A New Era of Experimental Political Science*

James N. Druckman
(druckman@northwestern.edu)
Northwestern University

Donald P. Green
(dpg2110@columbia.edu)
Columbia University

Abstract
Experimental political science has transformed in the last decade. The use of experiments has dramatically increased throughout the discipline, and technological and sociological changes have altered how political scientists use experiments. We chart the transformation of experiments and discuss new challenges that experimentalists face. We then outline how the contributions to this volume will help scholars and practitioners conduct high-quality experiments.

*We thank Nicolette Alayon, Robin Bayes, Jeremy Levy, Jacob Rothschild, and Andrew Thompson for research assistance. We thank Lynn Vavreck for excellent advice.

Experimental political science has changed. In two short decades, it evolved from an emergent method to an accepted method to a primary method. We are now entering a new era of experimental political science – what can be called experimental political science 2.0. We do not use the term "era" lightly. The new era reflects, in part, the expanded use of experiments throughout the discipline. But, more fundamentally, it reflects a radical shift in how social scientists design, analyze, and interpret experiments.

For most of social science history, the challenges for experimentalists concerned obtaining data beyond student subject pools and what to do with null results that typically landed in the "file drawer." This is no longer true. Data are plentiful thanks to internet panels, crowdsourcing platforms, social media, and electronic access to elites; partnerships between researchers and non-academic entities have also become a prolific source of experimental data. Computing advances have made the implementation and analysis of large-scale studies routine. Moreover, scholars now regularly discuss how to address issues of publication bias, replication, and data-sharing so as to ensure the production of credible experimental research.

The challenge now is to ensure that experimentalists design sound studies and implement them in ways that illuminate cause and effect. They must do so while also respecting ethical boundaries, interpreting results in a transparent manner, and sharing data and research materials to ensure others can build on what has been learned. Political science experimentalists, moreover, can capitalize on the widespread acceptance of the method, novel data sources, and evolving epistemological orientations. Making the most of these opportunities requires carefully choosing an appropriate design for a given research question, developing theoretically informative treatments and valid outcome measures, choosing a suitable setting, engaging in

sound analyses, cautiously generalizing, and addressing enduring debates. The goal of this volume is to shed light on best practices.

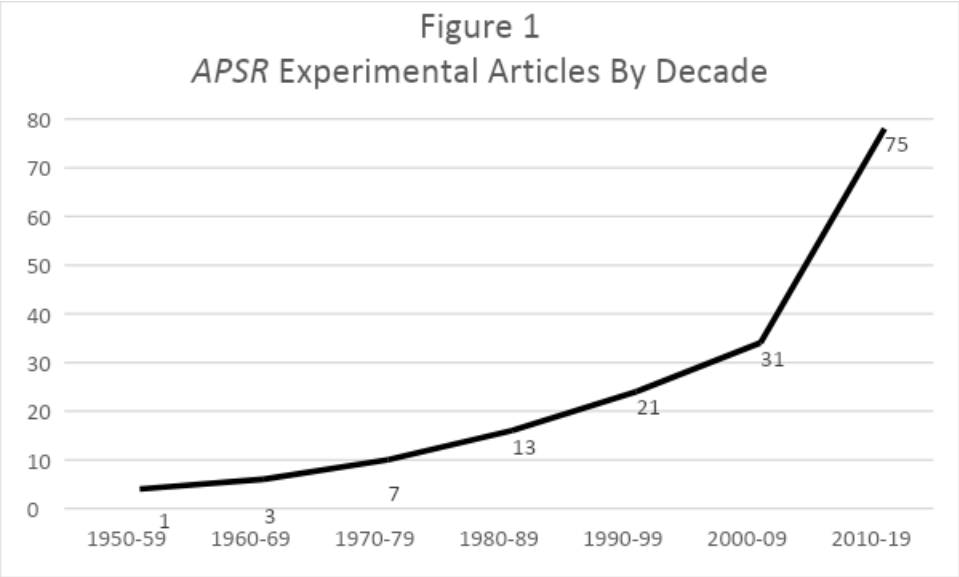
In what follows, we first describe the evolution of experiments in political science, focusing on quantitative trends, substantive reach, and institutional progression. This discussion documents a transformation in how political scientists think about and conduct experiments. We then turn to a discussion of recent developments in the social sciences, involving technological change and open science, an era we characterize as experimental political science 2.0. This new era is defined by the application of new designs; the introduction of novel data sources, measurement approaches, and statistical methods; the use of experiments in more areas; and discipline-wide discussions about the robustness, generalizability, and ethics of experiments in political science. This volume explores these new opportunities while also highlighting the concomitant challenges. The goal is to help scholars and practitioners conduct high-quality experiments that make important contributions to knowledge.

The Evolution of Experiments in Political Science

One way to document the evolution of experiments in political science is by counting the number of articles in general political science journals. We do that by focusing on the discipline's flagship journal, the *American Political Science Review* (*APSR*). We identified all published articles from the launch of the journal in 1906 through any article posted online in May 2019.[1] The first experiment appeared in 1956, 50 years after the journal started. In Figure 1, we plot the number of articles by decade, starting in the 1950s. To be clear, this is not a

[1] In so doing, we extend the timeline from our prior work (Druckman et al. 2006; also see Rogowski 2015). So as to accommodate how political scientists from varying perspectives define "experiment," we counted an experiment as a study involving random assignment to conditions or entailing an economic game that applies induced value theory. That said, we assert that "experiment" should only be used when the study employs random assignment (contrary to usage in many economic game studies) (see, e.g., Green and Tusicisny 2012).

cumulative count of articles but rather the specific number by decade. For example, from 2000 through 2009, 31 articles in the *APSR* used an experimental approach; this number jumped to 75 in the most recent decade. The figure supports the claim that experiments moved from being a marginalized method to an accepted method to a central method.



Has the recent surge in experimental articles spanned subfields in political science? In 2006, Druckman et al. (627) observed, "To date, the range of application remains narrow, with most experiments pertaining to questions in the subfields of political psychology, electoral politics, and legislative politics. An important question is the extent to which experiments or experiment-inspired research designs can benefit other subfields." The last decade has answered that question decisively: experiments have become common throughout the discipline. For example, in international relations there now exists a sizeable experimental literature on "audience costs," which refers to a process whereby governments publicly threaten to use force to induce a change in opposing countries' actions. The public nature of such threat makes it credible since the opponent recognizes a failure to use force would lead to domestic backlash (e.g., at the voting booth). Experiments show that, indeed, citizens have a distaste for empty

threats (e.g., Tomz 2007; although see Kertzer and Brutger 2016). The emergence of experimental research has also been apparent in other international relations domains, such as election monitoring, which has seen dramatic growth in the number and sophistication of randomized evaluations (Hyde and Marinov 2014; Ichino and Schündeln 2012; Buzin et al. 2016).

This momentum is especially noteworthy in comparative politics; since 2010, 45% 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). Some of these articles fall at the intersection of comparative politics and international relations, as in Beath et al.'s (2013) study of a massive aid program designed to empower Afghan women within the context of a civil war against the Taliban. Others span comparative politics and political psychology, as in Scacco and Warren's (2018) study of attempts to reduce prejudice between Muslims and Christians in Nigeria. Arguably the largest literature focuses on governance and accountability (see Dunning et al. 2019), typified by studies, such as Grossman and Michelitch (2018), that provide voters with job performance scorecards for randomly selected public officials over a series of election cycles. A final example of the reach of experiments concerns studies of whether and how public officials respond to queries from their constituents. In 2011, Butler and Brookman published their correspondence study of state legislators in 44 states. They sent email requests for information about voting registration, varying whether the email came from an ostensibly African-American or White constituent who was a Democrat, Republican, or did not mention a party. The binary outcome measure is whether the sender receives a reply from the state legislator's office. This study, which was patterned after correspondence experiments on job market discrimination (Pager 2003; Bertrand and Mullainathan 2004), spawned a literature that,

by 2017, included more than 50 audit experiments on the responsiveness of public officials (Costa 2017). It is also part of a growing experimental focus on elites – public officials or political leaders – as subjects (e.g., Grose 2014).

It is clear that political scientists think about and apply experiments in a very different way than a decade ago: they think of experimentation as a primary methodology and apply it in novel domains. These trends both reflect and spurred various institutional innovations. Here we point to three. First, in 2001, Time-sharing Experiments for the Social Sciences (TESS) was established with support from the National Science Foundation. TESS capitalizes on economies 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 400 experiments. Many of them are published in disciplinary flagship journals as well as *Science* and the *Proceedings of the National Academy of Science.* TESS also makes raw data from all experiments publicly available, regardless of whether the results are published.

The genesis of TESS in 2001 followed on the heels of what could be called a revolution in political science field experiments in 2000. In that year, a field experiment on voter mobilization was published in the *APSR* (Gerber and Green 2000). This publication was notable since it was the 47th experimental article in the journal but only the third field experiment, and the first field experiment in nearly 20 years.[2] This paper sparked burgeoning literatures on voter mobilization (e.g., Nickerson 2008) and vote choice (Wantchekon 2003); more generally, it

[2] We do not count Gosnell (1926) since he did not seem to employ random assignment.

ushered in the use of field experiments in other subfields (e.g. Findley et al. 2014; Hyde and Marinov 2014).[3]

The discipline established two other notable institutions about a decade later. In 2009, Evidence in Governance and Politics (EGAP) formed as a network for those engaged in field experiments on governance, politics, and institutions. EGAP played an important role in developing and advocating methodological practices such as pre-registration of experiments and professional standards concerning the public disclosure of results. As it grew in membership and capacity, it also expanded its worldwide outreach efforts to include instruction on experimental methods across the Global South. 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). These institutional innovations too were tracked 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 predecessor to this book, the *Cambridge Handbook of Experimental Political Science* (Druckman et al. 2011).[4]

These trends make clear that experiments now occupy a central place in political science. For reasons to which we turn next, the way researchers design, analyze, and present experiments is rapidly changing, leading to new challenges and opportunities.

Technological Change and Open Science

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

[4] Examples of other institutional developments include the launching of subject pools in more than a dozen political science departments (Druckman et al. 2018, 624) and a Routledge book series focused on experimental political science.

The initial rise of experiments 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., Kramer et al. 2014).

Consider four dynamics. First, as intimated above, data are now much cheaper and easier to obtain than ever before, thanks to the internet and the emergence of crowdsourcing platforms and commercial internet survey panels. These data are then easier to share due to the growing use of public data repositories, such as Dataverse and Github. The abundance of public data allowed, for example, Coppock et al. (2019) to use 27 studies to show that individual attributes such as age, gender, race, and ideology do not consistently condition how individuals process political messages: the effects of many messages do not vary across subgroups, implying that we can generalize about the impact of isolated experiments to large segments of the population.

Second, social media offer researchers access to behavioral data and the opportunity to intervene experimentally (e.g., Kramer et al. 2016), sometimes with literally millions of participants. Bond et al. (2012) conducted an experiment by delivering political mobilization messages to 61 million Facebook users, testing whether an "I Voted" widget that announces one's election participation to others increased turnout among Facebook users and their friends (see also Jones et al. 2017).

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. For instance, Kim (2018) used a truck equipped with mobile television monitors,

tablet computers, and chairs to conduct a lab-in-the-field study in three counties in rural Pennsylvania. The experiment shows that exposure to entertainment television with "rags-to-riches" narratives increases individuals' belief in the American Dream, particularly for Republicans (also see Busby 2018).

Fourth, advances in computing allow researchers to analyze high-dimensional data, which is to say data with large numbers of predictors or measurements. Computational requirements are especially demanding for algorithms that look for network effects (e.g., Grimmer et al. 2017). The same may be said for the rapidly growing list of techniques designed to automate the detection of treatment effect heterogeneity among subgroups in field experiments (Imai and Strauss 2011; Imai and Ratkovic 2013) and survey experiments (Green and Kern 2012). In the latter case, the authors revisit a large experimental literature based on General Social Surveys that have for decades asked national samples of Americans about their preferences regarding government spending. In the domain of social spending, question wording is varied randomly, and some respondents are asked about spending on "aid to the poor" while others are asked about spending on "welfare." These surveys consistently show "aid to the poor" to be much more popular than "welfare," but the question is what sorts of respondents are especially susceptible to this effect? Rather than manually search for treatment-by-covariate interactions with education, party, ideology, and a slew of other background attributes, the authors use machine learning methods to conduct an automated search that not only detects significant interactions but also cross-validates the results using respondents who were randomly excluded from the initial round of exploration.

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 counter-intuitive 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). In one notable example, Franco et al. (2014) show that of 221 Time-sharing Experiments in the Social Sciences surveys, strong results are 40 percentage points more likely to be published than null results and 50 percentage points more likely to be written up. This is clear evidence of a publication bias at the writing and submission stage (also see Franco et al. 2016).

One response to publication bias has been a call for more replications: emulating the extant study's procedures but with new data ("repeatability," as described in Freese and Peterson 2017). Massive replication efforts have had mixed results, with the most widely discussed being the Open Science Collaboration's (2015) effort where more than 250 scholars attempted to replicate 100 experiments in three highly ranked psychology journals from 2008. They reported that "39% of effects were subjectively rated to have replicated the original result," (Open Science Collaboration 2015, 943). This finding has led some to sound alarm bells of a replication crisis (Baker 2016); however, the extent of this crisis continues to be debated (e.g., van Bavel et al. 2016; Fanelli 2018), as other replication attempts, including those in political science, have had more success (e.g., Mullinix et al. 2015; Camerer et al. 2016; Coppock 2019).

These replication attempts are possible in part because of a push for scholars to make their procedures, stimuli, surveys, and data publicly available. In political science, most general and experimental oriented journals require data access upon publication (Lupia and Elman 2014). Growing public access to data is of enormous value to instructors and meta-analysts but also facilitates novel research. An example is Zigerell's (2018) re-analysis of 17 studies on racial

discrimination (e.g., attitudes towards White or Black political candidates or job applicants). He reports 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).

The opportunities that come from data sharing, replication debates, and related discussions have invigorated a call for "open science." Nosek et al. (2015) identify standards of transparency and openness, involving citation standards; transparency of data, material, and analyses; pre-registration of studies and analysis plans; and encouragement of replication studies. Interestingly, this move towards transparency has also generated some questions about respondent privacy as well as concerns about how respondents themselves react upon learning of data openness (Connors et al. 2019). In sum, fundamental technological *and* sociological changes have transformed the social sciences. The result, which coincided with the emergence of experiments as a primary method in political science, is what we call experimental political science 2.0.

Experimental Political Science 2.0

Experimental political science 2.0 is characterized by (1) the introduction of previously underutilized designs, (2) the explosion of new data sources, (3) the use of new measurement techniques, (4) advancements in statistical methods, (5) increased discussion about robustness and generalizability, and (6) applications to novel areas of study. To get some sense of these trends, we analyzed the content of all experimental articles in the *APSR* that made up Figure 1.[5] In reporting the results, we first distinguish three time periods: all articles prior to 2000 constitute the lead-up to the experimental era; 2001-2009 make up the first generation of widespread use;

[5] We thank Robin Bayes and Andrew Thompson for conducting the content analysis.

and 2010-present is what we call experimental political science 2.0. These cut-offs roughly coincide with the aforementioned institutional developments (e.g., TESS, EGAP, the American Political Science Association's Experimental Research section). Our interest is in the use and emergence of new approaches, and the statistics we present are the percentages of experimental articles in each era that used a given approach.

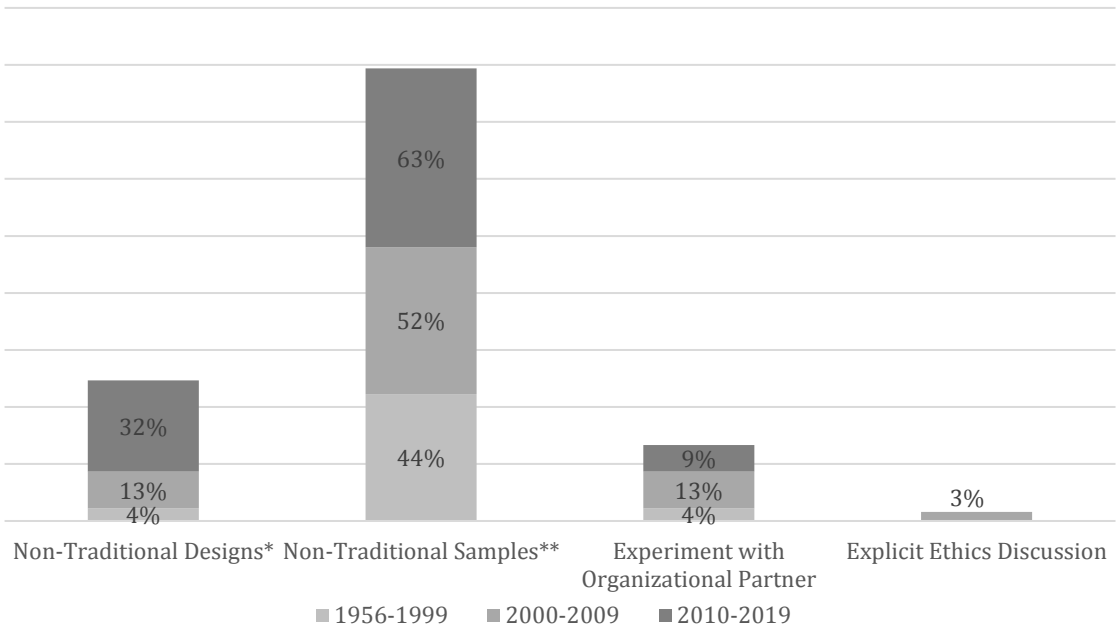
We start with what we might call "non-traditional designs" insofar as they are designs that received little application in early experiments in political science. We discuss them in more detail below, but they include conjoint surveys, audits, field experiments with surveys, lab-in-the-field studies, and natural experiments. In Figure 2, we report the percentage with which each of these designs were used out of all *APSR* experiments published in a time period. For instance, before 2000, of the 45 experiments published, 4% used one of the aforementioned designs. This number jumped to 13% in the second period and 32% in the most recent period – a clear trend towards increased application.

We see a similar upward trend when we look at the proportion of studies that use what we might call "non-traditional subject pools," including data from non-student convenience samples (e.g., crowdsourcing platforms), social media, or elites (e.g., legislators). The use of these subject pools jumped by 11 percentage points in the current era relative to the one that preceded it (52% to 63%).

Another change that came about largely with the rise of field experiments after 2000 was collaboration with organizational partners (e.g., nonprofits). The graph shows that such collaborations increased starting in 2000 but remain fairly minimal, perhaps due to a lack of guidance on how to develop such partnerships (a topic we take up in this volume). Another important issue that undoubtedly will be addressed more frequently in the future is discussion of

ethics. We identified only one experimental article in the *APSR* that included an explicit discussion of ethics in the main text of the paper (Paluck and Green 2009); growing recognition of ethical dilemmas in social science research (e.g., Teele 2014) will undoubtedly generate increased interest among both authors and audiences for further discussion of ethical issues.

Figure 2
Expeirmental Trends



*conjoint, audit, field + survey, lab-in-the-field, natural
**non-student convenience, social media, elite

In addition to these trends of design and data, the field continues to evolve when it comes to measurement and statistical methods. As in much of the social sciences, political scientists have embraced new measurement techniques and sources, such as administrative records, social media behaviors, physiological measures, and relatively unobtrusive measures of psychological processes. As for statistical methods, recent decades have seen growing sophistication in the use of techniques for detecting heterogeneous treatment effects (e.g., Ratkovic and Tingley 2017; Grimmer et al. 2017), spillovers between units (Aronow 2012; Bowers et al. 2016), and causal

mechanisms (e.g., Imai and Yamamoto 2013; Acharya et al. 2016). These methods feature in just over 30% of the articles appearing during the earliest time period and have become much more commonplace since 2000 (roughly 50% of experiments). A distinct trend worth noting concerns the use of visuals – nearly all experimental articles used visuals in the last decade, up from just more than half in the preceding period.

Another feature of experimental political science 2.0 echoes the aforementioned open science movement's concern with robustness and generalizability. This approach involves sustained discussion about reporting standards: one of the first actions of the American Political Science Association's Experimental Research section was to form a reporting committee (e.g., Gerber et al. 2014, 2015; Mutz and Pemantle 2015). At roughly the same time, the data access and research transparency (DA-RT) movement in political science gained prominence. It arose from growing concerns about scholars' failure to replicate a considerable number of empirical claims being made in top journals – often as a result of researchers' inability or unwillingness to provide information about how they drew conclusions from their data or to make the data available to others (Lupia and Elman 2014). The initiatives require authors, including experimentalists, to provide data access, production transparency (e.g., procedures about how the data were collected), and analytic transparency (American Political Science Association 2012, 9-10). There also are ongoing debates in the discipline about the need to register experiments so that researchers who later summarize literatures can see the extent to which research results went unreported. Another debate concerns pre-registration of analysis plans, an initiative designed to limit researcher discretion and to clarify which analytic decisions were made in advance of seeing the data and which grew out of data exploration (Monogan 2015). Judging from public websites that record the use of preregistration and pre-analysis plans, their use has grown

dramatically, and there seems to be an emerging norm among experimental researchers that best practices involve submitting these documents.

A distinct but related development concerns increased discussion of how to generalize from experiments. Generalization is fundamentally a theoretical issue but one that draws on empirical insights gleaned from the study of heterogeneous treatment effects across subjects, treatments, contexts, and outcomes. One way to advance this agenda is to conduct experiments in multiple contexts, as exemplified by EGAP's metaketa initiative that "funds and coordinates studies across countries, clustered by theme, to improve and incentivize innovative research alongside integrated analysis and publication" (http://egap.org/metaketa). This is an exciting advance given that, to-date, multi-country experiments are rare; our content analysis found only 6% of experiments included multiple countries in 2000-2009, and just 5% in the most recent decade. Of course, conducting experiments across countries requires careful thought about the comparability of measures across contexts; the qualitative data gathering that is used to validate and refine measurement reflects the disciplinary trend towards multi-methods research (e.g., Seawright 2016). The final feature of experimental political science 2.0 is the application of the method to novel areas that historically have not used randomized control trials. As will be highlighted in the volume, this includes topics such bureaucracy, corruption, and censorship – areas that can now be studied experimentally thanks to the aforementioned innovations in design, data access, and analysis.

We next turn to how this volume is structured so as to help scholars, students, and practitioners navigate experimental political science 2.0. Our goal is to help experimental political scientists thoughtfully design studies, analyze data, present results, and expand the application of experiments.

This Volume

We chose topics for the volume that are not only current but also emergent. We hope to stay one step ahead of the curve. Perhaps most importantly, we opted for areas and authors that connect with one another – this book is not a jumble of standalone chapters. Common themes surface throughout, such as the importance of connecting theory to design, making design choices that maximize generalizable inference, and using experiments to extend the frontier of knowledge, which means exploring difficult and even dangerous topics. We organized the book into seven sections, but the chapters intersect both within and across sections. Each chapter includes an abstract, so instead of summarizing them here, we highlight connections to provide readers with a roadmap of how the contributions relate to one another.

The first section includes discussions of experimental designs that are (relatively) newly applied in political science. Conjoint studies – covered in a chapter by Bansak, Hainmueller, Hopkins, and Yamamoto – ask participants to make choices across multi-dimensional descriptions of people, policies, or issues; for instance, this approach may involve soliciting opinions about immigrants who vary in their country of origin, religion, age, education, language skills, etc. Audit experiments, covered by Butler and Crabtree, involve sending correspondence to public officials, randomly varying the nature of the messages, and testing whether the different messages elicit different responses. For example, does a legislator's propensity to respond to constituent mail depend on whether the author has a putative White name or Black name? Both conjoint and audit designs allow political scientists to gauge difficult-to-isolate behaviors such as racial discrimination, gender biases, or illegal actions, because respondents remain unaware of what is being assessed (e.g., they are not directly asked about prejudice or corrupt behavior). The rigor and breadth of these experimental designs explain why they also play a central role in other

parts of the book that use experiments to illuminate hidden or corrupt activities and identity-based discrimination.

Applications of conjoint and audit designs depend on context – such as the level of scrutiny of hidden actions or the nature of gender norms. Two other designs focus even more on context. In their chapter, Kalla, Broockman, and Sekhon present a design that combines survey and field experiments – by first surveying respondents, then employing an ostensibly unrelated field intervention, and then surveying them again. This approach, which has clear cost advantages over other designs, is particularly germane to situations where field interventions seek to change attitudes and beliefs. Additionally, lab-in-the-field studies – where the lab is constructed in a field setting – allow researchers to study choices that reflect subjects' traits and strategic judgments. Eckel and Londono, in their chapter, detail several such examples while also explaining best design practices. All four of these designs – audit, conjoint, field-survey, and lab-in-the-field – constitute alternative approaches to measurement and casual inference across contexts. They also, in theory, could be combined – one could imagine a field-survey study where the survey component includes a conjoint design.

Stepping back from the details of specific designs, one can reflect on two larger issues. First, with one exception, experimental designs involve an intervention by the researcher. The exception is the so-called natural experiment, which has become popular in political science (e.g. Dunning 2012). But what counts as a natural experiment? What separates an experiment from a non-experimental study that is said to involve an "as-if" random assignment? This question is taken up in the chapter by Titunik. Her discussion clarifies what constitutes an actual experiment as opposed to a natural experiment and describes the advantages and disadvantages of each approach. Second, experimental interventions inherently involve ethical issues since the

researcher is changing the world in some way and, perhaps deceptively or unobtrusively, involving people in a research project. Teele's chapter offers a discussion of how to think about the ethics of consent in experiments.

The second section of the book covers data sources that have become more widely used in the last decade. Each of these chapters connects directly to themes raised in the design section. For instance, the goal of many audit studies is to explore racial or ethnic discrimination by political elites. This aim requires using elite samples, a topic covered by Grose in his chapter. Grose also discusses other designs (e.g., natural experiments) that have been used to study the behaviors of those who govern. Apart from elite samples, perhaps, the most notable development when it comes to data sources is, as mentioned, the use of crowdsourcing platforms and non-probability internet panels. These sources offer many research opportunities, but how to assess the impacts of these distinct samples is not always clear – this topic is addressed in the chapter by Krupnikov, Ham, and Style. Another recent data source comes from social media, which offer experimentalists opportunities for new samples and behavioral measures as well as a context within which to study social relationships. Guess's contribution provides one of the first overviews of the emerging experimental literature. Finally, the aforementioned explosion of field experiments of varying types (e.g. lab-in-the-field, field-survey) presents challenges to data collection with targeted populations. Partnering with organizations often can facilitate experimentation, but there is currently no "how to" guide for developing and sustaining collaborations. Levine offers this guidance in his chapter. Even if one does not anticipate using one of the data sources covered in this section, the reading is obligatory for anyone who wants to understand why a research program opts for a particular source of data.

The third section of the book contains just two chapters but touches on issues fundamental to nearly all experiments: once a research question is formulated, treatments and measures must be developed, which in turn presents questions of validity and generalizability. Perhaps ironically, given the rise of experiments in the discipline, there exists limited guidance on how to develop and deploy treatments. Mutz's chapter fills this gap, emphasizing the need to connect treatments to theory. For instance, if a lab-in-the-field study aims to explore the impact of emotion, the treatment needs to trigger emotion even if it does so in a way that does not resemble a stimulus in the "real world." Mutz stresses the importance of empirical verification that the intervention produces the intended change (e.g., in emotion) with no other unintended changes. This requirement involves delicate questions of measurement and conceptualization of the theoretically-specified treatment. As Mutz explains in her chapter, most work to date has not engaged in sufficient empirical verification. In their chapter, Peterson, Westwood and Iyengar also discuss ways to enhance treatments and measures, particularly in the context of survey experiments. A long-standing problem with many survey experiments concerns the use of vignettes that sometimes convey information beyond what the researcher intended (e.g. Dafoe et al. 2018); another problem is social desirability bias, which occurs when research participants confect responses that they hope will please the interviewer. These authors provide advice on how to develop more valid treatments and outcome measures. This advice is of particular importance for experimentation because the objective measures they discuss facilitate symmetric comparisons across treatment and control groups, which is crucial for unbiased inference.

The fourth section turns to longstanding methodological issues and recent advances in addressing them. One such challenge is understanding the causal mechanisms by which an experimental intervention influences an outcome. In his chapter, Glynn starts by pointing out the

formidable design and analysis challenges that arise when researchers attempt to isolate causal mechanisms; his review covers recent technical developments and their implications for applied research. Another burgeoning literature considers the challenges of drawing reliable inferences about which types of subjects are most responsive to treatment. Ratkovic's review of this literature calls attention to the growing role that machine learning methods are playing in the discovery and validation of subgroup differences in responsiveness to treatment. In their chapter, Aronow, Eckles, Samii, and Zonszein address an assumption that is typically invoked in experimental analysis, namely, that subjects respond exclusively to their own treatment assignment and no one else's. The chapter considers what happens when this assumption is relaxed and effects are transmitted across space or via a social network. The chapter's more advanced material reviews the ways in which experimental researchers across the social sciences have come to design and analyze experiments to detect spillovers of various types. The recurrent theme of analyzing data in ways that reflect the underlying experimental design culminates in Coppock's chapter on visual presentation, which offers a series of presentation principles to guide experimental researchers. We are grateful to the publisher for printing Coppock's chapter in color and hosting online the open-source code for his examples, so that readers can make the most of this work.

The volume's fifth section turns to foundational social science issues on how to conduct experimental political science research in a transparent, credible, and generalizable fashion. All of the chapters in this section are of relevance to social scientists who hope to use experiments in a credible manner going forward, regardless of design, sample, measurement, or method. Chapters by Boudreau and Malhotra assess the role of transparency and publication bias in experiments, respectively. A chapter by Seawright describes the benefits of taking a multi-

method approach to experimentation. This chapter amplifies and illustrates themes from previous chapters: how to develop valid treatments, measure outcomes accurately, and detect spillover effects. Two chapters grapple with the issue of generalization. Much of the history of experimental political science has focused the value of clear causal inference, but the newest generation of work asks for more – it wants to make broader statements that carry across samples and contexts. Hartman provides a discussion of the design assumptions that must be made to warrant generalization and discusses methods that attempt to meet these requirements. Blair and McClendon offer a framework for how communities of experimental researchers can learn from studies conducted in multiple contexts. They also explain how designs in particular contexts (e.g., countries) can be employed when the goal is to transport and generalize inferences about cause and effect. These kinds of ambitious designs are becoming increasingly common across subfields.

Finally, we include two sections on substantive areas that are of special prominence and tied to methodological issues discussed in the other parts of the book. The first explores topics related to ethnic identity, racial identity, and gender. These are not new topics, but they have attracted increasing attention from experimental researchers across the globe. In her chapter, Spry introduces readers to experiments on identity. Her discussion of measurement calls attention to promising approaches that allow respondents to express multiple ethnic identities and differentiate between demographic categories and identification with those categories. Valenzuela and Reny's chapter takes on the topic of ethnic and racial priming; while much has been learned on this topic, the authors point out that researchers have only begun to consider the range of priming effects and the contexts in which they occur. Klar and Schmitt, in their contribution, also discuss how political changes – in their case with regard to women in office

and gender stereotypes – have affected the design of experiments on gender in elections. These authors engage an old literature – going back forty years – and highlight some longstanding challenges of design and measurement. In their chapter, Clayton and Anderson-Nilsson review gender experiments in a comparative context, noting the empirical and theoretical challenges of explaining whether and when results generalize across settings. Addressing this question is difficult, and the authors discuss a host of design challenges, including ethical ones.

The last section of the book continues the theme of applying experiments to complex topics that have only recently featured active experimentation. The authors discuss design and data obstacles, robust findings and gaps, and theoretical implications. Nathan and Whites's chapter on experiments on street-level bureaucrats (e.g., social service administrators, election officials, police officers) complements earlier chapters on audit experiments and experiments involving elites. Their chapter instructs scholars on how to design studies to address a host of challenges involving statistical power, the potential for spoiling the sample pool, spillover between subjects, and ethical constraints. Lagunes and Seim's chapter takes up a related and similarly nettlesome topic for experimenters – corruption and corruption control. Corruption by its very nature is designed to elude detection, which makes social science measurement difficult and sometimes dangerous. Nonetheless, the authors offer a way forward that sheds light on micro motives and institutional mechanisms to control corruption. Pan's chapter looks at distinct governmental activities meant to be hidden, such as censorship and repression. Validity and ethical questions abound in this area, and Pan lays these out in a systematic manner, highlighting connections with other chapters such as Butler and Crabtree's, Nathan and White's and Lagunes and Seim's. In her chapter, Matanock considers the challenges of using experiments to understand post-conflict contexts. Addressing the vast literature on peace stabilization and peace

consolidation, she highlights the role of experiments in understanding enduring peace. McGrath's essay on climate change highlights a multi-layered world problem that involves citizens' opinions and behaviors, policies, and international collaboration. Experiments are perhaps the most promising method to disentangling the causal processes that may help address one of the most pressing world challenges.

The book concludes with reflections from Lynn Vavreck. She details the evolution of the field from narrow interventions to complex and ambitious experiments designed to elaborate theories. The result is that experiments now form a central part of the science of studying politics.

Conclusion

Political science has come a long way since A. Lawrence Lowell's 1909 presidential address to the American Political Science Association where he notably stated "We are limited by the impossibility of experiment. Politics is an observational, not an experimental science…" (Lowell 1910, 7). The last decade has made clear that experiments are in fact possible in virtually all areas of the discipline. The question no longer is whether one can use experiments but rather how to use them thoughtfully to shed light on political phenomena of theoretical and practical interest. This volume aims to ensure that experimentalists employ the method in ways that provide for the optimal accumulation of knowledge.

References

Acharya, Avidit, Matthew Blackwell and Maya Sen. 2016. "Explaining Causal Findings Without Bias: Detecting and Assessing Direct Effects." *American Political Science Review* 110 (3): 1-18.

American Political Science Association. 2012. *A Guide to Professional Ethics in Political Science. Second Edition.* District of Columbia.

Aronow, Peter M. 2012. "A General Method for Detecting Interference Between Units in Randomized Experiments." *Sociological Methods & Research* 41 (1): 3-16.

Baker, Monya. 2016. "Is There a Reproducibility Crisis?" *Nature* 533 (7604): 452–54.

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 (3): 540–57.

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 (3): 351-68.

Bertrand, Marianne, and Sendhil Mullainathan. 2004. "Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination." *American Economic Review* 94 (4): 991-1013.

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: 295-98.

Bowers, Jake, Mark M. Fredrickson, and Peter M. Aronow. 2016. "Research Note: A More Powerful Test Statistic for Reasoning about Interference between Units." *Political Analysis* 24 (3): 395-403.

Brown, Andrew W., Tapan S. Mehta, and David B. Allison. 2017. "Publication Bias in Science." In *The Oxford Handbook of the Science of Science Communication*, eds. Kathleen Hall Jamieson, Dan M. Kahan, and Dietram A. Scheufele. New York: Oxford University Press.

Busby, Ethan C. 2018. *It's All about Who You Meet: The Political Consequences of Intergroup Experiences with Strangers*. Ph.D. dissertation, Northwestern University.

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 (3): 463-77.

Buzin, Andrei, Kevin Brondum, and Graeme Robertson. 2016. “Election Observer Effects: A Field Experiment in the Russian Duma Election of 2011.” *Electoral Studies* 44: 184-91.

Camerer F. Colin, Anna Dreber, Eskil Forsell, Teck-Hua Ho, Jürgen Huber, Magnus Johannesson, Michael Kirchler, Johan Almenberg, Adam Altmejd, Taizan Chan, et al. 2016. “Evaluating Replicability of Labor Experiments in Economics.” *Science* 351 (6280): 1433-36.

Connors, Elizabeth C., Yanna Krupnikov, and John Barry Ryan. 2019. “How Transparency Affects Survey Responses.” *Public Opinion Quarterly* 83 (1): 185-209.

Coppock, Alexander. 2019. “Generalizing from Survey Experiments Conducted on Mechanical Turk: A Replication Approach.” *Political Science Research and Methods* 7 (3): 613-28.

Costa, Mia. 2017. “How Responsive are Political Elites? A Meta-Analysis of Experiments on Public Officials.” *Journal of Experimental Political Science* 4 (3): 241-54.

Dafoe, Allan, Baobao Zhang, and Devin Caughey. 2018. “Information Equivalence in Survey Experiments.” *Political Analysis* 26 (4): 399-416.

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: 389-95.

Druckman, James N., Adam J. Howat, and Kevin J. Mullinix. 2018. “Graduate Advising in Experimental Research Groups.” *PS: Political Science & Politics* 51 (3): 620-24.

Druckman, James N., Donald P. Green, James H. Kuklinski, and Arthur Lupia. 2006. “The Growth and Development of Experimental Research in Political Science.” *American Political Science Review* 100 (4): 627-35.

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.

Dunning, Thad. 2012. *Natural Experiments in the Social Sciences: A Design-Based Approach. Strategies for Social Inquiry*. New York: Cambridge University Press.

Dunning, Thad, Guy Grossman, Macartan Humphreys, Susan D. Hyde, Craig McIntosh, and Gareth Nellis, eds. 2019. *Information, Accountability, and Cumulative Learning: Lessons from Metaketa I*. New York: Cambridge University Press.

Fanelli, Daniele. 2018. “Is Science Really Facing a Reproducibility Crisis?” *Proceedings of the National Academy of Sciences of the United States of America.* 115 (11): 2628-31.

Findley, Michael G., Daniel L. Nielson, and Jason Campbell Sharman. 2014. *Global Shell Games: Experiments in Transnational Relations, Crime, and Terrorism*. New York: Cambridge University Press.

Franco, Annie, Neil Malhotra, and Gabor Simonovits. 2014. "Publication Bias in Social Science: Unlocking the File Drawer." *Science* 345 (6203): 1502-05.

Franco, Annie, Neil Malhotra, and Gabor Simonovits. 2016. "Underreporting in Psychology Experiments from a Study Registry." *Social Psychology and Personality Science* 7 (1): 8-12.

Freese, Jeremy, and David Peterson. 2017. "Replication in Social Science." *Annual Review of Sociology* 43: 147-165.

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 (1): 81-98.

Gerber, Alan S., Kevin Arceneaux, Cheryl Boudreau, Conor Dowling, and 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 (2): 216-29.

Gerber, Alan S., and Donald P. Green. 2000. "The Effects of Canvassing, Telephone Calls, and Direct Mail on Voter Turnout: A Field Experiment." *American Political Science Review* 94 (3): 653-63.

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.

Gosnell, Harold F. 1926. "An Experiment in the Stimulation of Voting." *American Political Science Review* 20 (4): 869-74.

Green, Donald P., and Holger L. Kern. 2012. "Modeling Heterogeneous Treatment Effects in Survey Experiments with Bayesian Additive Regression Trees." *Public Opinion Quarterly* 76 (3): 491-511.

Green, Donald P., and Andrej Tusicisny. 2012. "Statistical Analysis of Results from Laboratory Studies in Experimental Economics: A Critique of Current Practices." Paper presented at the North American Economic Science Association (ESA) Conference, Tucson, AZ.

Grimmer, Justin, Solomon Messing, and Sean J. Westwood. 2017. "Estimating Heterogeneous Treatment Effects and the Effects of Heterogeneous Treatments with Ensemble Methods." *Political Analysis* 25 (4): 413-34.

Grose, Christian R. 2014. “Field Experimental Work on Political Institutions.” *Annual Review of Political Science* 17: 355-70.

Grossman, Guy, and Kristin Michelitch. 2018. “Information Dissemination, Competitive Pressure, and Politician Performance between Elections: A Field Experiment in Uganda.” *American Political Science Review* 112 (2): 280-301.

Groves, Robert M. 2011. “Three Eras of Survey Research.” *Public Opinion Quarterly* 75 (5): 861-871.

Humphreys, Macartan, and Jeremy M. Weinstein. 2009. “Field Experiments and the Political Economy of Development.” *Annual Review of Political Science* 12: 367-78.

Hyde, Susan D. 2015. “Experiments in International Relations: Lab, Survey, and Field.” *Annual Review of Political Science* 18: 403-24.

Hyde, Susan D., and Nikolay Marinov. 2014. “Information and Self-Enforcing Democracy: The Role of International Election Observation.” *International Organization* 68 (2): 329-59.

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 (1): 292-307.

Imai, Kosuke, and Marc Ratkovic. 2013. “Estimating Treatment Effect Heterogeneity in Randomized Program Evaluation.” *The Annals of Applied Statistics* 7 (1): 443-70.

Imai, Kosuke, and Aaron Strauss. 2011. “Estimation of Heterogeneous Treatment Effects from Randomized Experiments, with Application to the Optimal Planning of the Get-Out-the-Vote Campaign.” *Political Analysis* 19 (1): 1-19.

Imai, Kosuke and Teppei Yamamoto. 2013. “Identification and Sensitivity Analysis for Multiple Causal Mechanisms: Revisiting Evidence from Framing Experiments.” *Political Analysis* 21 (2): 141-71.

Jones, Jason J., Robert M. Bond., Dean Eckles, and James H. Fowler. 2017. “Social Influence and Political Mobilization: Further Evidence from a Randomized Experiment in the 2012 U.S. Presidential Election.” *PloS ONE* 12 (4). doi: 10.1371/journal.pone.0173851.

Kertzer, Joshua D., and Ryan Brutger. 2016. “Decomposing Audience Costs: Bringing the Audience Back into Audience Cost Theory.” *American Journal of Political Science* 60 (1): 234-49.

Kim, Eujin. 2018. “Entertaining Beliefs in Economic Mobility.” Working Paper, University of Pennsylvania.

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 (24): 8788-90.

Lowell, A. Lawrence. 1910. “The Physiology of Politics.” *American Political Science Review* 4 (1): 1–15.

Lupia, Arthur, and Colin Elman. 2014. “Openness in Political Science: Data Access and Research Transparency.” *PS: Political Science and Politics* 47 (1): 19-42.

Monogan, James. E. 2015. “Research Preregistration in Political Science: The Case, Counterarguments, and a Response to Critiques.” *PS: Political Science and Politics* 48 (3): 425-29.

Mullinix, Kevin J., Thomas J. Leeper, James N. Druckman, and Jeremy Freese. 2015. “The Generalizability of Survey Experiments.” *Journal of Experimental Political Science* 2 (2): 109-38.

Mutz, Diana C. 2011. *Population-Based Survey Experiments*. Princeton, NJ: Princeton 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 (2): 192–215.

Nickerson, David W. 2008. “Is Voting Contagious? Evidence from Two Field Experiments.” *American Political Science Review* 102 (1): 49-57.

Nosek, Brian A., Geroge Alter, Geroge C. Banks, Denny Borsboom, Sara D. Bowman, Steven J. Breckler, Stuart Buck, et al. 2015. “Promoting an Open Research Culture.” *Science* 348 (6242): 1422-25.

Open Science Collaboration. 2015. “Estimating the Reproducibility of Psychological Science.” *Science* 349: aac4716-1-aac4716-8.

Pager, Devah. 2003. “The Mark of a Criminal Record.” *American Journal of Sociology* 108 (5): 937-75.

Paluck, Elizabeth Levy, and Donald P. Green. 2009. “Deference, Dissent and Dispute Resolution: An Experimental Intervention Using Mass Media to Change Norms and Behavior in Rwanda.” *Political Science Review* 103 (4): 622-44.

Ratkovic, Marc, and Dustin Tingley. 2017. “Sparse Estimation and Uncertainty with Application to Subgroup Analysis.” *Political Analysis* 25 (1): 1-40.

Rogowski, Ronald. 2015. “The Rise of Experimentation in Political Science.” In *Emerging Trends in the Social and Behavioral Sciences: An Interdisciplinary, Searchable, and Linkable Resource*, eds. Robert A. Scott and Stephen M. Kosslyn. John Wiley & Sons. doi: 10.1002/9781118900772.etrds0409.

Seawright, Jason. 2016. *Multi-Method Social Science: Combining Qualitative and Quantitative Tools*. Cambridge: Cambridge University Press.

Scacco, Alexandra, and Shana S. Warren. 2018. “Can Social Contact Reduce Prejudice and Discrimination? Evidence from a Field Experiment in Nigeria.” *American Political Science Review* 112 (3): 654-77.

Sniderman, Paul M., and Douglas B. Grob. 1996. “Innovations in Experimental Design in Attitude Surveys.” *Annual Review of Sociology* 22 (1): 377-99.

Teele, Dawn Langan. ed. 2014. *Field Experiments and Their Critics: Essays on the Uses and Abuses of Experimentation in the Social Sciences.* New Haven: Yale University Press.

Tomz, Michael. 2007. “Domestic Audience Costs in International Relations: An Experimental Approach.” *International Organization* 61 (4): 821-40.

Van Bavel, Jay J., Peter Mende-Siedleckia, William J. Bradya, and Diego A. Reinero. 2016. “Contextual Sensitivity in Scientific Reproducibility.” *Proceedings of the National Academy of Sciences of the United States of America* 113 (23): 6454-59.

Wantchekon, Leonard. 2003. “Clientelism and Voting Behavior: Evidence from a Field Experiment in Benin.” *World Politics* 55 (3): 399-422.

Zigerell, L.J. 2018. “Black and White Discrimination in the United States: Evidence from an Archive of Survey Experiment Studies.” *Research and Politics* 5 (1): 1-7.