Online Experimental Methods

Thomas J. Leeper
Associate Professor in Political Behaviour
Department of Government
London School of Economics and Political Science

Summary

Online experimental methods have become a major part of contemporary social science research. Yet the method is also controversial and experiments are frequently misunderstood. This chapter introduces online experimentation as a method, by explaining the logic of experimental design for causal inference. While experiments can be deployed in almost any setting, online experiments tend to take two forms: online survey experiments and experiments in naturalistic online environments. Discussing the advantages and disadvantages of these types relative to each other and relative to their offline analogues, the chapter demonstrates ways that experimentation has been used to learn about political behavior, media and campaign dynamics, and public opinion.

Emphasizing trade-offs between internal validity, experimental realism, and external validity, I demonstrate how researchers have used online platforms in tandem with randomization to gain insights into both online and offline phenomena. Though experiments are sometimes seen as trading off external validity for internal validity, this is not an accurate depiction of all experimental work. Rather, online experiments exist on spectrums that trade-off these features to varying degrees. And with those trade-offs come key challenges related to experimental control, the generalizability of experimental results across settings, units, treatments, and outcomes, and the ethics of online experimentation. The chapter concludes by suggesting how future research might innovatively push beyond existing work.

Keywords: experiments, research design, survey experiment, field experiment, causal inference, social media, internal validity, external validity, research ethics

Prepared for inclusion in the *Oxford Handbook of Electoral Persuasion*

6 August 2018

In the methodological toolkit of contemporary social scientists, online experimentation has become ordinary, even banal. We arrived at this point shockingly quickly. In 1990, this chapter would not have been written; the internet barely existed. In 2000, this chapter would not have been written; while widespread, the world wide web hardly resembled the diffuse and omnipresent institution we know today and hardly any research was taking place online. In 2010, this chapter would have no doubt been preoccupied with questions about the validity of online samples for studying social processes and the debate would have been about online platforms and marketing research firms that no longer exist (see Baker et al. 2010). Much has changed in the past decade and will change in the coming years.

Yet readers will come to this *Handbook* looking for answers about what has been studied using online experiments, what could plausibly be studied in this way, and what they should consider when embarking down this methodological path. In what follows, I introduce online experimentation as a method and then discuss a few key considerations when pursuing online experiments, with special emphasis on much-debated trade-offs between internal validity, experimental realism, and external validity. Throughout, I demonstrate how researchers have used online platforms in tandem with randomization to gain insights into both online and offline phenomena, discuss some of the key challenges of conducting experimental research online, and conclude by suggesting how future research might innovatively push beyond existing work.

Online Experimentation in a Nutshell

An online experiment is a research design involving randomly assigned interventions applied to units in an internet-mediated environment such as website, app, or online survey. The diversity of samples, interventions, and online environments and combinations thereof means it can be difficult to succinctly characterize the method or readily identify all of its extant applications. Units might be users of particular websites or apps (Bond et al. 2012), or they

might be samples of individuals recruited from online survey panels (Mutz 2011; Callegaro et al. 2014) or crowdsourcing websites (Mason and Suri 2011; Berinsky et al. 2012; Leeper 2013), or the unit of analysis might websites or social media pages themselves. Online experiments might entail interventions that occur in the context of a research experience, like a survey, or they might occur at the behest of a website in the course of everyday business ("a/b testing"; see King, Churchill, and Tan 2017), or they might be interventions fielded via online platforms without any particular involvement from the platforms themselves (e.g., Feezell 2017; Munger 2017). Online experiments may entail recruitment of participants, the application of interventions, and the measurement of outcomes entirely online, or they may mix online and offline experiences and measurements such as survey-based measures of online interventions or online behavioral measures of offline interventions (Broockman, Kalla, and Sekhon 2017). The possibilities are almost limitless.

But with so many possibilities, what is the point of an online experiment and why might researchers opt for any particular kind of online experiment? Like any experiment, one that occurs – at least in part – in an online environment should have one primary purpose: to leverage the random assignment of stimuli across units to infer the causal effect of those stimuli on outcomes of interest (see McDermott 2002; Gerber and Green 2012). This may seem obvious, but is something non-obvious to many researchers with a background in the survey-based study of political behavior. In survey research, questionnaire methods are used to construct quantitative and qualitative measurements of human respondents that are then analyzed using a variety of multivariate techniques (Groves et al. 2009). Experimental research therefore differs considerably from survey research in that the latter aims to provide inference primarily about descriptive relationships between measures; the former serves only one principle purpose: to establish the direction and magnitude of causal effects through design-based inference. The kinds of research questions that can be readily answered by

online experiments therefore differ from the questions that can be readily answered either by online surveys or observational methods such as textual analysis of website content or network analyses of social media users. These latter examples tend to focus on questions of sentiment, followership, network structure, and so forth that require computational approaches sufficiently complex that descriptive answers are interesting per se. Causal inference using textual data (Egami et al. 2017) and network structures (Taylor and Eckles 2017) are at this stage still emergent literatures.

How do experiments help address questions about causality? Randomization solves inferential challenges that would otherwise need to be addressed by mathematical modelling. Consider, for example, Valenzuela's (2013) study of the effects of social media use on participation in political protests in Chile. Here, assessing the direction and magnitude of influence of the allegedly causal variable—social media use—requires statistically controlling for those factors which might confound the relationship between it and the outcome—factors like education, age, and political engagement. Yet building such a complete model is tricky due to incomplete knowledge of the complete causal process that is at work between all variables. Randomization overcomes this problem by trading off depth of knowledge of the broader causal processes at stake (and assumptions that it can be and has been fully modelled) in exchange for superior internal validity (see McDermott 2011). Because individuals entering the experiment are randomly assigned to each of the experimental conditions, they are – and their broad experiences of the experiment are – identical, on average, save for the unique stimulus provided to their experimental group.

While their post-treatment outcomes are affected by many factors outside of the context of the experiment, those influences are balanced on average across the groups such that any difference in average outcomes between the experimental conditions can safely be attributed to the impact of the experimentally manipulated stimuli – with usual caveats of

uncertainty about the precise size of that effect. This inference can proceed without any broader knowledge of the causal processes involved in a phenomenon. For example, a researcher interested in how exposure to particular online content affects whether individuals follow political candidate accounts on social media is very unlikely to have access to either complete knowledge of the drivers of followership decisions or the means of operationalizing many of most likely causes. An online experimental design can readily assess the causal effect of information on followership by either supplying that information in a survey-experimental intervention and measuring intended or actual followership, or by deploying that information to users directly on the platform and measuring real followership directly. The difference in the proportion of treated users following the account less the proportion of untreated users following the account provides a direct estimate of the causal effect of information on followership without the need for further statistical control. 'Design trumps analysis,' to borrow a phrase from Rubin (2008). Gerber and Green (2012) provide a thorough and useful overview of the statistical issues involved in such inferences.

The use of an online mode carries with it clear trade-offs compared to other common modes of experimentation, namely laboratory settings and field settings. While all three modes leverage randomization in the same way to obtain internal validity, the insights gained from each depend on the strengths and limitations afforded by the change of mode. Laboratory experiments offer a heightened degree of control compared with other experimental sites as laboratories can be regulated in almost any imaginable way: the type of room involved, the environmental conditions such as temperature and sound, the social conditions of the experiment through isolation or the use of confederates, and the timing and duration of the experiment. Field studies relax control considerably either to gain external validity in terms of the representativeness of the experimental sample or setting (Mutz 2011),

or to additionally gain mundane realism in the way stimuli are applied and/or outcomes measured (Morton and William 2010; McDermott 2011; Dickson 2011). Online experiments rather than falling strictly in between extremes of maximal control and maximal realism instead exist on continuum between experiments that in many ways resemble laboratory studies save for their physical location to experiments that very much resemble field studies save for their use of an online device to deliver stimuli or measure outcomes or both. For example, Druckman and Leeper (2012) conducted an online experiment with a sample of undergraduate students in which they randomly assigned respondents to receive different textual stimuli and then measured how repeated exposure to those stimuli affected political preferences over time. Reliance on a convenience sample of respondents means the experiment likely could have also been performed in a laboratory setting. Similarly, Esterling, Neblo, and Lazer (2011) conduct an experiment in which representative samples of American participants were randomly assigned to have an opportunity to deliberate with their Member of Congress using an online "e-townhall" platform. The principal stimulus (the deliberative experience) could have alternatively been an in-person event given participants were all drawn from relatively small Congressional districts, changing the online experiment into a field experiment (e.g., Barabas 2004). Online experiments present a diverse array of research possibilities with no strict and universally applicable advantages or disadvantages relative to other experimental approaches.

Online Survey Experiments and Online Field Experiments

Online experiments can be seen as broadly of two classes: 1) those that are administered online but could be administered through another mode and 2) those that uniquely leverage features of an online environment in a way that would be impossible in other modes. Experiments of the first kind tend to be those that study general social or political phenomena and operate as survey experimental experiences. Experiments of the second kind tend to

focus online phenomena and online environments per se (though they may sometimes leverage digital technologies to gain broader insights). Both kinds are obviously useful.

Experiments of the First Kind

Online experiments of the first kind are what many behavior researchers are likely to deploy. These experiments tend to begin with questions of general political science interest and evolve into projects that occur online. For example, researchers interested in how citizens decide for whom to vote could use numerous methods such as surveys or decision-making experiments. They could just as easily administer a dynamic process tracing experiment (Lau and Redlawsk 2006) online or in a laboratory (see Utych and Kam 2014). That research is not about online behavior per se but rather may adopt an online mode for reasons of convenience, cost, feasibility, or technological features afforded by an online as opposed to in-person, telephone, or other mode.

Whereas behavioral aspects of politics have historically been studied using survey interviews and laboratory experiments, these kinds of online experiments now dominate due to the shifting costs of survey interviewing. Mutz (2011) provides useful insights into how online, population-representative survey experimentation has become feasible and prevalent since the late 1990s (see also Sniderman 2011). Leeper (2014), for example, used participants recruited both in a laboratory setting and from an online crowdsourcing platform to study information selection behavior and its effects on political polarization, balancing control of the laboratory with the sample diversity provided by online participants. In two experiments, Messing and Westwood (2014) recruit undergraduate and online research participants to participate in an experimental study of how social recommendations interact with news sources to affect online media consumption. The digital mode in these studies is important but not essential to how the studies were conducted; a move online serves goals of time, costs, or sample composition. The move online involves obvious changes in mode and less

obvious changes in sampling frames. But these changes may not be as consequential as they seem. Questionnaires and stimuli that might have worked in a laboratory or over the telephone may not work as well in an online setting, where respondents are less attentive and devices are more variable. But many established techniques continue to perform well. The reason is that a research participant completing an online survey experiment is interacting with a survey environment essentially identical to one they would complete in a laboratory or in a CASI component of a face-to-face interview. Compared to interviewer-based surveys, self-interviewing dramatically reduces social desirability biases (see, for example, Villarroel et al 2006). The cost, however, comes in the form of a loss of experimental control, with the risk of heightened item nonresponse, satisficing, and attrition. These risks are not qualitatively new. They are merely more severe.

Though the mode and its properties are ultimately familiar, the samples used in online survey experiments are distinct from those common in earlier periods. Indeed, the sampling frame for online participants is often murkier muddier than for traditional, offline populations. And often there is no sampling frame at all. Online survey experiments frequently rely on a combination of online panels of participants and, less frequently, model-based post-survey adjustments (such as post-stratification weighting) to achieve sample representativeness. Where telephone surveys have historically involved probability based random-digit-dial samples and laboratory experiments have historically involved relatively homogeneous convenience samples (e.g., Druckman and Kam 2011; Kam et al. 2007), online survey experiments strike a middle ground. The move online enables access to groups of participants more diverse than those available in a laboratory setting, but not strictly representative of any offline or online population. Research shows online convenience samples yield demographically diverse participants (Levay, Freese, and Druckman 2016), who behave in ways similar to respondents from general population samples (e.g., Clifford et

al. 2015), and respond to treatment in a manner similar to those from nationally representative samples (Mullinix et al. 2015).

Though the particular details of online experiments and online survey samples present new challenges, they ultimately need to be held to familiar standards. The quality of an online survey experiment depends on whether the treatments behaved as intended, whether the outcomes were appropriate, whether the settings were not too heavily localized, and whether the participants resemble population(s) of interest.

Experiments of the Second Kind

More challenging are online experiments of the second kind: namely, those that rely upon the online environment to study a phenomenon or apply experimental techniques that are uniquely digital. Take, for instance, an experiment by Messing and Westwood (2013) in which the researchers recruit student participants into an online environment that is built directly upon the Facebook Application Programming Interface (API), a programmatic interface to participants' Facebook content and friends lists. The environment, while artificial, was populated by real news content shared by their actual Facebook friends, with the composition of the feed randomly manipulated. This experiment, which administers experimental stimuli amidst real online content provides clear advantages in terms of realism for studying processes of social influence on political behavior. The trade-off, however, is the technical difficulty of constructing an entire environment backed by live interaction with a social media site.

Feezell (2017) achieves insight into a similar dynamic: namely the degree to which social media platforms serve an agenda-setting function for users. Feezell's study uses the live Facebook platform rather than an artificial environment, recruiting student participants into a panel study with pre- and post-experimental measures of issue importance. In between

these two time measurements, participants were randomly encouraged to join one of two seemingly identical Facebook groups, the content of which was subtly varied over a 75-day period. By embedding the experimental manipulation into participants' naturalistic experience of the social media platform, the study displays a high degree of realism while relying on a simple experimental manipulation to gain insight into the magnitude of agenda-setting effects. Theocharis and Lowe (2015) desired to understand not just the effects of a particular social media experience but the impact of using social media per se. To do so, they randomly encouraged individuals without Facebook accounts to join the platform and then measured their political participation downstream. Rather than attempt to orchestrate an artificial social media experience, they leverage the platform itself as the experimental stimulus. All of these examples show how online experiments that leverage uniquely digital tools to study uniquely online phenomena and processes can provide insights that would be impossible to obtain in an offline context or in a survey-experimental setting.

In addition to connecting an experimental environment to a social media platform via API integrations or through recruitment of participants to participate in an isolated environment, social media websites can also serve as vehicles for delivering experimental stimuli to a much wider group of users beyond those who have actively agreed to participate. Broockman and Green (2013) use Facebook advertisements to deliver candidate advertisements to demographically targeted subgroups of users over short windows of time and then use telephone surveys to measure the impact of those advertisements on name recognition. Users are unaware they are in an experiment and the experiment closely mimics the real-world behavior of candidate campaigns that might be interested in buying the exact same types of ads for delivery to exactly the same types of voters. Such advertising experiments are not necessarily restricted to social media sites; researchers could easily disseminate ads via other advertising platforms.

Going further, partnerships with social media platforms and commercial websites – particularly those with in-house research teams – provide even greater capacity for insight into such behavior. Bond et al. (2012) and Jones et al. (2017) are experiments that result from research partnerships between academic researchers and Facebook data scientists to study how the platform might be used to stimulate voter turnout. By embedding a social signaling widget within in the live Facebook site the researchers were able to assess the degree to which turnout signals from friends' stimulated voting behavior. While past work has suggested that social signals about voting might be very effective (e.g., Gerber, Green, and Larimer 2008), the Facebook experiments were embedded in a fully naturalistic online environment at a previously unimaginable scale (upwards of 60 million site users were engaged in the experiment).

Challenges

The immense possibilities presented by online experiments – both the survey-experimental kind but especially the naturalistic social media kind –have been under-exploited by social scientists. Jungherr (2016) shows that the vast majority of research into the use of Twitter in election campaigns involves observational analysis of "digital trace data" – namely, information about users and their posted content extracted from via API. Only 10 of 127 studies reviewed relied upon experimental designs. The reasons for this are likely numerous, including the relatively young age of most social media platforms, the technical challenges of conducting online experiments in some environments, and concerns about what can be learned. Indeed, all of the experiments discussed in this chapter raise questions about experimental control, generalizability, and research ethics. These issues are not unique to online experiments or even to experimental work in general, but online experimentation poses some unique or heighted concerns compared to research in other environments.

Online behavior is sometimes likened to digital truth serum – citizens' hidden interests and preferences might be revealed in online behavior in ways that all other existing research methods fail to capture (see, for example, Stephens-Davidowitz 2017). Yet it is not a panacea for the challenges that affect offline research. If anything, it only introduces new challenges; as Titiunik (2015) has argued, neither the number of cases nor the number of variables diminish the need for careful research design if one's goal is causal inference. Researchers need to apply the same broad principles of research design, analytic credibility, and ethics that apply offline to research that occurs online.

Experimental Control Online

While online platforms enable the design of experiments with a high degree of mundane realism and heightened external validity, this necessarily trades off control. Because social media sites are active, vibrant spaces that reflect the heavily personalized nature of participants' social networks and use of the sites, behavior in these live online environments is necessarily noisier and more heterogeneous than in a laboratory setting. Survey experimental interviews completed at home lack the control of a laboratory or the engagement of a face-to-face interview. Researchers therefore need to consider how to balance realism with the experimental control necessary for clean causal inference.

While a breakdown of control might come in a number of forms, three particularly salient threats to validity are posed by online experimentation: noncompliance, interference, and attrition. Noncompliance refers to the idea that experimental participants may not receive the stimulus to which they were randomly assigned, either due to intentional or unintentional non-exposure or due to cross-over wherein participants assigned to one experimental condition instead receive the stimulus for another experimental group. Noncompliance in the form of non-exposure is trivially easy to produce in an online environment: the web is noisy

and a subtle experimental manipulation may be easily lost in the cacophony of content. That noise may be precisely the point (see, again, Lau and Redlawsk 2006) but comes with a loss of obviousness typical of laboratory experimental stimuli. In survey experiments, respondents may be at home or work and multi-tasking during survey-taking. In naturalistic experiments, users may also have content filtering settings on news aggregators or social media that reduce the probability of seeing a particular stimulus, or may be algorithmically more or less likely to see a stimulus due to a combination of their previous web use behavior and the behavior of websites and apps they utilize.

Researchers need to decide how to deal with inattention and noncompliance. While it is instinctual to measure attention to experimental materials and stimuli ("attention checks") and to similarly measure whether participants experienced the treatment as intended ("manipulation checks"), care is needed when deploying these techniques. A common attention checking technique is to ask survey respondents a long question that ends with a clear instruction to select response option "B" and then provides four or five response options that correctly answer the question save for the final sentence instruction. Respondents answering anything but "B" are deemed in attentive. Measuring survey attention may be useful (see Berinsky et al. 2012; Clifford and Jerit 2015) but excluding participants based upon their inattention can be problematic. If excluding participants before randomizing treatment, then the sample composition will have shifted; the experiment will only be about the effects of stimuli on the subset of respondents who are attentive.

Excluding participants *after* treatment is generally a bad idea. For example, an experiment about exposure to an informational message versus a control condition with no information might measure whether respondents remember the information later on (as a manipulation check). Individuals who remember are said to "pass" and individuals who do not remember are said to "fail". Aside from obvious problems with expecting perfect recall, a

decision to exclude those who "fail" the manipulation check results in a broken experiment that compares the subset of treated individuals who remember to the subset of untreated individuals who do not already know the information. These subgroups are non-comparable, so the study is no longer an experiment. Montgomery, Nyhan, and Torres (2018) provide a detailed discussion of why such practices are problematic. Gerber et al. (2014) and Mutz and Pemantle (2015) provide a particularly thorough debate on the value of manipulation checks.

Attrition – the loss of participants from a study over time – presents a similar problem to noncompliance or intentional posttreatment exclusion of participants. The sample in an online experiment might be very large populations (e.g., the US adult population) or very localized ones (e.g., participants self-selected into a social media page). If participants leave a study only before stimuli are applied, then causal inference can safely proceed albeit about a reduced population and with reduced power. If participants leave a study after stimuli are applied, it becomes much more difficult to make credible causal inferences particularly if the intervention per se led participants to leave the study. Similarly, if the goal is study online behavior over a long period of time, continuous loss of participants has the potential to dramatically affect the credibility of inferences if drop-off is at all related to the experimental manipulations and/or outcomes of substantive interest.

Other breakdowns of experimental control may be less likely in survey contexts but much more likely in naturalistic environments. Cross-over is possible if participants cannot be strictly prevented from observing stimuli from other experimental conditions. Feezell (2017), for example, prevented crossover in a naturalistic experiment on Facebook through social media privacy settings, but experimental stimuli deployed in less restricted environments such as on more open platforms – like Twitter, Reddit, or Wikipedia – or more broadly on the web are not easily restricted only to users in one condition. Researchers engaging in these types of online experiments should be particularly cautious about

noncompliance and pre-specify how they will attempt to reduce probability of such occurrences and how they will analyze their resulting experimental data if noncompliance occurs (see Gerber and Green 2012 for guidance).

Interference between units is a particularly tricky problem in online environments. In contemporary causal inference, a strong but important assumption is known as the stable unit treatment value assumption or SUTVA (Holland 1986), which requires that units in an experimental condition all experience the same value of the treatment to which they have been assigned and additionally are unaffected by the experiences of other units in the experiment. Interference is a general problem that has received increasingly methodological interest in recent years (see, for example, Bowers, Fredrickson, and Panagopoulos 2013; Aronow 2012; Sinclair, McConnell, and Green 2012; Aronow and Samii 2015). Because online environments involve both networks between webpages and complex networks between human internet users, content and users are not easily isolated in a way they might be in the independent experience of a survey-like interview. Users in an experiment can easily move from content they have been encouraged to see to content they have been discouraged to see, either because that content is hyperlinked, because it appears together in search results, or because it has been shared by others online. In social media environments, interference is particularly difficult to prevent as the network distance between even seemingly isolated users actually tends toward small single digits: about four on Facebook (Ugander et al. 2011) and even lower on Twitter (Bakhshandeh et al. 2011).

The responses to this severe threat of interference across units are multiple. One can design experimental studies to sacrifice realism by leaving the native web environment in favor of something more artificial to enforce control and strictly prevent interference. Another is to design field studies that attempt as much as possible to avoid interference – for example by actively selecting users who are distant from one another within online networks

or by administering experimental stimuli in geographically isolated ways (e.g., on city-specific forums or news websites). A third possibility and one that has received too little attention within political science is to actively leverage networks to gain causal inference (e.g., Rogowski and Sinclair 2012; Fowler et al. 2011). Taylor and Eckles (2017) provide a general framework for designing and analyzing experiments within network settings.

Generalizability

Perhaps the most frequent concern about experiments generally is whether they *generalize*. These concerns are often about the sample of the units being studied, but also capture concerns about limitations of the setting, interventions, and outcomes. Users willing to participate in a survey experiment or other online experiment may not reflect the broader user population from which they were drawn nor the broader human population (see, for example, Barberá and Rivero 2014) so inferences reached from experimental participants may not apply more broadly. Experiments conducted at a given point in time may be colored by ongoing external events, such as campaigns or news cycles, in ways that may not be easy for researchers to observe or understand. Replication of these kinds of online experiments across times, contexts, geographies, and user groups is essential for reach such generalities if generalization is even possible. Indeed, Shadish, Cook, and Campbell (2002) go so far as to argue that external validity is a feature of *research literatures*, not individual studies, given the highly localized and particularistic nature of all aspects of any experiment (18; see also Cronbach 1986).

While all features of experiments are important for evaluating a study's external validity (Cartwright 2007), methodological writing about generalizability mostly dwells on sample representativeness and quality, reflecting the uncomfortable transition of much survey research from telephone sampling frames and interview modes to maintained online panels using self-interviewing (see Baker et al. 2010; Callegaro et al. 2014).Though non-probability

samples now dominate online behavioral research, not all samples are equivalent. Student convenience samples, which dominated (laboratory-based) behavioral research for most of the 20th century, have been heavily replaced by online convenience samples drawn from crowdsourced labor markets (see Mason and Suri 2011; Goodman and Paolacci 2017). Whether this has meant that researchers now have higher or lower quality samples is a matter of debate, but such samples are certainly more diverse with respect to age, income, educational experience, and some demographic characteristics than student samples. The rise of online survey panels has also meant a dramatic reduction in the cost of obtaining demographically representative samples (typically through "sample matching" or reweighting methods applied to opt-in online panels). Indeed, the price of such samples provided by market research firms is often comparable to that of crowdworkers paid minimum wage. Probability-based online samples remain far more expensive.

Rather than dwell upon the multitude of ways that a sample's "quality" might be evaluated—demographic representativeness, satisficing behaviors, etc.—my view is that researchers should focus not on surface features of samples but whether literatures of multiple studies generate similar results. Research suggests that online convenience samples provided by online crowdsourcing marketplaces generate experimental inferences that closely approximate those of more traditional and more representative samples (Mullinix et al. 2015; Coppock N.d.; Coppock, Leeper, and Mullinix N.d.). The surface dissimilarity of "low-quality" and "high-quality" samples in terms of respondent characteristics therefore seem much less important than they might initially appear.[1] Researchers working with online samples should consider the multitude of ways these samples may be similar or different *in terms of substantive inferences they might generate*, ideally by replicating experiments

[1] Bigger concerns with small online pools of participants relate to non-naivete and questions about respondent burden and what impact that might have on behavior (Stewart et al. 2015; Krupnikov and Levine 2014).

multiple times across samples and other contextual variations. The question of whether a sample generates similar inferences to a counterfactual sample is precisely what we want to know; all other questions about features of samples are secondary to this main criterion.

A final point needs to be made about research ethics. Online research is qualitatively different from laboratory, telephone, or field studies where researchers or research assistants have a direct, typically aural interaction with research participants. Because online participants are nameless, faceless individuals who live in an unspecified location on the other end of a vast computer network, it is easy to lose sight of their humanity and to invalidly justify avoidance of ethical standards for research conduct. As a prominent example, concerns about their treatment by researchers in the 2010s led crowdworkers to go so far as to publish a manifesto calling for improved ethics in online experiments (http://www.wearedynamo.org/). Online participants, being humans, are entitled to respect and the same standard of care given to participants in other settings. If participants in an offline setting would be compensated for their participation, online participants should likely be compensated in a similar manner. If participants in an offline setting would be given particular information prior to seeking their consent to participate, online participants should likely be given the same information. If offline participants would be given opportunities to leave a study or remove their data, online participants should likely be given the same opportunities. This should be obvious, but numerous experiences in the past decade suggest it is not. Though not exclusively about online experiments, Desposato's (2015) edited volume provides useful case studies and discussions about ethical practice.

Research Ethics Online

This transition from questions of generalizability to ethics is vital – efforts to be generalizable by seeking out new platforms and new types of participants almost necessarily invite new ethical questions. Indeed, all of the practical and methodological challenges posed

by online experimentation are relatively minor compared to the significant ethical issues at stake in online experimentation (see, for an extensive discussion, Salganik 2017, ch.6). While much ethical writing in the social sciences builds upon an implicit legal framework – typically the United States Common Rule – online environments are a uniquely difficult space to ethically navigate because the internet lives nowhere and everywhere and research into online environments is complicated by ethical obligations of researchers, national and international laws governing online data, and websites' terms of use. As all of these factors – laws, regulations, and terms of use – are prone to rapid change, it does not make sense to review their provisions specifically but rather engage with the broad issues raised by research in these spaces. Among the most important themes are consent, deception, data protection, and influence upon the political and social world. Researchers in some countries (e.g., the US) will be obligated to comply with ethical guidelines but these kinds of review boards do not exist in all countries.

Online experimentation, as has been shown, can vary between an archetype that closely resembles laboratory experimentation where participants are self-selected into the research, held temporarily captive in an artificial environment, and monitored only for a limited period of time, and a much different world in which participants are unaware of their participant in an experiment as their real lived experience of an online space is subtly manipulated in the course of everyday life. For this reason, there is almost no ethical rule-of-thumb or principle that cleanly maps onto all forms of online experimentation. Take, for instance, the matter of consent. Since 1947, when the Nuremberg Code formulated "the voluntary, well-informed, understanding consent of the human subject in a full legal capacity," informed consent has been seen as the core tenet of essentially all research ethics frameworks. Yet much research in online environments proceeds without any consent at all or with only some minimal or broadly stated consent. For example, terms of use for many

social media websites typically include explicit or implicit provisions allowing for content posted to the site to be used for research purposes. Twitter has become a dominant data source for the study of social networks precisely because of its permissive data use policies.

A fairly innocuous example of experimentation on Twitter without informed consent is provided by Coppock, Guess, and Ternovski's (2016) experimental use of "direct messages" sent by an environmental organization to users who followed the account (thus allowing for them to receive DMs from the organization). Their experiment randomly assigned users to receive a DM encouraging them to sign a petition and to tweet about the petition, with both measures used as outcomes. Users were already following the organization, the experiment involved activity the organization would have done otherwise, and the outcomes were public behaviors that were already observable.

Yet the form of implied, broad consent that makes this study ethically permissible differs dramatically from the forms of consent normally used in other types of research, like survey-style or semi-structured interviews where consent is explicit and narrow. Particular sites and particular legal frameworks may require forms of consent that exceed the terms of use of a given website, making the possibility of this kind of experiment dependent on the researcher's employment, location, and funding. Regardless, researchers should always consider what role consent does and ought to play in their online research.

The absence of a traditional process of consent in much online experimentation – online survey experiments as an exception – raises particularly important ethical considerations about the use of deception. While research participants entering a laboratory setting and providing explicit consent to participate in a psychological experiment might reasonably expect both the risk of small physical or emotional harm from their participation, they might also expect that the experience will be unusual and even deceptive (see, for a review of relevant research, Hertwig and Ortmann 2012). The risks are explicit. And

expectations about what they will accept are also colored by the context in which the research takes place. Online experiments, to the extent that they occur naturalistically, feel much closer to everyday life and research participants might—rightfully—have more conservative expectations about what will happen to them on a social media site, in a mere survey, or in the course of their browsing a website. While mild forms of deception – such as omissions of full information about the purpose of research or active communication of false information – might be perfectly acceptable in the laboratory where norms are different and where debriefing is guaranteed, naturalistic online environments carry different norms of conduct and different tolerance for harm, privacy, and deception.

Consider, for example, Munger's (2017) experimental study of prejudice reduction. Munger first detected uses of a racial slur on the platform and then used an army of Twitter bots with randomly varied profile characteristics (number of followers and implied race) to send the message: "Hey man, just remember that there are real people who are hurt when you harass them with that kind of language." He then measured subsequent uses of the racial slur. Participants did not opt-in to the study, yet Munger defended the ethics of the research:

> the intervention I applied falls within the "normal expectations" of their user experience on Twitter. The subjects were not debriefed. The benefits to their debriefing would not outweigh the risks to me, the researcher, in providing my personal information to a group of people with a demonstrated propensity for online harassment. (Munger 2017, 636n9)

In this case, where the purpose of the study is arguably positive and the intervention relatively trivial, this "normal expectations" is a potentially credible defense for neither obtaining informed consent nor debriefing participants about the deceptive nature of the experimental intervention into their lives. But a lack of informed consent in tandem with deception poses serious ethical concerns.

The Munger and Coppock et al. studies also highlight that even if the absence of consent and use of deception are problematic, Twitter users receive a degree of privacy above

and beyond what they might otherwise expect from their public posting of content. Twitter terms of use prevent the storage of message content by researchers, enabling these users to remove their data from future research simply by deleting posts. Similarly, legal frameworks in the United States, European Union, and many other places require anonymization of research data in such a way that even datasets constructed from online behavior should not be able to reveal the identities of the participants. Such privacy considerations are particularly important for online research where something as simple as a social media handle, a single-sentence post, or a profile photo may be enough to not only link an anonymized dataset to its source user but further link that user to other websites and datasets.

Researchers using online data will almost necessarily find themselves in a situation one-step away from a catastrophic risk to the privacy of participants, who may not have even explicitly consented to participating their research. Intentional data dumps like the release of Yahoo! individual-level, "anonymized" search histories or Netflix's release of subscribers' "anonymized" film ratings reveal the ease with which online data are re-identifiable. Web companies may impose terms of use on data meant to limit the risk of these kinds of re-identification or data reuse. But terms of use are a legal and thus imperfect protection.[2] And terms of use and website functionality can shift dramatically over time. Facebook, for example, once allowed developers (and researchers) access to the data of not only users who consented to use of an app, but also some data on those users' friends. Privacy risks presented by this situation led to changes in functionality and terms of use. Researchers working with online data must therefore be cognizant of a variety of restrictions on data use, including shifting contractual agreements and data handling norms that fall somewhere between strict legal protections and broader ethical frameworks. Negative public reception to high-profile

[2] Terms of use can also be surprising. Some websites restrict researchers even from using public content either by prohibiting scraping of content or prohibiting the storage thereof. The ability to gather data from a site because it is "public" does not per se give researchers the permission of the site or its users to use it for research.

instances of the now-common practice of website experimentation (such as Kramer et al. [2014]), highlight just how perilous public trust in science has become in the digital age.

One final ethical issue posed by online experimentation relates to the impact of researcher interventions that extend beyond the scope of the study itself. Online experiments sometimes involve interventions that extend over potentially long periods of time and with wide-reaching impacts. Any time a researcher intervenes in individuals' lives or in the social and political world, consideration must be given to whether that intervention has intended or unintended consequences. "Action research" intended to improve perceived deficiencies in the real world can be justifiable but is not value-neutral. What one researcher sees as an acceptable or even desirable impact of—or externality of—a research project may not be seen in the same light by other researchers, the media, or the mass public. Unintended impacts of research projects that impact upon election outcomes, the economic well-being of non-participants, or the business activities of firms are just some of the ways that online experimentation might impact reality in ways susceptible to intense scrutiny. Whether these kinds of externalities are ethically permissible remains an open question. Researchers should be cognizant of externalities and give consideration to how those impacts might be perceived by others.

Opportunities

The online world offers many exciting possibilities that should not be downplayed amidst a long discussion of difficulties involved in studying it. It can be a convenient environment for studying offline political behavior because of the ease of recruiting experimental participants from crowdsourcing platforms or online survey panels. It is also a venue for obtaining novel descriptive data about media content, political campaigning, and public opinion. And the online world is a political entity in its own right to be investigated. Striking a balance between the hype of online methods and the realistic challenges of

leveraging digital environments for effective causal inference is therefore critical. As I have argued so far, online experiments are an evolution, not a revolution, and need to be handled with the same care as offline experimentation.

Yet this evolution carries with in valuable new opportunities not previously available that are worth highlighting. Digital environments provide ways administering stimuli and measuring outcomes that were previously unimaginable. Smartphones provide an omnipresent access point to the world wide web, the ability to widely deploy research applications, the ability to engage SMS to provide brief, immediate contact with large numbers of potential participants. Leveraging these new technologies requires new technical skills – like web scraping, app development, and website design – that budding social scientists may wish to develop to take full advantage of the opportunities afforded by digital platforms.

Yet it is not only technical skills that may be required to take full advantage of online research environments. Many of the most interesting sites of political activity online take place in propriety environments where user data are not public such that access to those data is inequitable. News websites typically do not publish usage data, search engines provide only minimal topline information (see Stephens-Davidowitz 2017). And social media platforms vary radically in the degree to which they allow research access to user data. While major internet firms all employ sizable in-house teams of researchers and regularly partner with academic researchers, the information cultivated by these firms and the ability to experiment with their online environments is not natively available to all researchers. Building connections with these firms and developing new ways for researchers to access online platforms is a matter for the next generation of social scientists to tackle

Amidst this changing landscape, there are numerous promising avenues for research using online experimentation. A striking feature of extant work is that it typically generates context-specific answers that provide locally useful information about a particular online environment at a particular point in time. Methodological advancement in this area could develop new methods for evaluating sample representativeness, data quality, and other sample characteristics on-the-fly rather than requiring extensive research into particular platforms before they become seen as acceptable or unacceptable for particular research applications. Beyond sample-related considerations, other fruitful paths for methodological innovation involve the development of easily deployed software that can orchestrate complex experimental designs with block randomization, adaptive treatment allocation, or high dimensionality (e.g., such as conjoint designs). While these methods already exist (e.g., Moore 2013), the lack of accessible tooling to implement them limits their application.

A particularly important but almost entirely unexamined methodological aspect of online experiments is the degree of variation in how stimuli can be rendered across devices, operating systems, and web browsers. Users participating in an online experiment may complete that survey on a desktop computer, on a laptop, on a modestly sized tablet, or a small smartphone. The variation in the experience of stimuli – particularly audio and visual stimuli – across these devices introduces a substantial loss of control into the experimental environment that might affect compliance, engagement with stimuli, reading and response times, attrition, and so forth (de Bruijne and Wijnant 2014; Couper and Peterson 2016; Tourangeau et al. 2017). Social media users might similarly experience stimuli through desktop interfaces, smartphone applications, or mobile browsers. We know almost nothing about how these features of online technologies might impact experimental participants.

More substantively, online experiments are and will continue to be a powerful tool for studying offline and online political phenomena. Between-country variation in internet use

and the prevalence of online research platforms necessarily limits the degree to which online experimentation has generated knowledge about other country contexts. More research is needed to understand whether any phenomena investigated on conveniently studied online populations applies to online or offline populations in other contexts. Ambitious projects in low internet penetration contexts have tried to study the impact of the internet access on political engagement using randomized access to internet cafes (Bailard 2012) or by relying on web-like technologies (such as SMS) to mimic the research environment of an online experiment (Buntaine et al. 2018). More work is needed to deploy online experiments in novel contexts and, similarly, on how to understand political phenomena that take places across contexts with widely varied internet and social media usage. For example, while Facebook is massively diffuse in the United States and many other countries at the time of this writing, large and politically interesting countries including Russia and China have essentially no use of the platform with VKontakte and WeChat, respectively, dominating instead. These cross-national differences facilitate country-specific research questions like how the Chinese government censors social media content (King, Pan, and Roberts 2014) but hinder more comparative questions. Like cross-country comparisons, experiments into the impact of a platform or use of a platform by political actors are also restricted to the *temporal* context in which those experiments were performed and the features and populations of the online environments involved. Digital researchers need to consider how if at all to learn about such rapidly changing environments.

Conclusion

Online experiments are here to stay. Be they survey-, field-, or laboratory-like experiments performed in an online mode, the method has shown promise for providing uniquely useful insights into political phenomena. This chapter has described the logic of experimental methods, discussed trade-offs involved in online experimentation compared to other methods,

demonstrated numerous applications, highlighted a number of challenges posed by online experiments, and suggested several avenues for future research. The entirety of this discussion is informed—and limited—by how the internet exists in 2018 and the set of technologies available in this moment. The internet will no doubt look very different in the near future, with new technologies, new patterns of internet usage, new websites and platforms, and new questions posed by those evolutions. Researchers should be prepared to be unprepared for what lies ahead.

References

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

Aronow, Peter M., and Cyrus Samii. 2015. "Does Regression Produce Representative Estimates of Causal Effects?" *American Journal of Political Science* . Unpublished paper.

Bailard, Catie Snow. 2012. "A Field Experiment on the Internet's Effect in an African Election: Savvier Citizens, Disaffected Voters, or Both?" *Journal of Communication* 62(2): 330–344.

Baker, Reg, Stephen J. Blumberg, J. Michael Brick, Mick P. Couper, Melanie Courtright, J. Michael Dennis, Don A. Dillman, Martin R. Frankel, Philip Garland, Robert M. Groves, Courtney Kennedy, Jon A. Krosnick, Paul J. Lavrakas, Sunghee Lee, Michael Link, Linda Piekarski, Kumar Rao, Randall K. Thomas, and Dan Zahs. 2010. "Research Synthesis: AAPOR Report on Online Panels." *Public Opinion Quarterly* 74(4): 711–781.

Bakhshandeh, Reza, Mehdi Samadi, Zohreh Azimifar, and Jonathan Schaeffer. 2011. "Degrees of Separation in Social Networks." *Proceedings, The Fourth International Symposium on Combinatorial Search* .

Barbera, Pablo, and Gonzalo Rivero. 2014. "Understanding the Political Representativeness of Twitter Users." *Social Science Computer Review* , In press.

Barrera, Davide, and Brent Simpson. 2012. "Much Ado About Deception: Consequences of Deceiving Research Participants in the Social Sciences." *Sociological Methods & Research* 41(3): 383–413.

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): 351–368.

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(7415): 295–298.

Bowers, J., M. M. Fredrickson, and C. Panagopoulos. 2013. "Reasoning about Interference Between Units: A General Framework." *Political Analysis* 21(1): 97–124.

Broockman, David E., and Donald P. Green. 2013. "Do Online Advertisements Increase Political Candidates' Name Recognition or Favorability? Evidence from Randomized Field Experiments." *Political Behavior* 36(2): 263–289.

Broockman, David E., Joshua L. Kalla, and Jasjeet S. Sekhon. 2017. "The Design of Field Experiments With Survey Outcomes: A Framework for Selecting More Efficient, Robust, and Ethical Designs." *Political Analysis* .

Buntaine, Mark T., Sarah Bush, Ryan Jablonski, Daniel Nielson, and Paula Pickering. N.d. "Budgets, SMS Texts, and Votes in Uganda.". In Information and Accountability: A New Method for Cumulative Learning. Forthcoming, Cambridge University Press.

Callegaro, Mario, Reg Baker, Jelke Bethlehem, Anja S. Göritz, Jon A. Krosnick, and Paul J. Lavrakas, eds. 2014. *Online Panel Research: A Data Quality Perspective*. West Sussex, UK: Wiley.

Cartwright, Nancy. 2007. *Hunting Causes and Using Them*. Cambridge University Press.

Clifford, Scott, and Jennifer Jerit. 2015. "Do Attempts to Improve Respondent Attention Increase Social Desirability Bias?" *Public Opinion Quarterly* 79(3): 790–802.

Clifford, Scott, Ryan M Jewel, and Philip D. Waggoner. 2015. "Are samples drawn from Mechanical Turk valid for research on political ideology?" *Research and Politics*, 1–9.

Coppock, Alexander. N.d. "Generalizing from Survey Experiments Conducted on Mechanical Turk: A Replication Approach." *Political Science Research & Methods: In press*.

Coppock, Alexander, Andrew Guess, and John Ternovski. 2016. "When Treatments are Tweets: A Network Mobilization Experiment over Twitter." *Political Behavior* 38(1): 105–128.

Coppock, Alexander, Thomas J. Leeper, and Kevin J. Mullinix. N.d. "The Generalizability of Heterogeneous Treatment Effect Estimates Across Samples." Unpublished paper, Unpublished paper, Yale University.

Couper, Mick P., and Gregg J. Peterson. 2016. "Why Do Web Surveys Take Longer on Smartphones?" *Social Science Computer Review* 35(3): 357–377.

de Bruijne, Marika, and Arnaud Wijnant. 2014. "Mobile Response in Web Panels." *Social Science Computer Review:* In press.

Desposato, Scott, ed. 2015. *Ethics and Experiments*. Taylor & Francis.

Dickson, Eric S. 2011. “Economics versus Psychology Experiments.” In *Cambridge Handbook of Experimental Political Science*, eds. James N. Druckman, Donald P. Green, James H. Kuklinski, and Arthur Lupia. New York: Cambridge University Press.

Druckman, James N., and Cindy D. Kam. 2011. “Students as Experimental Participants: A Defense of the ‘Narrow Data Base’.” In *Cambridge Handbook of Experimental Political Science*, eds. James N. Druckman, Donald P. Green, James H. Kuklinski, and Arthur Lupia. New York: Cambridge University Press, 41–57.

Druckman, James N., and Thomas J. Leeper. 2012. “Learning More from Political Communication Experiments: Pretreatment and Its Effects.” *American Journal of Political Science* 56(4): 875–896.

Egami, Naoki, Christian J. Fong, Justin Grimmer, Margaret E. Roberts, and Brandon M. Stewart. N.d. “How to Make Causal Inferences Using Text.” Unpublished paper, Unpublished paper, Princeton University.

Esterling, Kevin M., Michael A. Neblo, and David M.J. Lazer. 2008. “Means, Motive, and Opportunity in Becoming Informed about Politics: A Deliberative Field Experiment with Members of Congress and Their Constituents.” *Public Opinion Quarterly* 75(3): 483–503.

Feezell, Jessica T. 2017. “Agenda Setting through Social Media: The Importance of Incidental News Exposure and Social Filtering in the Digital Era.” *Political Research Quarterly* , 106591291774489.

Fowler, James H., Michael T. Heaney, David W. Nickerson, John F. Padgett, and Betsy Sinclair. 2011. “Causality in Political Networks.” *American Politics Research* 39(2): 437–480.

Gerber, Alan S., and Donald P. Green. 2012. *Field Experiments: Design, Analysis, and Interpretation*. W. W. Norton & Company.

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(1): 33–48.

Gerber, Alan S., Kevin Arceneaux, Cheryl Boudreau, Conor M. Dowling, D. Sunshine Hillygus, Thomas R. 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.

Goodman, Joseph K., and Gabriele Paolacci. 2017. “Crowdsourcing Consumer Research.” *Journal of Consumer Research* 44(1): 196–210.

Groves, Robert M., Floyd J. Fowler, Mick P. Couper, James M. Lepkowski, Eleanor Singer, and Roger Tourangeau. 2009. *Survey Methodology*. Second ed. Wiley-Interscience.

Hertwig, Ralph, and Andreas Ortmann. 2008. “Deception in Experiments: Revisiting the Arguments in Its Defense.” *Ethics & Behavior* 18(1): 59–92.

Holland, Paul W. 1986. “Statistics and Causal Inference.” *Journal of the American Statistical Association* 81(396): 945–960.

Jones, Jason J., Robert M. Bond, Eytan Bakshy, 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): e0173851.

Jungherr, Andreas. 2016. “Twitter use in election campaigns: A systematic literature review.” *Journal of Information Technology & Politics* 13(1): 72–91.

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(4): 415–440.

King, Gary, Jennifer Pan, and Margaret E. Roberts. 2014. “Reverse-engineering censorship in China: Randomized experimentation and participant observation.” *Science* 345(6199): 1251722–1251722.

King, Rochelle, Elizabeth Churchill, and Caitlin Tan. 2017. *Designing with Data*. O’Reilly UK Ltd.

Kramer, Adam DI, 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* , 201320040.

Lau, Richard R., and David P. Redlawsk. 2006. *How Voters Decide: Information Processing in Election Campaigns*. New York: Cambridge University Press.

Leeper, Thomas J. 2013. “Crowdsourcing with R and the MTurk API.” *The Political Methodologist* 20(2): 2–7.

Leeper, Thomas J. 2014. “The Informational Basis for Mass Polarization.” *Public Opinion Quarterly* 78(1): 27–46.

Mason, Winter, and Siddharth Suri. 2011. “Conducting Behavioral Research on Amazon’s Mechanical Turk.” Unpublished paper, Yahoo! Research Unpublished paper. http://www.ncbi.nlm.nih.gov/pubmed/21717266

McDermott, Rose. 2002. “Experimental Methods in Political Science.” *Annual Review of Political Science* 5(1): 31–61.

McDermott, Rose. 2011. “Internal and External Validity.” In *Cambridge Handbook of Experimental Political Science*, eds. James N. Druckman, Donald P. Green, James H. Kuklinski, and Arthur Lupia. New York: Cambridge University Press.

Messing, Solomon, and Sean J. Westood. 2013. “Friends that Matter: How Social Influence Affects Selection in Social Media.” Unpublished paper, Stanford University.

Messing, Solomon, and Sean J. Westwood. 2014. “Selective Exposure in the Age of Social Media: Endorsements Trump Partisan Source Affiliation When Selecting News Online.” *Communication Research* 41(8): 1042–1063.

Montgomery, Jacob M., Brendan Nyhan, and Michelle Torres. 2018. "How Conditioning on Posttreatment Variables Can Ruin Your Experiment and What to Do about It." *American Journal of Political Science* 62(3): 760–775.

Moore, Ryan T. 2013. "Blocking for Sequential Political Experiments." *Political Analysis* 21(4): 507–523.

Morton, Rebecca B., and Kenneth C. Williams. 2010. *Experimental Political Science and the Study of Causality: From Nature to the Lab*. Cambridge University Press.

Mullinix, Kevin J., Thomas J. Leeper, James N. Druckman, and Jeremy Freese. 2015. "The Generalizability of Survey Experiments." *Journal of Experimental Political Science* 2: 109–138.

Munger, Kevin. 2017. "Tweetment Effects on the Tweeted: Experimentally Reducing Racist Harassment." *Political Behavior* 39(3): 629–649.

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.

Rogowski, Jon C., and Betsy Sinclair. 2012. "Estimating the Causal Effects of Social Interaction with Endogenous Networks." *Political Analysis* 20(03): 316–328.

Rubin, Donald B. 2008. "For Objective Causal Inference, Design Trumps Analysis." *The Annals of Applied Statistics* 2(3): 808–840.

Salganik, Matthew J. 2017. *Bit by Bit*. Princeton University Press.

Shadish, William R., Thomas D. Cook, and Donald T. Campbell. 2001. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston, MA: Houghton-Mifflin.

Sinclair, Betsy, Margaret McConnell, and Donald P. Green. 2012. "Detecting Spillover Effects: Design and Analysis of Multilevel Experiments." *American Journal of Political Science* 00(0): no–no.

Sniderman, Paul M. 2011. "The Logic and Design of the Survey Experiment: An Autobiography of a Methodological Innovation." In *Cambridge Handbook of Experimental Political Science*, eds. James N. Druckman, Donald P. Green, James H. Kuklinski, and Arthur Lupia. New York: Cambridge University Press.

Stephens-Davidowitz, Seth. 2017. *Everybody Lies: Big Data, New Data, and What the Internet Can Tell Us About Who We Really Are*. Dey Street Books.

Stewart, Neil, Christoph Ungemach, Adam J. L. Harris, Daniel M. Bartels, Ben R. Newel, Gabriele Paolacci, and Jesse Chandler. 2015. "The Average Laboratory Samples a Population of 7,300 Amazon Mechanical Turk Workers." *Judgment and Decision Making* 10(3): 479–491.

Taylor, Sean J., and Dean Eckles. 2017. “Randomized experiments to detect and estimate social influence in networks.” *Arxiv* .

Theocharis, Yannis, and Will Lowe. 2015. “Does Facebook increase political participation? Evidence from a field experiment.” *Information, Communication & Society* 19(10): 1465–1486.

Titiunik, Rocío. 2015. “Can Big Data Solve the Fundamental Problem of Causal Inference?” *PS: Political Science & Politics* 48(01): 75–79.

Tourangeau, Roger, Hanyu Sun, Ting Yan, Aaron Maitland, Gonzalo Rivero, and Douglas Williams. 2017. “Web Surveys by Smartphones and Tablets.” *Social Science Computer Review*.

Ugander, Johan, Brian Karrer, Lars Backstrom, and Cameron Marlow. 2011. “The Anatomy of the Facebook Social Graph.” *ArXiV* .

Utych, Strephen M., and Cindy D. Kam. 2014. “Viability, Information Seeking, and Vote Choice.” *The Journal of Politics* 76(1): 152–166.

Valenzuela, Sebastián. 2013. “Unpacking the Use of Social Media for Protest Behavior.” *American Behavioral Scientist* 57(7): 920–942.

Villarroel, Maria A., Charles F. Turner, Elizabeth Eggleston, Alia Al-Tayyib, Susan M. Rogers, M. Roman, Anthony, Philip C. Cooley, and Harper Gordek. 2006. “Same-Gender Sex in the United States: Impact of T-ACASI on Prevalence Estimates.” *Public Opinion Quarterly* 70(2): 166–196.