden, Daniel Storage, Runkun Su, Christina M. Tworek, Jay J. Van Bavel, Daniel Walco, Julian Wills, Xiaobing Xu, Kai Chi Yam, Xiaoyu Yang, William A. Cunningham, Martin Schweinsberg, Molly Urwitz, The Crowdsourcing Hypothesis Tests Collaboration, Eric L. Uhlmann. 2020. "Crowdsourcing Hypothesis Tests: Making Transparent How Design Choices Shape Research Results." *Psychological Bulletin* 146: 451-479.

Lau, Richard R., and David P. Redlawsk. 2001. "Advantages and Disadvantages of Cognitive Heuristics in Political Decision Making." *American Journal of Political Science* 45: 951-971.

Lau, Richard R., and David P. Redlawsk. 2006. *How Voters Decide: Information Processing in Election Campaigns*. Cambridge, MA: Cambridge University Press.

Lau, Richard R., Lee Sigelman, Caroline Heldman, and Paul Babbitt. 1999. "The Effects of Negative Political Advertisements: A Meta-Analytic Assessment." *American Political Science Review* 93: 851-875.

Lau, Richard R., Lee Sigelman, and Ivy Brown Rovner. 2007. "The Effects of Negative Political Campaigns: A Meta-Analytic Reassessment." *The Journal of Politics* 69: 1176-1209.

Lavrakas, Paul J., Michael W. Traugott, Courtney Kennedy, Allyson L. Holbrook, Edith D. de Leeuw, and Brady T. West, eds. 2019. *Experimental Methods in Survey Research: Techniques that Combine Random Sampling with Random Assignment*. Hoboken, NJ: John Wiley & Sons, Inc.

Lazarsfeld, Paul, Bernard Berelson, and Hazel Gaudet. 1948. *The People's Choice*. New York: Columbia University Press.

Leeper, Thomas J. 2011. "The Role of Protocol in the Design and Reporting of Experiments." *Newsletter of the American Political Science Association Experimental Section* 2: 6-10.

Leeper, Thomas J., Sara B. Hobolt, and James Tilley. 2020. "Measuring Subgroup Preferences in Conjoint Experiments." *Political Analysis* 28: 207-221.

Lelkes, Yphtach and Sean Westwood. 2017. "The Limits of Partisan Prejudice." *The Journal of Politics* 79: 485-501.

León, Federico R., Rebecka Lundgren, Irit Sinai, Ragini Sinha, and Victoria Jennings. 2014. "Increasing Literate and Illiterate Women's Met Need for Contraception via Empowerment: A Quasi-Experiment in Rural India." *Reproductive Health* 11: 74.

Lepper, Mark R., Jennifer Henderlong, and Isabelle Gingras. 1999. "Understanding the Effects

of Extrinsic Rewards on Intrinsic Motivation—Uses and Abuses of Meta-Analysis: Comment on Deci, Koestner, and Ryan." *Psychological Bulletin* 125*:* 669-676.

Levay, Kevin E., Jeremy Freese, and James N. Druckman. 2016. "The Demographic and Political Composition of Mechanical Turk Samples." *SAGE Open* 6: 1-17.

Levendusky, Matthew. 2010. "Clearer Cues, More Consistent Voters: A Benefit of Elite Polarization." *Political Behavior* 32: 111-31.

Levendusky, Matthew. 2013. *How Partisan Media Polarize America*. Chicago: University of Chicago Press.

Levine, Adam Seth. 2021. "How to Form Organizational Partnerships to Run Experiments" In James N. Druckman, and Donald P. Green, eds., *Advances in Experimental Political Science*. New York: Cambridge University Press.

Liebman, Jeffrey, Jens Ludwig, Lawrence Katz, Raj Chetty, Jeffrey Kling, Lisa Sanbonmatsu, Emma Adam, Greg J. Duncan, Ronald C. Kessler, Lisa A. Gennetian, Stacy Tessler Lindau, Thomas W. McDade, Joshua C. Pinkston, Robert C. Whitaker, Paul Hirschfield, and Nathaniel Hendren. 2020. "Evaluating the Impact of Moving to Opportunity in the United States." The Abdul Latif Jameel Poverty Action Lab (J-PAL). Accessed at: https://www.povertyactionlab.org/evaluation/evaluating-impact-moving-opportunity-united-states.

Lin, Winston, and Donald P. Green. 2016. "Standard Operating Procedures: A Safety Net for Pre-Analysis Plans." *PS: Political Science & Politics* 49): 495–500.

Liyanarachchi, Gregory A. 2007. "Feasibility of Using Student Subjects in Accounting Experiments: A Review." *Pacific Accounting Review* 19: 47-67.

Lowell, A. Lawrence. 1910. “The Physiology of Politics.” *American Political Science Review* 4: 1-15.

Luce, R. Duncan, and John W. Tukey. 1964. “Simultaneous Conjoint Measurement: A New Type of Fundamental Measurement.” *Journal of Mathematical Psychology* 1: 1-27.

Luo, Yu, and Jiaying Zhao. 2019. “Motivated Attention in Climate Change Perception and Action.” *Frontiers in Psychology* 10: 1541.

Lupia, Arthur, and Colin Elman. 2014. “Openness in Political Science: Data Access and Research Transparency.” *PS: Political Science and Politics* 47: 19-42.

Lupia, Arthur, and Mathew D. McCubbins. 1998. *The Democratic Dilemma: Can Citizens Learn What They Need to Know?* New York: Cambridge University Press.

Lupton, Danielle L. 2019. “The External Validity of College Student Subject Pools in Experimental Research: A Cross-Sample Comparison of Treatment Effect Heterogeneity.” *Political Analysis* 27: 90-97.

Lynott, Dermot, Katherine S. Corker, Jessica Wortman, Louise Connell, M. Brent Donnellan, Richard E. Lucas, and Kerry O’Brien.2014. “Replication of “Experiencing Physical Warmth Promotes Interpersonal Warmth.” *Social Psychology* 45: 216-222.

MacDonald, Paul. 2003. “Useful Fiction or Miracle Maker: The Competing Epistemological Foundations of Rational Choice Theory.” *American Political Science Review* 97: 551-565.

MacInnis, Bo, Jon A. Krosnick, Annabell S. Ho, and Mu-Jung Cho. 2018. “The Accuracy of Measurements with Probability and Nonprobability Survey Samples: Replication and Extension.” *Public Opinion Quarterly* 82: 707-744.

Mahoney, James. 2010. “After KKV: The New Methodology of Qualitative Research.” *World Politics* 62: 120-147.

Mahoney, Robert, and Daniel Druckman. 1975. “Simulation, Experimentation, and Context.” *Simulation & Games* 6: 235-270.

Malhotra, Neil. 2021. “The Scientific Credibility of Experiments.” In James N. Druckman, and Donald P. Green, eds., *Advances in Experimental Political Science*. New York: Cambridge University Press.

Malhotra Neil, and Alexander G. Kuo. 2008. “Attributing Blame: The Public’s Response to Hurricane Katrina.” *The Journal of Politics* 70: 120-135.

Malhotra, Neil, and Elizabeth Popp. 2012. “Bridging Partisan Divisions over Antiterrorism Policies: The Role of Threat Perceptions.” *Political Research Quarterly* 65: 34-47.

Maniaci, Michael R. and Ronald D. Rogge. 2014. “Caring about Carelessness: Participant Inattention and its Effects on Research.” *Journal of Research in Personality* 48: 61-83

Mansbridge, Jane, 2003. “Rethinking Representation.” *American Political Science Review* 97: 515-528.

Matanock, Aila M. 2021. “Experiments in Post-Conflict Contexts.” In James N. Druckman, and Donald P. Green, eds., *Advances in Experimental Political Science*. New York: Cambridge University Press.

McDermott, Rose. 2002. “Experimental Methodology in Political Science.” *Political Analysis* 10: 325-342.

McDermott, Rose. 2011. “Internal and External Validity.” In James N. Druckman, Donald P. Green, James H. Kuklinski, and Arthur Lupia, eds., *Cambridge Handbook of Experimental Political Science*, New York: Cambridge University Press.

McFadden, Daniel. 2017. “Stated Preference Methods and Their Applicability to Environmental Use and Non-use Valuations.” In Daniel McFadden, and Kenneth Train, eds., *Contingent Valuation of Environmental Goods: A Comprehensive Critique*. Cheltenham, UK Edward Elgar Publishing Limited.

McKelvey, Richard D., and Peter C. Ordeshook. 1990. “A Decade of Experimental Research on Spatial Models of Elections and Committees.” In James M. Enelow, and Melvin J. Hinich, eds., *Advances in the Spatial Theory of Voting*. Cambridge: Cambridge University Press.

McShane, Blakeley B., Jennifer L. Tackett, Ulf Böckenholt, and Andrew Gelman. 2019. “Large-Scale Replication Projects in Contemporary Psychological Research.” *The American Statistician* 73: 99-105.

Meredith, Marc. 2009. “Persistence in Political Participation.” *Quarterly Journal of Political Science* 4: 187–209.

Messick, Samuel. 1998. “Test Validity: A Matter of Consequence.” *Social Indicators Research* 45: 35-44.

Mettler, Suzanne, and Joe Soss. 2004. “The Consequences of Public Policy for Democratic Citizenship: Bridging Policy Studies and Mass Politics.” *Perspectives on Politics* 2: 55–73.

Meyer, Michelle N. Patrick R. Heck, Geoffrey S. Holtzman, Stephen M. Anderson, William Cai, Duncan J. Watts, and Christopher F. Chabris. 2019a. “Objecting to Experiments That Compare Two Unobjectionable Policies or Treatments.” *Proceedings of the National Academy of Sciences* 116: 10723-10728.

Meyer, Michelle N. Patrick R. Heck, Geoffrey S. Holtzman, Stephen M. Anderson, William Cai, Duncan J. Watts, and Christopher F. Chabris. 2019b. "Reply to Mislavsky et al.: Sometimes People Really Are Averse to Experiments." *Proceedings of the National Academy of Sciences* 116: 23885-23886.

Miguel, E., C. Camerer, K. Casey, J. Cohen, K. M. Esterling, A. Gerber, R. Glennerster, D. P. Green, M. Humphreys, G. Imbens, D. Laitin, T. Madon, L. Nelson, B. A. Nosek, M. Petersen, R. Sedlmayr, J. P. Simmons, U. Simonsohn, M. Van der Laan. 2014. "Promoting Transparency in Social Science Research." *Science* 343: 30-31.

Miles, Eleanor, and Richard J. Crisp. 2014. "A Meta-Analytic Test of the Imagined Contact Hypothesis." *Group Processes & Intergroup Relations* 17: 3–26.

Miller, David, ed. 1985. *Popper Selections*. Princeton: Princeton University Press.

Miller, David. 1994. *Critical Rationalism: A Restatement and Defense*. Chicago: Open Court.

Miller Joanne M., and Jon A. Krosnick. 2000. "News Media Impact on the Ingredients of Presidential Evaluations: Politically Knowledgeable Citizens Are Guided by a Trusted Source." *American Journal of Political Science* 44: 295-309.

Miratrix, Luke W., Jasjeet S. Sekhon, Alexander G. Theodoridis, and Luis F. Campos. 2018. "Worth Weighting? How to Think About and Use Weights in Survey Experiments." *Political Analysis* 26: 275-291.

Mislavsky, Robert, Berkeley J. Dietvorst, and Uri Simonsohn. 2019. "The Minimum Mean Paradox: A Mechanical Explanation for Apparent Experiment Aversion." *Proceedings of the National Academy of Sciences* 116: 23883-23884.

Mislavsky, Robert , Berkeley Dietvorst, Uri Simonsohn. 2020. “Critical Condition: People Don’t Dislike a Corporate Experiment More Than They Dislike Its Worst Condition.” *Marketing Science* 39:1092-1104.

Monroe, Kristen Renwick. 2018 “The Rush to Transparency: DA-RT and the Potential Dangers for Qualitative Research.” *Perspectives on Politics* 16: 141–148.

Montgomery, David B, and Dick R. Wittink. 1979. “Predictive Validity of Trade-Off Analysis for Alternative Segmentation Schemes.” *Proceedings of the American Marketing Association Educators' Conference*. Research Collection Lee Kong Chian School Of Business.

Mook, Douglas G. 1983. “In Defense of External Invalidity.” *American Psychologist* 38: 379-387.

Morgenstern, Oskar. 1976. “The Collaboration Between Oskar Morgenstern and John von Neumann on the Theory of Games.” *Journal of Economic Literature* 14: 805-816.

Morton, Rebecca B. and Kenneth C. Williams. 2008. “Experimentation in Political Science.” In Janet M. Box-Steffensmeier, Henry E. Brady, and David Collier, eds., *The Oxford Handbook of Political Methodology*. Oxford: Oxford University Press.

Morton, Rebecca B. and Kenneth C. Williams. 2010. *Experimental Political Science and the Study of Causality: From Nature to the Lab.* New York: Cambridge University Press.

Motyl, Matt, Alexander P. Demos, Timothy S. Carsel, Brittany E. Hanson, Zachary J. Melton, Allison B. Mueller, J. P. Prims, Jiaqing Sun, Anthony N. Washburn, Kendal M. Wong, Caitlyn Yantis, Linda J. Skitka. 2017. “The State of Social and Personality Science: Rotten to the Core, Not So Bad, Getting Better, or Getting Worse?” *Journal of Personality and Social Psychology* 113: 34-58.

Mullinix, Kevin J., Thomas J. Leeper, James N. Druckman, and Jeremy Freese. 2015. "The Generalizability of Survey Experiments." *Journal of Experimental Political Science* 2: 109-38.

Mummolo, Jonathan, and Erik Peterson. 2019. "Demand Effects in Survey Experiments: An Empirical Assessment." *American Political Science Review* 113: 517-529.

Munger, Kevin, Jonathan Nagler, Joshua Tucker and Mario Luca. 2019. "Everyone On Mechanical Turk is Above a Threshold of Digital Literacy: Sampling Strategies for Studying Digital Media Effects." Working Paper, Pennsylvania State University. Available at http://kmunger.github.io/pdfs/clickbait_mturk.pdf.

Mutz, Diana C. 2005. "Social Trust and E-Commerce: Experimental Evidence for the Effects of Social Trust on Individuals' Economic Behavior." *Public Opinion Quarterly* 69: 393-416.

Mutz, Diana C. 2007. "Effects of 'In-Your-Face' Television Discourse on Perceptions of a Legitimate Opposition." *American Political Science Review* 101: 621-635.

Mutz, Diana C. 2011. *Population-Based Survey Experiments*. Princeton University Press.

Mutz, Diana C. 2021. "Improving Experimental Treatments in Political Science." In James N. Druckman, and Donald P. Green, eds., *Advances in Experimental Political Science*. New York: Cambridge University Press.

Mutz, Diana C., and Robin Pemantle. 2015. "Standards for Experimental Research: Encouraging a Better Understanding of Experimental Methods." *Journal of Experimental Political Science* 2: 192-215.

Nathan, Noah L., and Ariel White. 2021. “Experiments on and with Street-Level Bureaucrats.” In James N. Druckman, and Donald P. Green, eds., *Advances in Experimental Political Science*. New York: Cambridge University Press.

National Academies of Sciences, Engineering, and Medicine. 2009. *On Being a Scientist: A Guide to Responsible Conduct in Research: Third Edition*. Washington, DC: The National Academies Press.

National Academies of Sciences, Engineering, and Medicine. 2019. *Reproducibility and Replicability in Science*. Washington DC: The National Academies Press.

Nature. 2014. “Journals Unite for Reproducibility.” *Nature* 515: 7.

Neblo, Michael A., Kevin M. Esterling, and David M. J. Lazer. 2018. *Politics with the People: Building a Directly Representative Democracy*. Cambridge: Cambridge University Press.

Nelsen, Matthew D. 2019. “Cultivating Youth Engagement: Race & the Behavioral Effects of Critical Pedagogy.” *Political Behavior* https://doi.org/10.1007/s11109-019-09573-6.

Nelson, Thomas E., Rosalee A. Clawson, and Zoe M. Oxley. 1997. “Media Framing of a Civil Liberties Conflict and Its Effect on Tolerance.” *American Political Science Review* 91: 567-583.

Neumark, David. 2012. “Detecting Discrimination in Audit and Correspondence Studies. *The Journal of Human Resources* 47: 1128-1157.

Neumark, David. 2018. “Experimental Research on Labor Market Discrimination.” *Journal of Economic Literature* 56: 799-866.

Neyman, Jerzy. 1923 [1990]. “On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9.” *Statistical Science* 5: 465-472. Trans. Dorota M. Dabrowska and Terence P. Speed.

Nickerson, David W. 2008. “Is Voting Contagious?: Evidence from Two Field Experiments.” *American Political Science Review* 102: 49-57.

Nosek, B. A., G. Alter, G. C. Banks, D. Borsboom, S. D. Bowman, S. J. Breckler, S. Buck, C.D. Chambers, G. Chin, G. Christensen, M. Contestabile, A. Dafoe, E. Eich, J. Freese, R. Glennerster, D. Goroff, D. P. Green, B. Hesse, M. Humphreys, J. Ishiyama, D. Karlan, A. Kraut, A. Lupia, P. Mabry, T. Madon, N. Malhotra, E. Mayo-Wilson, M. McNutt, E. Miguel, E. Levy Paluck, U. Simonsohn, C. Soderberg, B. A. Spellman, J. Turitto, G. VandenBos, S.Vazire, E. J. Wagenmakers, R. Wilson, and T. Yarkoni 2015. “Promoting an Open Research Culture.” *Science* 348: 1422-1425.

Nosek, Brian A., Charles R. Ebersole, Alexander C. DeHaven, and David T. Mellor. 2018. “The Preregistration Revolution.” *Proceedings of the National Academy of Sciences* 115: 2600-2606.

Nosek, Brian A., and Timothy M. Errington. 2019. “What is Replication?” Working Paper, Center for Open Science.

Nyhan, Brendan. 2015. “Increasing the Credibility of Political Science Research: A Proposal for Journal Reforms.” *PS: Political Science & Politics* 48: 78-83.

Oliver, Jack E. 2004. *The Incomplete Guide to the Art of Discovery*. Ithaca, NY: Internet-First University Press.

Olken, Benjamin A. 2015. “Promises and Perils of Pre-Analysis Plans.” *Journal of Economic Perspectives* 29: 61-80.

Open Science Collaboration (OSC). 2015. “Estimating the Reproducibility of Psychological Science.” *Science* 349: aac4716-1-aac4716-8.

Oppenheimer, Daniel M., Tom Meyvis, Nicolas Davidenko. 2009. “Instructional Manipulation Checks: Detecting Satisficing to Increase Statistical Power.” *Journal of Experimental Social Psychology* 45: 867-872.

Oreskes, Naomi. 2019. *Why Trust Science?* Princeton: Princeton University Press.

Oschatz, Corinna, and Caroline Marker. 2020. “Long-Term Persuasive Effects in Narrative Communication Research: A Meta-Analysis.” *Journal of Communication* 70: 473-496.

Packer, Milton. 2017. “Are Meta-Analyses a Form of Medical Fake News?: Thoughts About How They Should Contribute to Medical Science and Practice.” *Circulation* 136: 2097-2099.

Pager, Devah. 2003. “The Mark of a Criminal Record.” *American Journal of Sociology* 108: 937-975.

Pager, Devah, and Hana Shepherd. 2008. “The Sociology of Discrimination: Racial Discrimination in Employment, Housing, Credit, and Consumer Markets.” *Annual Review of Sociology* 34: 181-209.

Paluck, Elizabeth Levy, Seth Green, and Donald P. Green. 2019. “The Contact Hypothesis Re-Evaluated.” *Behavioural Public Policy* 3: 129-158.

Paolacci, Gabriele, Jesse Chandler, and Panagiotis G. Ipeirotis. 2010. “Running Experiments on Amazon Mechanical Turk.” *Judgment and Decision Making* 5: 411-419.

Patil, Prasad, Roger D. Peng, and Jeffrey T. Leek. 2016. “What Should Researchers Expect When They Replicate Studies? A Statistical View of Replicability in Psychological Science.” *Perspectives on Psychological Science* 11: 539-544.

Pearl, Judea. 2000. *Causality: Models, Reasoning, and Inference*. Cambridge: Cambridge University Press.

Pearl, Judea, with Dana Mackenzie. 2018. *The Book of Why: The New Science of Cause and Effect*. New York: Basic Books.

Persson, Mikael, and Maria Solevid. 2014. "Measuring Political Participation—Testing Social Desirability Bias in a Web-Survey Experiment." *International Journal of Public Opinion Research* 26: 98-112.

Perugini, Marco, Marcello Gallucci, and Giulio Costantini. 2018. "A Practical Primer to Power Analysis for Simple Experimental Designs." *International Review of Social Psychology* 31: 20, 1-23.

Peterson, Erik. 2017. "The Role of the Information Environment in Partisan Voting." *The Journal of Politics* 79: 1191-1204.

Peterson, Erik, Sharad Goel, and Shanto Iyengar. 2017. "Echo Chambers and Partisan Polarization: Evidence from the 2016 Presidential Campaign." Unpublished Paper, Stanford University.

Peyton, Kyle, Gregory A. Huber, and Alexander Coppock. 2020. "The Generalizability of Online Experiments Conducted During The COVID-19 Pandemic." Working Paper, Yale University.

Pfaff, Steven, Charles Crabtree Holger L. Kern, and John B. Holbein. N.d. "Does Religious Bias Shape Access to Public Services?: A Large-Scale Audit Experiment among Street-level Bureaucrats." *Pubic Administration Review*. Forthcoming.

Piazza, Thomas. 2010. "Fundamental of Applied Sampling." In Peter V. Marsden, and James D. Wright. *Handbook of Survey Research*. Bingley: Emerald.

Pirlott, Angela G., and David P. MacKinnon. 2016. "Design Approaches to Experimental Mediation." *Journal of Experimental Social Psychology* 66: 29-38.

Plott, Charles R. 1991. “Will Economics Become an Experimental Science?” *Southern Economic Journal* 57: 901-919.

Plott, Charles R., and Kirill Pogorelskiy. 2017. “Call Market Experiments: Efficiency and Price Discovery through Multiple Calls and Emergent Newton Adjustments.” *American Economic Journal: Microeconomics* 9: 1-41.

Popper, Karl R. 1959. *The Logic of Scientific Discovery*. Abingdon-on-Thames, UK: Routledge.

Popper, Karl R. 1962. *Conjectures and Refutations: The Growth of Scientific Knowledge*. New York: Basic Books.

Popper, Karl R. (edited by Paul A. Schilpp). 1974. *The Philosophy of Karl Popper*. Chicago: Open Court.

Prior, Markus, Gaurav Sood, and Kabir Khanna. 2015. “You Cannot Be Serious: The Impact of Accuracy Incentives on Partisan Bias in Reports of Economic Perceptions.” *Quarterly Journal of Political Science* 10: 489-518.

Quadlin, Natasha. 2018. “The Mark of a Woman’s Record: Gender and Academic Performance in Hiring.” *American Sociological Review* 83: 331-360.

Quattrone, George A., and Amos Tversky. 1988. “Contrasting Rational and Psychological Analyses of Political Choice.” *American Political Science Review* 82: 719-736.

Quillian, Lincoln, Anthony Heath, Devah Pager, Arnfinn H. Midtboen, Fenella Fleischmann, Ole Hexel. 2019. “Do Some Countries Discriminate More than Others?: Evidence from 97 Field Experiments of Racial Discrimination in Hiring.” *Sociological Science* 6: 467-496.

Quillian, Lincoln, John J. Lee, Mariana Oliver. 2020. “Evidence from Field Experiments in Hiring Shows Substantial Additional Racial Discrimination after the Callback.” *Social Forces* 99: 732-759.

Quillian, Lincoln, Devah Pager, Ole Hexel, and Arnfinn H. Midtbøen. 2017. "Meta-analysis of Field Experiments Shows No Change in Racial Discrimination in Hiring Over Time." *Proceedings of the National Academy of Sciences* 114: 10870-10875.

Ratkovic, Marc. 2021. "Subgroup Analysis: Pitfalls, Promise, and Honesty." In James N. Druckman, and Donald P. Green, eds., *Advances in Experimental Political Science*. New York: Cambridge University Press.

Ratkovic, Marc, and Dustin Tingley. 2017. "Sparse Estimation and Uncertainty with Application to Subgroup Analysis." *Political Analysis* 25: 1-40.

Redish, A. David, Erich Kummerfeld, Rebecca Lea Morris, and Alan C. Love. 2018. "Reproducibility Failures are Essential to Scientific Inquiry." *Proceedings of the National Academy of Sciences* 115: 5042-5046.

Robison, Joshua, Randy T. Stevenson, James N. Druckman, Simon Jackman, Jonathan N. Katz, and Lynn Vavreck. 2018. "An Audit of Political Behavior Research." *SAGE Open* 8: 1-14.

Rogowski, Ronald. 2016. "The Rise of Experimentation in Political Science." In Robert A Scott and Stephan M. Kosslyn, eds., *Emerging Trends in the Social and Behavioral Sciences*. Hoboken, New Jersey: John Wiley & Sons, Inc.

Rosenfeld, Bryn, Kosuke Imai, and Jacob N. Shapiro. 2016. "An Empirical Validation Study of Popular Survey Methodologies for Sensitive Questions" *American Journal of Political Science* 60: 783-802.

Rosenthal, Robert. 1979. "The File Drawer Problem and Tolerance for Null Results." *Psychological Bulletin* 86: 638-641.

Roth, Alvin E. 1995. “Introduction to Experimental Economics.” In John H. Kagel, and Alvin E. Roth, eds., *The Handbook of Experimental Economics*. Princeton: Princeton University Press.

Rubin, Donald B. 1974. “Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies.” *Journal of Educational Psychology* 66: 688-701.

Ruggeri, Kai, Sonia Alí, Mari Louise Berge, Giulia Bertoldo, Ludvig D. Bjørndal, Anna Cortijos-Bernabeu, Clair Davison, Emir Demić, Celia Esteban-Serna, Maja Friedemann, Shannon P. Gibson, Hannes Jarke, Ralitsa Karakasheva, Peggah R. Khorrami, Jakob Kveder, Thomas Lind Andersen, Ingvild S. Lofthus, Lucy McGill, Ana E. Nieto, Jacobo Pérez, Sahana K. Quail, Charlotte Rutherford, Felice L. Tavera, Nastja Tomat, Chiara Van Reyn, Bojana Većkalov, Keying Wang, Aleksandra Yosifova, Francesca Papa, Enrico Rubaltelli, Sander van der Linden, and Tomas Folke. 2020. “Replicating Patterns of Prospect Theory for Decision Under Risk.” *Nature Human Behavior* 4: 622-633.

Schneider, Sandra L., ed. 2013. *Experimental Design in the Behavioral and Social Sciences*. London: Sage Publications.

Schwarz, Susanne, and Alexander Coppock. 2020. “What Have We Learned About Gender From Candidate Choice Experiments? A Meta-analysis of 67 Factorial Survey Experiments.” *The Journal of Politics,* Forthcoming.

Sears, David O. 1986. “College Sophomores in the Laboratory: Influence of a Narrow Data Base on Social Psychology’s View of Human Nature.” *Journal of Personality and Social Psychology* 51: 515-530.

Seawright, Jason. 2016. *Multi-Method Social Science: Combining Qualitative and Quantitative Tools*. Cambridge University Press.

Sekhon, Jasjeet S., and Roćio Titiunik. 2012. "When Natural Experiments Are Neither Natural Nor Experiments." *American Political Science Review* 106: 35-57.

Settle, Jaime E. 2018. *Frenemies: How Social Media Polarizes America*. Cambridge: Cambridge University Press.

Shadish, William R., and Thomas D. Cook. 2009. "The Renaissance of Field Experimentation in Evaluating Interventions." *Annual Review of Psychology* 60: 607-629.

Shadish, William, R, Thomas D. Cook, and Donald T. Campbell. 2002. *Experimental and Quasi-Experimental Designs for Generalized Causal Inferences*. Boston: Houghton Mifflin.

Shafranek, Richard M. 2019 "Political Considerations in Nonpolitical Decisions: A Conjoint Analysis of Roommate Choice." *Political Behavior*. https://doi.org/10.1007/s11109-019-09554-9.

Sherif, Muzafer, and Carolyn W. Sherif. 1953. *Groups in Harmony and Tension: An Integration of Studies on Intergroup Relations*. New York: Harper.

Simmons, Joseph P., Leif D. Nelson, and Uri Simonsohn. 2011. "False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant." *Psychological Science* 22: 1359-1366.

Simon, Herbert A. 1963. "Problems of Methodology Discussion." *American Economic Review Proceedings* 53: 229-231.

Simon, Herbert A. 1979. "Rational Decision Making in Business Organizations." *The American Economic Review* 69: 493-513.

Simonsohn, Uri. 2015. "Small Telescopes: Detectability and the Evaluation of Replication Results." *Psychological Science* 26: 559-569.

Sinclair, Betsy. 2012. *The Social Citizen: Peer Networks and Political Behavior*. University of Chicago Press.

Slothuus, Rune. 2016. "Assessing the Influence of Political Parties on Public Opinion: The Challenge from Pretreatment Effects." *Political Communication* 33: 302-327.

Slothuus, Rune, and Martin Bisgaard. 2020. "How Political Parties Shape Public Opinion in the Real World." *American Journal of Political Science*, https://doi.org/10.1111/ajps.12550.

Smaldino, Paul. 2019. "Better Methods Can't Make Up for Mediocre Theory." *Nature* 575: 9.

Smaldino Paul E., and Richard McElreath. 2016. "The Natural Selection of Bad Science." *Royal Society Open Science* 3: 160384.

Smith, Rogers M. 1993. "Beyond Tocqueville, Myrdal, and Hartz: The Multiple Traditions of America." *American Political Science Review* 87: 549-566.

Smith, Rogers M. 2020. "What Good Can Political Science Do?: From Pluralism to Partnerships." *Perspectives on Politics* 18: 10-26.

Smith, Vernon L. 1976. "Experimental Economics: Induced Value Theory." *American Economic Review* 66: 274-279.

Smith, Vernon L. 1982. "Microeconomic Systems as an Experimental Science." *The American Economic Review* 72: 923-955.

Sniderman, Paul M. 1995. "Evaluation Standards for a Slow-Moving Science." *Political Science and Politics* 28: 464-467.

Sniderman, Paul M. 2018. "Some Advances in the Design of Survey Experiments." *Annual Review of Political Science* 21: 259-275.

Sniderman, Paul M., Richard A. Brody, and Philip E. Tetlock. 1991. *Reasoning and Choice: Explorations in Political Psychology*. Cambridge: Cambridge University Press.

Sniderman, Paul M., and Douglas B. Grob. 1996. "Innovations in Experimental Design in Attitude Surveys." *Annual Review of Sociology* 22: 377-399.

Sniderman, Paul M., and Sean M. Theriault. 2004. "The Structure of Political Argument and the Logic of Issue Framing." In Willem E. Saris and Paul M. Sniderman, eds., *Studies in Public Opinion*. Princeton: Princeton University Press.

Sokal, Alan, and Jean Bricmont. 1997. *Intellectual Impostures: Postmodern Philosophers' Abuse of Science.* London: Profile Books.

Sollors, Werner, Caldwell Titcomb, and Thomas A. Underwood, eds. 1993. *Blacks at Harvard: A Documentary History of African-American Experience at Harvard and Radcliffe*. New York: New York University Press.

Sondheimer, Rachel Milstein. 2011. "Analyzing the Downstream Effects of Randomized Experiments." In James N. Druckman, Donald P. Green, James H. Kuklinski, and Arthur Lupia, eds., *Cambridge Handbook of Experimental Political Science*, New York: Cambridge University Press.

Stark, Tobias H., Henning Silber, Jon A. Krosnick, Annelies G. Blom, Midori Aoyagi, Ana Belchior, Michael Bosnjak, Sanne Lund Clement, Melvin John, Guðbjörg Andrea Jónsdóttir, Karen Lawson, Peter Lynn, Johan Martinsson, Ditte Shamshiri-Petersen, Endre Tvinnereim, Ruoh-rong Yu. 2018. "Generalization of Classic Question Order Effects Across Cultures." *Sociological Methods & Research*. doi:10.1177/0049124117747304.

Stewart, Neil, Christoph Ungemach, Adam J.L. Harris, Daniel M. Bartels, Ben R. Newell, Gabriele Paolacci, and Jesse Chandler. 2015. "The Average Laboratory Samples a Population of 7,300 Amazon Mechanical Turk Workers." *Judgment & Decision Making* 10: 479-491.

Stroebe, Wolfgang, and Fritz Strack. 2014. "The Alleged Crisis and the Illusion of Exact Replication." *Perspectives on Psychological Science* 9: 59-71.

Swedberg, Richard. 2020. "Exploratory Research." In Colin Elman, John Geering, and James Mahoney, eds., *The Production of Knowledge: Enhancing Progress in Social Science*. Cambridge: Cambridge University Press.

Taber, Charles S., and Milton Lodge. 2006. "Motivated Skepticism in the Evaluation of Political Beliefs." *American Journal of Political Science* 50: 755-769.

Tappin, Ben M. 2020. "Estimating the Between-Issue Variation in Party Elite Cue Effects." Working paper. Massachusetts Institute of Technology. https://doi.org/10.31234/osf.io/p48zb.

Thelen, Kathleen, and James Mahoney. 2015. "Comparative-historical Analysis in Contemporary Political Science." In James Mahoney, and Kathleen Thelen, eds., *Advances in Comparative-Historical Analysis*. Cambridge: Cambridge University Press.

Tipton, Elizabeth, Jessaca Spybrook, Katie Fitzgerald, Qian Wang, and Caryn Davidson. 2019. "The Convenience of Large Urban School Districts: A Study of Recruitment Practices in 37 Randomized Trials." Working Paper, Northwestern University.

Titiunik, Roćio. 2016. "Drawing Your Senator from a Jar: Term Length and Legislative Behavior." *Political Science Research and Methods* 4: 293-316.

Titiunik, Roćio. 2021. “Natural Experiments.” In James N. Druckman, and Donald P. Green, eds., *Advances in Experimental Political Science*. New York: Cambridge University Press.

Tomz, Michael. 2007. “Domestic Audience Costs in International Relations: An Experimental Approach.” *International Organization* 61: 821-840.

Tourangeau, Roger, and Ting Yan. 2007. “Sensitive Questions in Surveys.” *Psychological Bulletin* 133: 859-883.

Twenge, Jean M., Sara Konrath, Joshua D. Foster, W. Keith Campbell, and Brad J. Bushman. 2008. “Egos Inflating Over Time: A Cross-Temporal Meta-Analysis of the Narcissistic Personality Inventory.” *Journal of Personality* 76: 875-902.

Valenzuela, Ali, and Tyler Reny. 2021. “Evolution of Experiments on Racial Priming.” In James N. Druckman, and Donald P. Green, eds., *Advances in Experimental Political Science*. New York: Cambridge University Press.

Van Bavel, Jay J., Peter Mende-Siedlecki, William J. Brady, and Diego A. Reinero. 2016a. “Contextual Sensitivity in Scientific Reproducibility.” *Proceedings of the National Academy of Sciences* 113: 6454-6459.

Van Bavel, Jay J., Peter Mende-Siedlecki, William J. Brady, and Diego A. Reinero. 2016b. “Reply to Inbar: Contextual Sensitivity Helps Explain the Reproducibility Gap Between Social and Cognitive psychology.” *Proceedings of the National Academy of Sciences* 113: E4935-E4936.

VanderWeele. Tyler J. 2015. *Explanation in Causal Inference: Methods for Mediation and Interaction*. Oxford: Oxford University Press.

Vannette, David L., and Jon A. Krosnick, eds. 2018. *The Palgrave Handbook of Survey Research*. Cham, Switzerland: Palgrave Macmillan.

Vavreck, Lynn, and Douglas Rivers. 2008. "The 2006 Cooperative Congressional Election Study." *Journal of Elections, Public Opinion and Parties* 18: 355-366.

Wagenmakers, E.J. , T. Beek, L. Dijkhoff , Q. F. Gronau, A. Acosta, R. B. Adams, Jr., D. N. Albohn, E. S. Allard, S. D. Benning, E.-M. Blouin-Hudon, L. C. Bulnes, T. L. Caldwell, R. J. Calin-Jageman, C. A. Capaldi, N. S. Carfagno, K. T. Chasten, A. Cleeremans, L. Connell, J. M. DeCicco, K. Dijkstra, A. H. Fischer, F. Foroni, U. Hess, K. J. Holmes, J. L. H. Jones, O. Klein, C. Koch, S. Korb, P. Lewinski, J. D. Liao, S. Lund, J. Lupianez, D. Lynott, C. N. Nance, S. Oosterwijk, A. A. Ozdog˘ru, A. P. Pacheco-Unguetti, B. Pearson, C. Powis, S. Riding, T.-A. Roberts, R. I. Rumiati, M. Senden, N. B. Shea-Shumsky, K. Sobocko, J. A. Soto, T. G. Steiner, J. M. Talarico, Z. M. van Allen, M. Vandekerckhove, B. Wainwright, J. F. Wayand, R. Zeelenberg, E. E. Zetzer, and R. A. Zwaan. 2016. "Registered Replication Report: Strack, Martin, & Stepper (1988)." *Perspectives on Psychological Science* 11: 917-928.

Wantchekon, Leonard. 2003. "Clientelism and Voting Behavior." *World Politics* 55: 399-422.

Westreich, Daniel, Jessie K. Edwards, Catherine R. Lesko, Stephen R. Cole, and Elizabeth A. Stuart. 2019. "Target Validity and the Hierarchy of Study Designs." *American Journal of Epidemiology* 188: 438-443.

Westwood, Sean J., Erik Peterson, Yphtach Lelkes. 2019. "Are there Still Limits on Partisan Prejudice?." *Public Opinion Quarterly* 83: 584-597.

White, Ariel R., Noah L. Nathan, and Julie K. Faller. 2015. “What Do I Need to Vote? Bureaucratic Discretion and Discrimination by Local Election Officials.” *American Political Science Review* 109: 129-142.

Williamson, Vanessa. 2016. “On the Ethics of Crowdsourced Research.” *PS: Political Science & Politics* 49: 77–81.

Willis, Allison W. 2015. “Using Administrative Data to Examine Health Disparities and Outcomes in Neurological Diseases of the Elderly.” *Current Neurology and Neuroscience Reports* 15: 75.

Wong, Vivian C., and Peter M. Steiner. 2018. “Replication Designs for Causal Inference.” EdPolicyWorks Working Paper Series No. 62, University of Virginia.

Wood, Abby K., and Christian R. Grose. N.d. “Campaign Finance Transparency Affects Legislators' Election Outcomes and Behavior.” *American Journal of Political Science* Forthcoming.

Wright, James D., and Peter V. Marsden. 2010. “Survey Research and Social Science: History, Current Practice, and Future Prospects.” In Peter V. Marsden, and James D. Wright. *Handbook of Survey Research*. Bingley: Emerald.

Yeager, David S., Jon A. Krosnick, Penny S. Visser, Allyson L. Holbrook, Alex M. Tahk. 2019. “Moderation of Classic Social Psychological Effects by Demographics in the U.S. Adult Population: New Opportunities for Theoretical Advancement.” *Journal of Personality and Social Psychology* 117: e84-e99.

Zigerell, L.J. 2017. “Reducing Political Bias in Political Science Estimates.” *PS: Political Science & Politics* 50: 179-183.

Zigerell, L.J. 2018. “Black and White Discrimination in the United States: Evidence from an Archive of Survey Experiment Studies.” *Research & Politics* 5: 205316801775386.