



Contents lists available at ScienceDirect

Social Science Research

journal homepage: www.elsevier.com/locate/ssresearch

The past, present, and future of experimental methods in the social sciences

Trenton D. Mize^{a,*}, Bianca Manago^b^a Departments of Sociology & Advanced Methodologies, Purdue University, USA^b Department of Sociology, Vanderbilt University, USA

ARTICLE INFO

Keywords:

Experiments

Causality

Random assignment

Sampling

ABSTRACT

In the midst of the current causal revolution, experimental methods are increasingly embraced across the social sciences. We first document the growth in the use of the experimental method and then overview the current state of the field along with suggestions for future research. Our review covers the core features of experiments that facilitate causal inference, and we offer numerous examples of different experimental designs and the types of research questions they are well-suited to test. We then harmonize popular but differing perspectives on validity in regard to experiments focused on theoretical considerations. Next, we detail a new framework of purposive sampling for social science experiments, suggesting researchers could benefit from moving beyond only two stock choices of representativeness or convenience. We then discuss issues of analysis unique to experimental data before ending with a critical discussion of null findings, open-science practices, social desirability, and other open debates in the experimental methods literature.

Experimental methods have a long history in the social sciences as they provide great utility for building and testing causal theories, for identifying the causal effects of interventions, and for determining mechanisms underlying causal effects. For at least the last 60 years, laboratory experiments have been key to the development of many core theories in sociology, economics, political science, and related fields (Falk and Heckman 2009; Gärxhani and Miller 2022; Webster and Jane Sell, 2007). Increasingly, social science experiments are also being conducted outside the lab in surveys, online environments, field settings, and other innovative platforms in order to develop theories and to test the generalizability of experimental findings to broad populations of interest (Baldassarri and Abascal 2017; Druckman and Green 2021b; Gaddis 2018; Mutz 2011).

The experimental method is not a panacea—but there is a reason it is common across most scientific fields. When determinations of causality are necessary and experiments are feasible, they are an invaluable tool. For example, medical trials usually rely on experiments (AKA randomized controlled trials) to determine whether a new vaccine, medication, or procedure is effective (for a critical discussion, see Deaton and Cartwright 2018). In business, experiments (AKA A/B tests) are often used to determine the most effective form of a product design, advertising effort, website layout, or any other miscellaneous choice (Gallo 2017). Within the social sciences, the breadth of topics studied and the types of experimental designs utilized has expanded rapidly (Druckman and Green 2021b; Gärxhani and Miller 2022). When an experimental design is feasible and ethical, there is rarely a more elegant and convincing method for causal inference.

In this article, we first overview the impressive growth in the use of the experimental method across the social sciences. Second, we

* Corresponding author. 700 W State St West Lafayette, IN 47907

E-mail address: tmize@purdue.edu (T.D. Mize).

detail key features of experimental design, citing myriad examples of effective and creative experimental designs currently used throughout the social sciences. Next, we discuss validity and the various purposes experiments can serve integrating multiple influential but sometimes antagonistic views of validity and the experimental method. We then propose a new framework of purposive sampling with an eye towards encouraging sampling choices that best match specific research questions and theoretical goals. Finally, we discuss issues of analysis unique to experimental data and then end with some challenging questions for experimentalists and avenues for continued research on the method. Throughout, we provide suggestions for how the experimental method should grow in the future. In so doing, we advocate for the wider adoption of experimental methods into social science research and provide guidance about best practices.

1. The rise of experiments in the social sciences

1.1. Tracking the increasing presence of experiments in top social science journals

The experimental method has long been part of social scientists' methodological toolkit (for a history of the experimental method in various contexts, see Druckman et al., 2006; Salsburg 2001; Thye 2014a). Despite experiments' long history in social science, for years they remained mostly a niche methodology in the fields of sociology, political science, and economics (Druckman et al., 2006; Druckman and Green 2021a; Gërxhani and Miller 2022; Jackson and Cox 2013). For multiple reasons, likely including a renewed focus on causality across the social sciences (Imbens and Rubin 2015; Morgan and Winship 2014; Pearl 2009), use of the experimental method has increased over the last 30 years. In particular, there was substantial growth in the use of the experimental method during the 2000s in political science and economics (Druckman and Green 2021a; Gereke and Gërxhani 2019; Jackson and Cox 2013).

To determine current trends in the use of the experimental method across the social sciences we tracked the number of journal publications using experimental methodology across sociology, political science, and economics from the years 1990–2020. As there are notable subdiscipline differences in the use of experiments (e.g., social psychology within sociology has traditionally used experiments more heavily than other subdisciplines; (Webster and Jane Sell, 2007)), we focus on top generalist journals for each field.¹ For sociology, we examined publications in the *American Sociological Review* (ASR) and the *American Journal of Sociology* (AJS). For political science, we examined the *American Political Science Review* (APSR) and the *American Journal of Political Science* (AJPS). For economics, we examined the *American Economic Review* (AER) and *Econometrica* (Econ). Collectively, these represent some of the most highly visible and oft-cited journals in the social sciences, with 2-year impact factors ranging from 4.503 (AJPS) to 12.444 (ASR).² In Fig. 1, we plot both the proportion of publications using experiments each year (as hollow dots) and smoothed trend lines tracking the proportion over time (solid lines; Cleveland 1994).³

The first trend to emphasize from Fig. 1 is that all three disciplines are utilizing the experimental method more today than they were 30 years ago. On average across the three disciplines, roughly four percent of articles published in the early 1990s used experimental methods; in contrast, over fourteen percent of published articles used experiments in the late 2010s, more than a threefold increase. However, there are distinct trends among the disciplines. Political science shows the most dramatic increase in the use of experiments, with a smoothed average increasing from a little over three percent to roughly 24 percent in the present day. Economics has also seen a

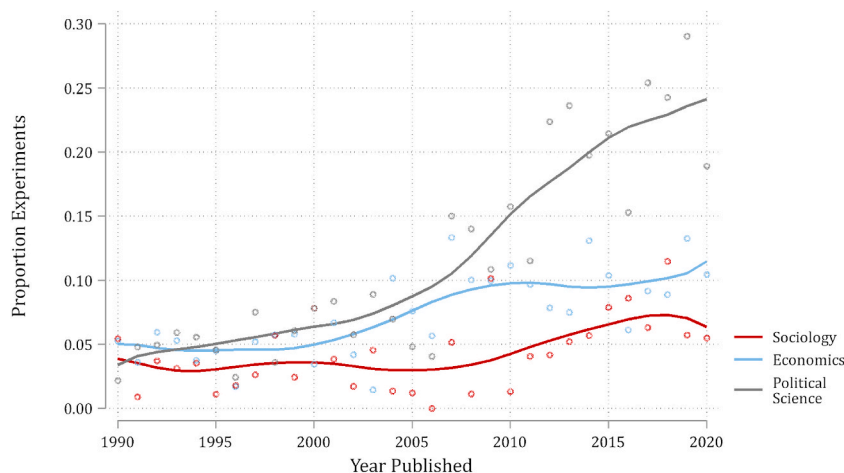


Fig. 1. Proportion of published articles using experiments in top two generalist sociology, economics, and political science journals (1990–2020).

¹ See Appendix A for details on our methodology.

² Based on InCites Journal Citation Reports' 2021 journal impact factor.

³ The data and coding files needed to reproduce Figs. 1–3 are available at <https://www.trentonmize.com/research>.

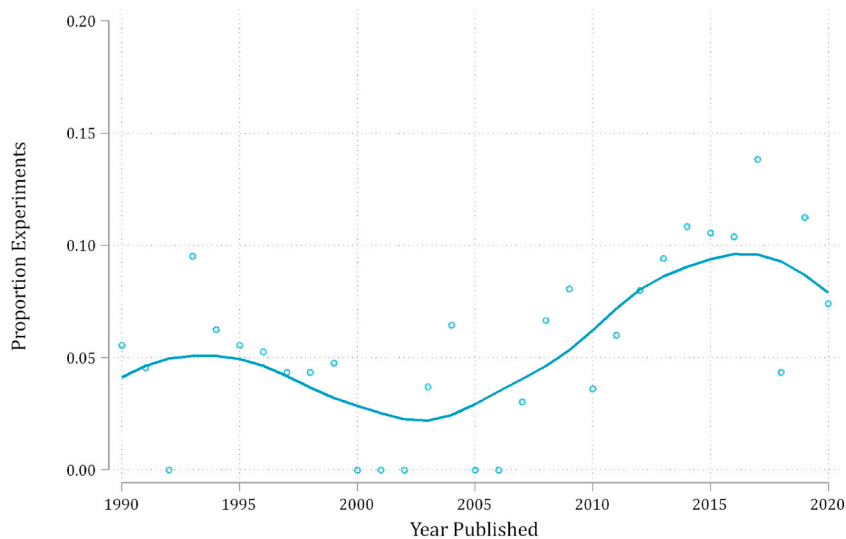


Fig. 2. Proportion of published articles using experiments in *Social Science Research* (1990–2020).

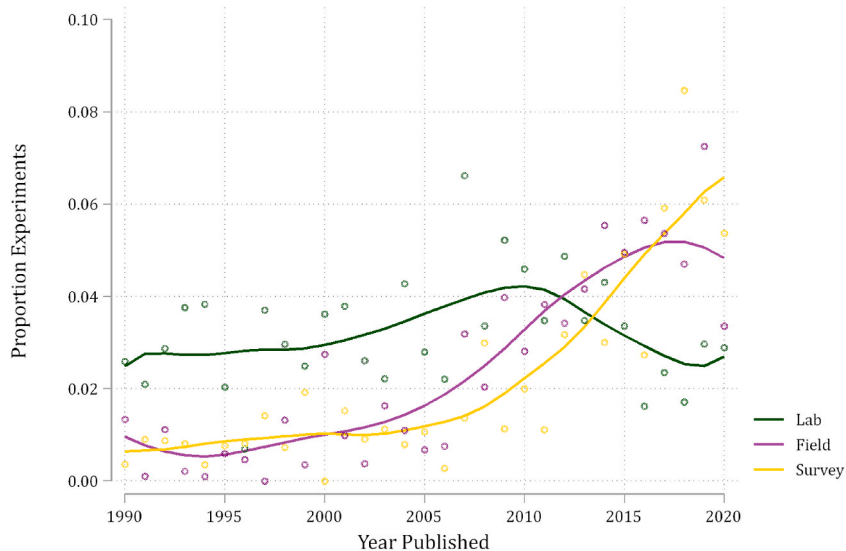


Fig. 3. Proportion of published articles using experiments across sociology, political science, and economics by type of experimental method used (1990–2020).

marked increase, from about five percent to nearly twelve percent today. Finally, sociology has more slowly begun adopting experiments. Until about 2006, between three to four percent of sociological publications used experimental methods; this number has doubled to approximately seven percent today.

Given the growth across sociology, economics, and political science, we would expect to see similar trends in the interdisciplinary *Social Science Research*. Indeed, as illustrated in Fig. 2, the proportion of experiments has doubled from roughly five percent to almost ten percent of *Social Science Research* articles over the last thirty years.

What explains the substantial growth documented above? The increase in the representation of experiments in the three fields is due to the embrace of a wider range of experimental methodologies than only the traditional behavioral laboratory experiment. Fig. 3 separates out experimental publications in the top social science journals by type: behavioral lab study, survey experiment, or field experiment (including audit studies). The number of laboratory experiments has held steady while representation of survey and field experiments has increased from almost zero to now representing around seven and five percent of all publications, respectively, in these top journals.

Next, we discuss core features of the experimental method and their benefits, which likely explain the growth documented above.

2. Core features of experiments

Experiments are a powerful methodological tool for testing hypotheses and demonstrating causality. Despite considerable advances in statistical methods for causal inference in observational data, experiments remain unrivaled as a “gold standard” of causal inference (Mutz 2011; Webster and Jane Sell, 2007). As Mutz (2011) succinctly states:

“For many research questions, experiments are simply the most effective means of evaluating competing causal hypotheses. As much as we would love it if it were true, there simply are no statistical techniques for observational data that provide the power and elegance of an experimental design.” (p.14).

We believe the rise in the use of experiments documented in the last section is due not only to an increased appreciation of this point, but also due to the embrace of multiple types of experiments and innovation in applying experimental methods to topics once thought not amenable for study experimentally. Despite the breadth of different types of experiments and topics studied, there are two key features of all experiments that facilitate the testing of causal processes: manipulation of an independent variable and random assignment.⁴ Importantly, both are needed for a study to be considered an experiment and to establish causality.

2.1. Manipulation of an independent variable

By manipulating the independent variable (IV) in a research study, researchers have taken the first step towards establishing causality. Specifically, by manipulating a specific IV, researchers control the temporal order of the cause relative to effect and control the selection process by which an observation comes to have a given value of the IV. By manipulating the cause, researchers can determine that the outcome is caused by the manipulation as opposed to something that often co-occurs with that change.

For example, consider the hypothesis that people who attend elite universities are more likely to be hired, even when their qualifications are otherwise identical to someone from a non-elite college (Gaddis 2015). It may be difficult to test this hypothesis in observational data, as individuals from elite universities may also have broader social networks and may have more knowledge about how to market their skills. Using an experiment, a researcher could provide one of two resumes to research participants. The resumes would be identical, except for the IV, i.e., the name of the university that the job candidate attended (elite vs. non-elite). Researchers would then ask participants their likelihood of hiring the job candidate. By only changing the name of the university and keeping all else equal, researchers could account for confounding variables and temporality. This control allows researchers to be more confident that any effect was attributable to the university’s elite status.

2.1.1. Categorical and continuous manipulation of an independent variable

In experimental studies, the independent variable (which is manipulated) can be either categorical or continuous. Most often, researchers manipulate experimental variables categorically (e.g., a person is either in the experimental or control group, elite or non-elite university, etc.). Sometimes the experimental IV is fundamentally a categorical concept—e.g., how specific degrees influence perceptions of competence (e.g., high school, Bachelors, Masters, etc.). In contrast, if the theory suggests a continuous independent variable, researchers have multiple choices of how to manipulate the IV.

As an example, consider an experiment designed to test individuals’ likelihood of being hired based on their college grade point average (GPA). Although GPA is a continuous variable, a researcher may wish to choose *ideal types*—i.e., values of the continuous variable that match useful theoretical abstractions (Long and Freese 2014; Weber 1922). E.g., a “C student” or a “B student” are ideal types that could be tested by the researcher by randomly assigning respondents to either read about a person with a 2.5 or a 3.5 GPA and compare rates of hiring. If, however, the theory is that the higher the GPA, the more likely an individual is to be hired, then a continuous manipulation of the IV may be more appropriate. For example, to continuously manipulate GPA, the researchers may assign respondents to read about a person with a GPA ranging uniformly between 1.0 and 4.0 (e.g., Quadlin 2018). The researchers would then analyze the IV as continuous. An advantage of continuous manipulations of the IV is their ability to reveal nonlinear effects. For example, Kincaid (2022) manipulated age as a continuous variable ranging from 18 to 90 and found highly nonlinear effects of age on perceived status. Men’s status is highest in middle-age but lower at young and old ages while women’s status peaks around age 40, declines, but then begins increasing again at old ages. Such complex nonlinearities would be difficult to assess without the continuous manipulation of age.

2.1.2. Manipulating immutable characteristics

Many social science research questions involve immutable—or at least hard to change—characteristics. E.g., someone’s gender, race-ethnicity, childhood SES, or country of birth are factors that do not usually change over the adult life course.⁵ This fact poses a challenge for the influential line of causal inference work that suggests “no causation without manipulation” (Holland 1986:959). How, then, can we assess the causal effects of things like a person’s race-ethnicity? Some in the causal inference literature have come to

⁴ Control is also an important feature of experiments – see Section 3.3 below.

⁵ Exceptions exist, of course. For example, someone can transition to a different gender identity or someone’s racial-ethnic identity (or perceptions of it) may change as other aspects of their life change (Freeman et al., 2011; Saperstein and Penner 2014). Our examples address classic causal inference questions like “are cisgender women treated differently than cisgender men because of their gender?” rather than a question such as “does transitioning to a different gender identity change how someone is treated”—though we note experiments could be used to address both questions.

view these effects in the manner social science experimentalists have long viewed them: while we cannot manipulate things like a person's race-ethnicity, we can manipulate the perception of someone else's race-ethnicity and observe the causal effects of those perceptions (Greiner and Rubin 2011; Morgan and Winship 2014).

Experiments usually do not manipulate the characteristics of participants but instead the levels of the independent variable—which can be immutable human characteristics of someone being evaluated.⁶ For example, the names on resumes are often used to signal the gender or race-ethnicity of a job applicant (Gaddis 2018).⁷ The difference in evaluations of each resume is the effect of an applicant's race-ethnicity. Thus, experiments can be used to answer causal questions about immutable characteristics with a shift in perspective to how perceptions of these characteristics change attitudes, opinions, and behavior towards individuals who possess certain immutable characteristics.

2.1.3. Studying immutable characteristics through abstraction

Another route for studying immutable characteristics is to abstract their theoretically important features and create an experimental test of the theoretical constructs themselves. For example, status construction theory seeks to understand how distinctions among individuals—e.g., their gender identities or race-ethnicities—come to carry status value and affect how people are treated in interactions (Ridgeway 2019; Ridgeway and Correll 2006). But constructs like gender and race-ethnicity contain many connotations other than just status. To test the theoretical model of status construction, Ridgeway and colleagues (e.g., 2006) create status characteristics in the lab: categorical distinctions among individuals, which are then associated with differential status value. This line of research has provided a pure test of how status characteristics become associated with differential status value and then affect subsequent interactions (see Ridgeway 2019 for an overview). While the status characteristics tested in the lab are abstract, they inform our understanding of how real status characteristics like gender and race likely obtained their current status value and how they influence interaction.⁸

2.2. Random assignment

The second core feature of an experiment is random assignment. By random assignment, we are referring to the probability that participants are exposed to one manipulation vs. another; usually, each participant has an equal probability of being assigned to each level of the IV (i.e., each experimental condition). As an example, consider a study designed to examine the effect of caffeine on happiness. The researchers decide to use two conditions: one in which the participants must drink two cups of coffee each morning, and another in which they are forbidden from drinking any coffee at all. The researchers recruit a sample, represented as the dots within the large circle in Fig. 4. Each dot in the circle represents a participant—with the different colors representing different participant characteristics (e.g., different educational degrees).

If the researchers let participants choose their condition—as the middle panel of Fig. 4 shows, there might be some confounding factors. For example, imagine that people who choose to drink coffee are, for some reason, also more anxious. If the colors of the dots

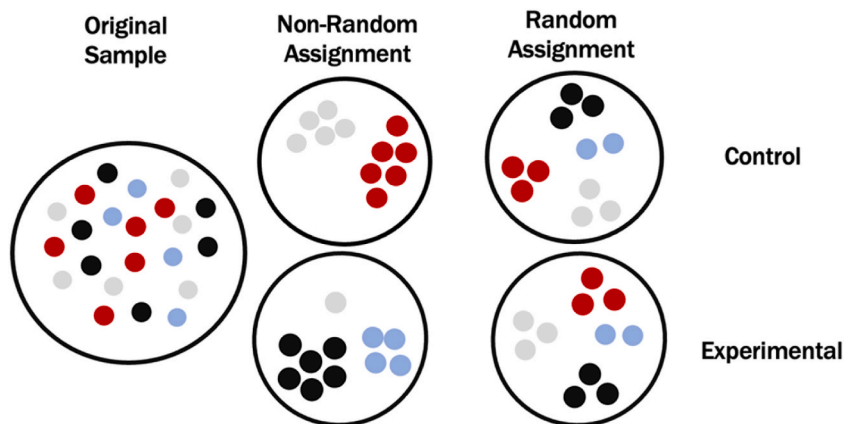


Fig. 4. Illustration of random assignment's ability to balance participant characteristics across experimental conditions.

⁶ Experiments can also be used to manipulate self-perceptions. E.g. Willer et al. (2013) manipulated feedback about participant's masculinity and femininity and observed their reactions.

⁷ But note recent work cautioning that names alone are unreliable signals of race-ethnicity and are often confounded with social class connotations (Gaddis 2017).

⁸ For another example, see Manago and Mize (2022) who create fictitious mental illness labels that while artificial, retain the key theoretical features of different categories of mental illness.

represent anxiety levels and participants are allowed to choose their experimental condition, anxiety would be confounded with values of the IV. As a result, any effects on happiness are not only attributable to coffee allocation as, e.g., those with high anxiety are overrepresented in the high coffee (experimental) condition (and this anxiety may affect their happiness). In contrast, if the participants are randomly assigned to conditions as the right panel of Fig. 2 shows, then any effect of coffee on happiness is caused by the experimental manipulation because participant characteristics are balanced across conditions—as illustrated by an equal number of light grey, blue, red, and black dots in each condition.

Random assignment is also useful for accounting for other unmeasured and/or yet undiscovered factors relevant to the research question. For example, consider again the coffee experiment. If there is a (currently undiscovered) gene that makes individuals particularly sensitive to caffeine, random assignment will control for this factor by evenly balancing the number of those with and without the gene in each condition. Random assignment randomly distributes an infinite number of differences, providing confidence that a difference in the dependent variable is attributable to the IV that has been experimentally manipulated.

Random assignment is essential for experimental design because it controls for all aspects of a participant—including those that are measured, measurable but not measured, and unmeasurable. In observational data, researchers can use statistical procedures such as controlling for confounding factors or matching subjects on characteristics that are known to be theoretically relevant (Morgan and Winship 2014). However, such analyses are limited by what was measured and there are, at times, characteristics that are not yet known to affect the relationship between an independent and dependent variable (for example, with the undiscovered coffee gene mentioned above). In addition, the correct statistical model must be chosen and specified to effectively "control" for the confounding factors—a difficult task. In contrast, with random assignment and experimental manipulation of an IV researchers can be confident that the independent (manipulated) variable is the true cause of any difference in the dependent variable even when confounds are not known or measured.

2.3. Complications to manipulation and random assignment

The discussion above of manipulated independent variables and random assignment describes an experiment where all progresses as intended. As social science experiments are usually conducted by and on people, complications can occur. For example, it is common for some participants to not effectively receive the treatment level because they are distracted, do not wish to participate, do not understand the manipulation, or any other reason. It is common to include manipulation check questions after the dependent measures have been assessed to determine which participants agreed to, noticed, understood, and complied with the treatment as intended (for details and advice see Foschi 2014; Mutz 2021). For example, if racialized names are used on a resume to convey race-ethnicity a manipulation check would ask participants what race-ethnicity they thought the person was (at the end of the study).

There is some debate among experimentalists on how to proceed with data analysis based on answers to manipulation checks. One school of thought suggests omitting those who fail manipulation checks from subsequent analyses. The theoretical case suggests this approach because those who fail manipulation checks do not provide useful data to test the theoretical question at hand as they do not properly represent the effect of the level of the independent variable (e.g., if the participant did not read the name on the resume as denoting a black man their evaluation is not representative of how they would evaluate a black man). For example, Foschi (2014) suggests this approach but gives the important caveat that specific exclusion criteria should be decided prior to data collection. Others suggest omitting those who fail manipulation checks for statistical power reasons (e.g., Oppenheimer et al. 2009).

Increasingly, many experimental methodologists suggest including manipulation checks to assess the quality of the experimental manipulation but argue against excluding participants who fail these checks (Aronow et al. 2019; Kotzian et al., 2020; Mutz 2021). The power of random assignment to set all else equal across participants only holds on the full randomized sample. Excluding participants who fail manipulation checks could cause experimental conditions to become unbalanced, with an overrepresentation of certain types of participants in certain conditions (see section 6.1 for a discussion of balance). For example, if older participants are more likely to not understand the treatment condition (compared to their understanding of the control condition), older participants would be underrepresented in the treatment condition if those who fail manipulation checks are excluded from analysis.

Our view is that both approaches have merit and the decision should be driven by the theoretical question to be answered but that *not* excluding participants who fail manipulation checks should be the default. In most cases, results should be very similar in either case. In the case of a well-designed experiment, very few participants will fail manipulation checks because the experimental manipulation will be clear and effective. Further, if each experimental condition is appropriately similar in terms of length and complexity, there should be no systematic difference in manipulation check failures across conditions. In this case, leaving the few participants who fail manipulation checks in the analysis sample will simply add some noise to the estimates but not bias them. However, excluding those participants has the potential to introduce bias when manipulation check failures are systematic across conditions and characteristics of the participants (Aronow et al., 2019).

However, we do not think researchers should always include those who fail manipulation checks without further thought. For a concrete example, consider the experiment by Pedulla (2014) which manipulated race and sexuality by items on a resume (names [Brad Miller or Darnell Jackson] and organizational affiliations [gay student advisory council or student advisory council]). Both race and sexuality are difficult variables to manipulate effectively as illustrated by the 14–15% of the sample who misunderstood the

sexuality manipulation and the 15–60% who misunderstood the race manipulation (depending on the condition). Pedulla (2014) argues for excluding those who fail the manipulation checks which we think is appropriate if the question to be assessed is “when both race and sexuality are read as intended, what is the effect on evaluations?”.⁹ However, if the question is more broadly “how are resumes with signals of race and sexuality evaluated?” then excluding those who did not read those signals as intended is inappropriate.¹⁰ That is, in almost any realistic situation race-ethnicity and sexuality will be read imperfectly and thus their practical effect will depend partially on this fact.

To summarize, manipulation checks are important to include in most experimental studies to assess the effectiveness of manipulations; failure rates are one important indication of the quality of the experimental design. In general, the best default is to report manipulation check failure rates but not to exclude participants who fail them from subsequent analyses. However, there are situations where excluding these participants is appropriate and we recommend a well-justified *a priori* decision be provided in these cases along with tests of balance (see Section 6.1).

3. Additional features of experiments

As noted above, the core features that define an experiment are manipulation of an independent variable and random assignment. In addition to these core features, there are other designations based on: (1) the way participants were assigned to conditions (within- or between-subjects), (2) types of dependent variables (attitude or behavior), and (3) the location where the study occurred (laboratory, survey, natural, audit, etc.). Below, we discuss these additional features of experiments and emphasize the importance of theory for driving methodological decisions about which to use.

3.1. Experimental designs

3.1.1. Between- and within-subjects designs

All experiments involve random assignment to conditions—thus, all provide causal estimates of the effects of experimentally manipulated independent variable(s). Experimental designs can be broadly classified as between- or within-subjects designs—though hybrid designs that mix the two approaches are possible. In a between-subjects design, each participant (subject) is assigned a single level of the independent variable. An example is a resume evaluation study where participants evaluate only one resume (e.g., Pedulla 2014). Similarly, Moderna and Pfizer’s COVID-19 vaccine trials randomly assigned participants to either a control/placebo condition or to the experimental treatment condition (vaccine)—but never to both.

In contrast, a within-subjects design assigns participants to multiple levels of the experimentally manipulated independent variable. For example, if participants evaluate multiple experimental resumes and provide ratings for each (e.g., Benard and Correll 2010; Correll, Benard, and Paik 2007), it is a within-subjects design. Depending on the within-subjects design, participants may participate in all or just a subset of the experimental conditions. Within-subjects designs have the advantage of being more economical because participants can act as their own control and because a single participant provides data on multiple conditions; thus, all else equal similar effect sizes can be detected with far fewer total participants in a within-subjects study than a between-subjects study (see section 5.8 for additional details).

Hybrid designs are also possible. For example, because of social desirability bias concerns, Correll et al., (2007) manipulated the race of job applicants between subjects—where participants rated multiple resumes that were either all white applicants or all black applicants.¹¹ However, their key manipulation of parental status was manipulated within-subjects: e.g., a single participant evaluated both a mother and a woman without children.

Many permutations on these basic designs exist. In particular, multiple types of within-subjects experimental designs have increased in popularity in recent years, which we outline below.

3.1.2. Factorial designs

A factorial design includes multiple experimentally manipulated independent variables that are usually fully crossed—meaning all possible combinations are included.¹² For example, a 2 (Gender: Man, Woman) X 4 (Race-Ethnicity: White, Black, Latino, Asian) X 3 (Age: 25, 55, 85) full factorial design would include all $2 * 4 * 3 = 24$ combinations. Such designs are of fundamental importance to social science questions that are conditional in nature. For example, questions about intersectionality can be addressed by examining each unique combination rather than assuming outcomes are a simple summation of main effects (e.g., Benard et al., Forthcoming; Pedulla, 2014; Tinkler et al., 2019). Factorial designs can be implemented in between- or within-subjects experiments, though they are most commonly combined with a within-subjects design with participants rating multiple experimental conditions (Auspurg and Hinz 2014). Such designs can be particularly useful for combatting social desirability bias; because so many factors are varying across each

⁹ This quantity to be estimated is sometimes referred to as the average treatment effect on the treated as only the effectively “treated” are included in the analysis.

¹⁰ This quantity is the average treatment effect.

¹¹ By social desirability bias, we mean the possibility that participants give answers that they perceive to be in line with expectations and norms rather than their true beliefs when they perceive their true beliefs as socially undesirable.

¹² The exception is when some combination is implausible, e.g., a combination of age and work experience indicates a 25-year-old with 30 years of work experience. Such implausible combinations should generally not be fielded (Auspurg and Hinz 2014).

experimental condition, it is difficult for participants to guess what the study is about or to single out a particular experimental manipulation.

Another reason factorial designs often utilize a within-subjects approach is the number of combinations becomes large quickly and having participants rate multiple conditions reduces the total sample size needed (e.g., participants rated between 10 and 30 conditions in [Auspurg et al., 2017](#)). In the case where the universe of combinations is large, simply randomly sampling from among the possible experimental manipulation combinations to determine which to field to each participant is inefficient. More advanced methods such as D-efficient/D-optimal designs—which guide which subsets of combinations to field to different participants—are recommended ([Auspurg and Hinz 2014](#); [Su and Steiner 2020](#)).

3.1.3. Conjoint designs

The aforementioned designs involve participants rating each experimental condition individually—even if they rate multiple conditions sequentially with a within-subjects design. In a conjoint design, participants evaluate multiple conditions (usually two at a time) simultaneously, which encourages direct comparison. Such comparisons can reveal even subtle preferences and acceptable trade-offs ([Bansak et al., 2021](#); [Green et al. 2001](#)). For example, [Hainmueller et al. \(2015\)](#) had participants view two citizenship applications side-by-side and then had them choose which of the two they thought were more deserving of citizenship.¹³

Recent methodological work suggests conjoint designs are particularly effective at combatting social desirability ([Hainmueller et al., 2015](#); [Horiuchi et al. 2021](#)). Somewhat surprisingly, one study showed that having participants compare two conditions led to less socially desirable responses than viewing a single condition—and this held even in the absence of a forced choice between the two ([Hainmueller et al., 2015](#)). While more methodological work is needed, we find factorial and conjoint designs very promising for the study of topics prone to social desirability. See section 7.5 for a more thorough and critical discussion of social desirability concerns along with a description of some other popular experimental designs for combatting social desirability.

3.2. Type of dependent variable: attitudes vs. behavior

Although attitudes/behavioral intentions are generally the best available predictors of actual behavior ([Vaisey 2014](#)), their relationship with behavior is not perfect ([Collett and Childs 2011](#); [Melamed et al., 2019](#); [Mutz 2011](#); [Pager and Quillian 2005](#); [Quadlin 2018](#)). Due to potential disparities between attitudes and behavior, and the importance of behavior for many social science research questions, researchers are often interested in measuring behavior directly. Although researchers typically think of laboratory studies as behavioral (and surveys as attitudinal), this is not always the case ([Manago et al. 2021](#); [Mutz 2011](#); [Peterson et al. 2021](#)). For example, many laboratory studies include measures of both attitudes and behavior. E.g., Researchers often ask questions about how participants perceive their partners in laboratory experiments or measure other aspects of their beliefs, using these measures for further evidence of a theory's mechanisms (e.g., [Manago et al., 2021](#); [Melamed et al., 2019](#); [Melamed and Savage 2016](#)). Additionally, some survey experiments are designed to directly assess behavior (e.g., [Doyle 2021](#); [Manago et al., 2021](#); [Peterson et al., 2021](#)). For example, if researchers are interested in individuals' prosocial behavior, researchers can tell individuals they have the option to donate a certain amount of money and ask if/how much they are willing to share. The act of choosing the amount of money given to the shared pool or to donate to charity is a behavioral measure that does not require researchers to witness someone's behavior in the lab (e.g., [Abascal 2015](#); [Sanders 2017](#); also see the more complex design of [Simpson et al., 2018](#)).

3.3. Study location and control

Experiments can occur in a range of locations. For example, experiments can take place in a highly controlled setting such as a laboratory, embedded in a survey taken on the phone or online, in a natural or field setting such as a classroom, and/or in the online environment. Control is a desirable feature of experimental research because it allows researchers to eliminate other possible explanations for research findings and to eliminate distractions for participants; however, the amount of control varies greatly across settings. For example, outside of the laboratory, participants may be somewhat distracted or experiencing a wide array of life circumstances that can affect findings.

Increasingly, researchers are conducting behavioral studies in the online setting ([Peterson et al., 2021](#)). Many use online panels of participants for studies that measure attitudes or require reading vignettes or similar treatments and answering a series of questions (e.g., similar to those that were previously administered via survey either over the phone or by mail). Some studies use online settings like social media sites as their platform of interest (e.g., [Bail et al., 2018](#); [Kramer et al. 2014](#)) or design their own online environment (e.g., [Salganik et al. 2006](#); [Salganik and Watts 2008](#)). Increasingly, researchers are conducting behavioral research online adapting designs that were traditionally confined to the lab (see, for example, [Arechar et al. 2018](#); [Doyle 2021](#); [Giamattei and Lambsdorff 2019](#); [Horton et al. 2011](#); [Manago et al., 2021](#); [Melamed et al. 2018](#); [Simpson et al., 2018](#)). The use of the online setting is particularly attractive because it requires fewer resources than in-person experiments. Specifically, researchers do not need: to develop their own participant pool, physical space, as much support from research assistants, or as much time to complete the study ([Kraut et al., 2004](#); [Litman et al. 2017](#); [Manago et al., 2021](#); [Suri and Watts 2011](#)). One potential downside to non-laboratory-based experiments (i.e., those in natural settings, online, via survey, or via phone), however, is a lack of experimenter control ([Crump et al. 2013](#); [Reips 2000](#)). Yet there are

¹³ The condition we are describing is labeled “paired conjoint forced choice” in their study; they field and compare other experimental designs in the same study.

many promising ways for researchers to implement control outside of the lab (see, for example, the suggestions in [Manago et al., 2021](#)).

4. Theory testing and internal/external validity

Issues of validity are key to assessing evidence provided by any method. Experiments are often cited in discussions of validity and traditionally viewed as a tradeoff between high internal but low external validity. Experimentalists from multiple disciplines have challenged this traditional view but from different angles. We discuss both below with a focus on theoretical considerations.

4.1. Identifying individual processes for experimental intervention

In natural settings, social processes are multifaceted and complex. Like any large or complex issue, it can be useful to examine social processes in smaller pieces. For example, there are multiple processes that generate social inequality. To understand how to intervene in social inequality, however, it is helpful to identify individual processes and examine them separately. By identifying a single process, and learning how it generates inequality, researchers can begin to identify ways to intervene in that process. Over time, researchers can chip away at the myriad processes that generate inequality—identifying specific interventions, and then combining these interventions for a more comprehensive response to social inequality.

To identify these individual processes, how they function, and how to intervene, experiments are ideal. Specifically, as noted above, many social processes are multifaceted, meaning that there are many causes of any specific outcome. By using an experiment, researchers can examine one specific process at a time, isolating it from other causes, and seeing how it functions by itself. This kind of isolation cannot always be reproduced in natural settings because it is often possible to control for all other factors. But, without this isolation, it can be difficult to understand the relative importance of different factors on an outcome of interest.

For example, stigma is produced by multiple paths. One such path is labeling. When a label, for example of a certain mental illness like schizophrenia, is applied to an individual, research suggests it will result in stigma. Although a label results in stigma, concurrent behavior such as the individuals' deviant (symptomatic) behavior suggestive of mental illness may also result in stigma even in the absence of an explicit mental illness label. This leads to the question of whether the deviant behavior or the label itself is the true source of stigma from others. A well-designed experiment can control for other factors and manipulate labels and deviant behavior separately to examine the contribution of a label separately from (or in addition to) deviant behavior for stigma ([Manago and Mize, 2022](#)). Such processes are extremely difficult to examine in observational data because of the usual concurrence of deviant (symptomatic) behavior and a label for the mental illness. Studied experimentally to separate the two effects, researchers can see if labels individually contribute to stigma. If so, researchers can design interventions into these processes. If, however, labels have little to no effect on stigma, researchers can focus instead on the other factors that contribute to stigma. Importantly, researchers must know *a priori* what kinds of factors to manipulate and isolate. Theoretical considerations can and should guide such decisions.

4.2. The theory and goals of experimental design

Experiments are ideal for establishing causal inference. That is, an experiment can test a prediction that a specific independent variable generates an effect on a dependent variable. Sometimes, experiments are used *atheoretically* to determine whether a possible intervention matters and the causal inference is valid even in this case. However, absent a theoretical explanation it is difficult to know how (or if) we can apply the experimental intervention outside of the narrow context that the experiment was conducted in. For these reasons, social science experiments generally aim to test theoretical processes and are most useful for testing a pre-specified, theoretically derived (*a priori*) hypothesis.

To generate a testable and falsifiable hypothesis researchers rely on prior theory. A theory used to generate a hypothesis may include some, or all, of the following features: (1) abstraction, (2) exact class/formal definitions, (3) formal logic, and (4) initial and scope conditions ([Markovsky 1995](#); [Robinaugh et al., 2021](#); [Sell 2018](#)).¹⁴ First, abstraction means that rather than describe a specific phenomenon, the theory describes a class of phenomena. Second, exact class (or formal) definitions frame what is necessary and sufficient for the identification of a given concept ([Sell 2018](#)).¹⁵ Third, formal logic includes a set of interrelated propositions that lead to a derived hypothesis, thereby allowing for falsification ([Popper 1959](#)).¹⁶ Finally, "Initial conditions define particular circumstances, such as the time and place, under which a theory will be relevant ([Cohen 1980](#); [Sell and Martin 1983](#); [Walker and Cohen 1985](#)) ... [In contrast] scope conditions are the abstractly defined conditions under which a theory is expected to be true ([Cohen 1980, 1989](#); [Foschi 1997](#); [Freese 1980](#); [Lucas 2003](#); [Sell and Martin 1983](#); [Walker and Cohen 1985](#); [Webster and Kervin 1971](#))" ([Manago 2020:95–98](#)). The extent to which these features are incorporated into a theory, and the extent to which they are being tested in a specific experiment, should help guide the specific experimental design used.

Importantly, experiments are conducted for different, or even multiple, reasons. In social science research, researchers often use experiments to test the extent to which a theory may apply to a specific situation. Other times, researchers use experiments to test the abstract conditions of a theory. Finally, under some circumstances, an experiment might also test both the abstract conditions of a

¹⁴ For a textbook example of a formal theory containing all of these aspects see [Appendix B](#).

¹⁵ Although exact class definitions facilitate the use of formal theories, as noted by [Sell \(2018\)](#), formal definitions can be useful even without the use of a formal theory.

¹⁶ Falsifiability is the capacity for a hypothesis and/or theory to be proven wrong. For more information, see [Popper \(1959\)](#).

theory and its application under certain conditions.

4.3. Goals of experimental design, validity, and generalizability

To evaluate the validity of a given experiment, it is important to consider the goals of the experimental design (Lucas 2003; Mook and Douglas 1983; Thye 2014b). There are two primary types of validity: internal and external. Internal validity refers to the integrity of the research study/experiment itself. A study is said to have internal validity if it, as designed, is an appropriate test of the theory and research questions at hand.

Causality is key to assessments of internal validity and, because well-designed experiments can establish causality, experiments are generally seen as having high internal validity. External validity refers to be the extent to which a study's findings can generalize beyond that study, typically to different settings or populations. Traditionally, experiments are thought to have low external validity though we find this overgeneralization unhelpful. On one hand, experiments increasingly use diverse or nationally representative samples so an outdated view of experiments' external validity hinging only on undergraduate samples is often incorrect (Mutz 2011). In addition, there has long been an argument for the external validity of experiments through their strict test of theoretical relationships, regardless of the diversity of the sample (Lucas, 2003; Zelditch, 1969).

We agree with Gërxhani and Miller (2022) that experiments' high degree of internal validity is a strength and that, in general, we should focus on internal validity first before turning to concerns about external validity:

"For many, internal and external validity cannot be in a tradeoff relationship because internal validity is a pre-requisite of external validity. In other words, there is no point in worrying about the generalization of some experimental results until one is confident about the internal validity of that result." (p.313).

Below, we outline a more nuanced view than the traditional one for how to assess the external validity of experiments.

4.3.1. External validity

A helpful way to think about external validity is generalizability, or asking: to whom or what can the findings of this experiment apply? Broadly speaking, there are four forms of generalizability: participants (perhaps better referred to as population), settings, measures, and treatments (as defined by Mutz (2011)). Below, we detail these forms of generalizability, and discuss which are most relevant based on the goals of the experimental design.

Participant or population generalizability refers to the process of generalizing from those who are in the study to some other population. For example, if a study was conducted only on undergraduate students, can researchers confidently say that the same processes would occur for the general population (see Section 5 for an overview of methodological research on this question)? Setting generalizability refers to the process of generalizing from one setting to another setting. For example, if studying helping behavior based on European football (soccer) affiliation, would the same processes occur in an American football setting (Levine and Prosser 2005)? Measure generalizability refers to the ability for different operationalizations of the same variables to be used and produce the same outcome. For example, if researchers are interested in decreasing inequality, does the intervention work for multiple operationalizations of inequality (so long as they all fit the definition of the concept)? Finally, treatment generalizability is the ability to implement different, but theoretically similar, interventions to test the effect of a manipulation on an outcome of interest. For example, if researchers are interested in the effect of a label on stigma, do all labels have the same effect on stigma, or does the effect on stigma vary by the type of label (e.g., mental illness vs. criminal)?

Using experimental designs, researchers can answer different questions and will therefore have different considerations for generalizability. Not every experiment needs to be equally concerned with all forms of generalizability. Rather, as we describe below, concerns of generalizability should relate to the research question.

4.3.2. Example 1 of the relationship between theory, experimental design, and validity

An oft-cited example of external *invalidity*, and specifically, a lack of participant generalizability, is a (pilot) study in which researchers were said to examine the impact of obesity on uterine and breast cancer using male subjects (Dresser 1992; Jaschik 1990; Keville 1993).¹⁷ This paper is often used as an example of bias against women in medical science. On its face, this seems like a reasonable criticism—how can researchers study uterine cancer with male subjects who generally do not have uteruses?

Importantly, however, the actual study (which is not generally cited in criticisms of the work) is examining *estrogen metabolism* (Anderson et al., 1984; Jaschik 1990; Keville 1993). That is, in reality, the authors are trying to understand and generalize findings to cellular processes regarding metabolism (the theory). Importantly, the authors are *not* trying to generalize to any specific individuals, or even to the relationship between metabolism and uterine or breast cancer. In fact, the only discussion of the link to uterine or breast cancer comes in the "Discussion" section of the paper, in which the researchers note the potential broader impacts of their findings. In this section, the authors also address the use of males as subjects in the study:

¹⁷ Although gender is a complex social construct that spans binary categorizations, here, we are talking about biological sex (which is also non-binary) and studying breast and uterine cancer on people without uteruses. We use the same language (men/women) as the researchers, as we do not know how they identified sex vs. gender.

“In previous studies we demonstrated by radiometric methods significant quantitative differences in estrogen oxidations in men and women. The present investigations were limited to men because a more complex protocol would be required in women to separate the effects of diet from potential effects of hormonal changes associated with the ovulatory cycle. It will be important in the future to determine whether the regulation of estrogen oxidations by the ratio of dietary protein to carbohydrate is quantitatively or qualitatively different in men and women.” (Anderson et al., 1984:106)

On a scientific level, the researchers are stating that, although past research has shown some differences in estrogen oxidation (the dependent variable) among men and women, they do not know if there would be a difference in *the effect of diet on estrogen oxidation* for men and women (the relationship between the independent and dependent variable).¹⁸ The researchers leave this as an open question, noting the weaknesses of the design and the current unexplored scope.

In this instance, despite first impressions and the critical media attention surrounding the study, the choice of male participants is not necessarily problematic. Rather, researchers were trying, incrementally, to learn more about the relationship between diet, estrogen, and cancer. To reduce confounding variables, the researchers used male participants who have more stable levels of estrogen.¹⁹

This argument has also been used in the social sciences, especially at early stages of theory development. Specifically, to avoid possible confounding factors, some social scientists prefer to use samples of undergraduate students or other—relatively homogeneous—groups (Druckman and Kam 2011; Lucas, 2003). These kinds of samples may be appropriate when trying to generalize to the theory, rather than any specific population (Krupnikov et al. 2021; Lucas, 2003; Mutz 2011; Zelditch Jr. 1969). Future research can then test the participant/population generalizability.

4.3.3. Example 2 of the relationship between theory, experimental design, and validity

In another example, consider the famous Asch experiment in which individuals were shown a series of lines and asked about their length (Asch 1956). In the study, confederate participants were told to make the same false assertions about the length of lines.²⁰ The goal was to see if the (non-confederate) participant would conform to the opinions of others, even if they were demonstrably false.

Asch's experiment has been criticized for lacking a form of setting generalizability known as *mundane realism*, i.e., similarity to the outside world (Mutz 2011). For example, a study in which individuals were highly influenced about perceived line length may not provide insights on deeply entrenched views about topics such as abortion (Mutz 2011). As a result, some may question the value of Asch's contributions—noting that it lacks mundane realism. Due to a lack of specific theoretical cornerstones (e.g., abstraction, formal definitions, formal logic, and scope conditions), this criticism may be well-founded.

If, however, Asch had defined the scope of the theory and was seeking to test/generalize to this theory, the criticism of mundane realism would be irrelevant (Webster and Sell 2007; 2014; Zelditch 2007, 2014). Said otherwise, the key to using Asch's insights to generalize to another kind of issue or setting (such as deeply entrenched political beliefs) may be in identifying *why* the scope is limited and *how* it might be expanded. For example, perhaps this process of influence only works when individuals do not yet have strongly formed opinions about a topic. So, perhaps when new social issues are coming to the fore, exposure to one group's ideas may lead individuals to have shared opinions with the majority ingroup. These are scope-defining conditions that, when considered and specified in advance, can help clarify the theory and lead to a more accurate evaluation of the experiment's contribution (Webster and Sell 2007; 2014; Zelditch 2007, 2014; Zelditch Jr. 1969). As it stands, Asch's theory cannot be used to make broad claims about influence. If researchers want to make a broader claim, they can do so by conducting experiments that are designed to test a more completely specified theory.

Below, we extend this discussion of generalizability and the goals of experimental design to considerations of sampling.

5. Sampling

In our reading of past experimental studies, most authors either do not justify their sampling strategy at all or use one of two dominant justifications: (a) desiring a nationally representative sample of some target population or (b) convenience. Popular methodological texts are similar; e.g., Mutz's (2011) excellent *Population-Based Survey Experiments* is focused almost exclusively on nationally representative samples of the U.S. In contrast, the similarly exemplary *Laboratory Experiments in the Social Sciences* (Webster and Sell, 2014) is primarily focused on formal theory testing and includes little discussion of sampling as convenience sampling is normative for formal theory testing (see Druckman and Kam 2011; Krupnikov et al., 2021; Lucas 2003; Zelditch Jr. 1969 for defenses of this strategy). While we believe representative considerations and convenience concerns are two factors that researchers should sometimes consider, we recommend experimentalists embrace a larger array of considerations and be more explicit in justifying their sampling decisions.

We propose a new framework of *purposive sampling* for social science experiments, drawing on many prior frames, but most centrally Shadish et al. (2002). We propose five non-mutually exclusive motivations for sampling that, when considered in the context of a

¹⁸ On a gender-equality level, the researchers proceeded to study cancers that affect breasts and uteruses, on people with breasts and uteruses, thereby providing people with these sex organs options for cancer treatment. These efforts were designed to improve and/or extend lives.

¹⁹ Alternatively, Mutz (2011) makes the interesting point that if effect sizes are too small to be detected with this extra “noise” the effects may not be particularly meaningful. That is, Mutz (2011) argues that if the effect is too minor to be identified during women's hormonal fluctuations, perhaps the effect is not that important. Thus, the ability to control for certain phenomena may need to be balanced with considerations of effect sizes (see Section 6.3).

²⁰ In experimental parlance, confederate refers to research assistants who pose as participants for part of the experimental manipulation.

given theory, should guide researchers' sampling decisions: (1) sampling for convenience, (2) sampling a specific target population, (3) sampling for range, (4) sampling for [equal] representation, and (5) sampling for representativeness. We then discuss ways to combine multiple motivations; in particular, some aspect of convenience or representativeness is often fruitfully combined with another motivation. We end the section with a discussion of options for implementing these sampling criteria and some general guidelines for sample size in experimental studies.

5.1. Sampling for convenience

Some research questions can likely be answered with any sample. That is, some experiments are interested in testing theoretical processes that should apply to any group or culture. In these cases, one can convincingly argue for sampling for convenience. Second, some measurement techniques or behavioral interventions may only be possible within a laboratory, and the benefit of the design may be considered to outweigh the limitation of the sample to the local area. For example, [Willer and colleagues' \(2009\)](#) lab experiment of norm enforcement used an elaborate group wine tasting design that would be difficult to recreate in a survey experiment. Third, convenience samples are also useful for pilot and pre-tests or for initial tests of a theory. In these cases, it is useful to note potential limitations of the sample and to not overclaim the scope of the theory's application until further tests are conducted.

There is a long tradition within the social sciences of using convenience samples (usually undergraduate students) to develop and test formal theory ([Druckman and Kam 2011](#); [Falk and Heckman 2009](#); [Krupnikov et al., 2021](#); [Zelditch 2014](#)). In this case, the external validity of the findings are argued to hinge on the formality—and specifically scope conditions—of the theory (see Section 4 above), making a convenience sample appropriate ([Lucas, 2003](#)).

In any of these cases, our advice is to acknowledge and justify the appropriateness of the convenience sample for the specific research questions being addressed (see additional considerations in [Krupnikov et al., 2021](#)). However, as we outline below, there are many cases where a more targeted sample would better answer the research question and—in most of these cases—could be obtained for the same or similar costs to the researcher.

5.2. Sampling a specific target population

Some research questions can only be answered adequately with a specific population. For example, if a researcher wishes to know whether a new textbook will improve test scores for 4th grade students, the researcher needs to sample 4th grade students. No other population can definitively answer the research question.

In most cases, a target population will not be well represented in either convenience or population-based samples (e.g., of U.S. adults). For example, if a researcher wants to study libertarians there will be few in a U.S. population sample; specifically recruiting libertarians is necessary. In these cases, a targeted sample will ensure the statistical power to test the research question in a sample of the population most relevant to the theoretical question.

5.2.1. Sampling for expertise

A sub-category of sampling a specific target population is sampling for expertise. Here, a researcher may expect that the processes they wish to study only apply to those with specific skills and expertise. For example, [Stepanikova \(2012\)](#) tested whether a patient's race-ethnicity influences medical doctors' evaluation of their symptoms and suggestions for treatment—therefore, she sampled medical doctors exclusively. It is reasonable to expect that those without medical training would not have made the same patient evaluations. Similarly, those with hiring experience, training, and/or management experience might make different decisions about job candidates and employee performance evaluations than those without—so samples of people with hiring experience and/or managers are justified for these types of research questions (e.g., as implemented in [Castilla and Benard \(2010\)](#) and [Quadlin \(2018\)](#)).

5.3. Sampling for range

We borrow the concept “sampling for range” from qualitative methodologists ([Weiss 1995](#)). The key idea is that a given sample should be made intentionally diverse on theoretical variables of interest. For example, if it is reasonable to expect that people of different ages will respond differently to your treatment, then sampling to ensure diversity of ages of your participants is warranted. This ensures statistical power to test whether age moderates the effect of treatment. Age provides one example where convenience samples of undergraduates may be inappropriate as age of the participants tends to be relatively homogenous in these samples.

Popular online samples like Amazon mTurk and Prolific tend to be far more diverse than undergraduate samples (see section 5.7 below). However, these samples still underrepresent older adults, adults without college degrees, and political conservatives ([Levay et al. 2016](#)). For this reason, we often oversample each of these groups when using such samples to ensure adequate range to test whether results differ across the age, educational attainment, or political views of the participants—factors that often correlate strongly with topics social scientists' study (e.g., [Petts et al. 2022](#)). However, sampling for range is most effective when tailored to the specific research question, as prioritizing range on other factors (e.g., race-ethnicity) may better suit certain studies' goals.

5.4. Sampling for [equal] representation

Social scientists are often interested in comparing across groups: e.g., gender, race, class, age, political party, etc. In many cases, these groups are not equally represented in the general population or on college campuses. For example, if a researcher wants to determine whether their treatment is responded to similarly by White, Black, Latino, and Asian participants in the U.S., neither a convenience nor population-based sample are likely to yield adequate—and certainly not ideal—statistical power for each group.

In these cases, we suggest researchers sample to ensure adequate representation of each group of interest. Ideally, the sample would have equal representation of each racial-ethnic group—thus helping ensure any observed statistical differences across groups is due to real differences and not an artifact of statistical power.

Doan et al. (2014) provide an exemplar for sampling for (equal) representation. Their study focused on attitudes toward different- and same-gender couples with specific hypotheses about how heterosexual vs. gay/lesbian participants will view each. To answer their research questions, they equally sampled heterosexual men, heterosexual women, gay men, and lesbian women. This approach was necessary to answer their research questions as a general population sample would yield far too few sexual minorities ($\approx 7.1\%$ of the U.S. adult population identifies as LGBT with only 2.5% identifying specifically as gay or lesbian (Jones 2022)).

5.5. Sampling for representativeness

In some cases, population estimates are necessary, and in this case, sampling for representativeness is warranted. First, a population of interest is identified, and then the usual strategies for ensuring representative samples are followed. The exact sampling methods are beyond the scope of our review, but random sampling, stratified random sampling, quota sampling, and similar methods are common (for details see Lohr 2021).

A country's adult population is often used as the target population when sampling for representativeness. While this sometimes matches a research question well, we suggest this is a fairly arbitrary population and is not best for every—or even most—research questions. For example, consider a research question about campaign ads and presidential voting intentions in the U.S. (e.g., Coppock et al. 2020). A nationally representative sample is necessary to determine the proportion of the population that plans to vote for a given candidate with and without the intervention of the campaign ad. In this case however, the entire U.S. adult population isn't of interest—only likely voters are of interest. First, not all U.S. adults are eligible to vote and some who are eligible are not registered to vote. In addition, among the subset of eligible and registered voters not all do vote—e.g., 44.3% of eligible voters did not vote in the 2016 presidential election. Moreover, the U.S. presidential election provides a particularly useful example as the electoral college determines the winner, meaning it is actually the population of voters within each swing state that are of interest. Thus, while representativeness may be a fundamental concern, the entire country's population is likely not the population of interest.

5.5.1. Do I need a nationally representative sample?

Consider a two-condition between-subjects experiment (e.g., a control and a treatment condition). With a continuous dependent variable, we may fit a linear regression and obtain estimates of the form:

$$\hat{y} = \hat{\alpha} + \hat{\beta}_{treat} treat_i \quad (1)$$

Where $\hat{\alpha}$ is the mean for the control condition which serves as a baseline reference and $\hat{\beta}_{treat}$ is the difference in the mean of the treatment group from the mean of the control condition. When the population value of α (e.g., how many people in the population plan to vote for a candidate without any intervention) or of $\alpha + \beta_{treat}$ (e.g., how many people in the population plan to vote for a candidate with the intervention) is of interest, sampling for representativeness is necessary. While there are certainly situations where these quantities are of key interest (necessitating an accurate estimate of α for the population in either case), we believe most social science research tends to be interested primarily only in β_{treat} ; i.e., what is the effect of the experimental manipulation?

Much recent methodological work has studied the replicability of treatment effects across samples. Several projects have replicated social science experiments conducted on both nationally representative and convenience samples. This work, which includes a wide range of experimental designs and social science topics has overwhelmingly tended to find that while $\hat{\alpha}$ estimates differ across the samples—as we would expect—the estimate of $\hat{\beta}_{treat}$ tends to replicate (e.g., Coppock 2018; Mullinix et al., 2015; Weinberg et al. 2014). Similarly, the “many labs” project in psychology that replicated myriad experimental effects across a range of samples and settings concluded “... replicability is more dependent on the effect itself than on the sample and setting used to investigate the effect” (p.142; Klein et al., 2014, 2018).

The primary reason $\hat{\beta}_{treat}$ tends to replicate is the general absence of heterogeneous treatment effects (see Section 6.4 for details on heterogeneous treatment effects; Coppock 2018; Coppock et al., 2020; Coppock et al. 2018).²¹ That is, while participants tend to enter studies with different baseline attitudes and opinions, they tend to react similarly to treatments (e.g., small shifts in favorability after viewing a political ad; Coppock et al., 2020).

We should caveat that the replication of $\hat{\beta}_{treat}$ across representative and convenience samples is an empirical finding and not

²¹ Note that we are referring to effects of treatments across different characteristics of participants (e.g., their age, gender, or race-ethnicity). There are many treatment effects that are heterogeneous across other treatments—but this is not directly related to sampling decisions.

necessarily some universal truth.²² In addition, most of this methodological work has been done on U.S. samples. Nonetheless, the evidence is consistent and compelling and suggests many social science questions can be profitably answered in much less costly samples than a representative sample. In addition, researchers can use our strategy of sampling for range to ensure adequate statistical power to test whether their treatment effects are homogenous across factors of theoretical import (see Section 6.4 for details on these tests).

5.6. Additional considerations and combining multiple motivations

In practice, a given research question may require some combination of the five purposes outlined above. For example, if there is a known sampling frame of experts (a specific target population), a researcher could sample for representativeness by randomly sampling from within that list of experts. However, more often a convenience sample of experts will be the only feasible option and preferable to a representative sample of a broader population (e.g., of all U.S. adults among which there are few experts on the topic of interest).

Our examples above primarily discuss sampling on stable characteristics of the participants (e.g., their demographic traits). However, the same logic can be used to sample on levels of an outcome of interest. For example, if one wishes to know whether a new treatment is effective against depression, the specific target population may be depressed individuals—with the eventual outcome of interest being depression levels after the treatment. In contrast, one may wish to ensure there is a range of depression levels represented depending on the exact research question.

5.7. Options for implementing our purposive sampling framework

The rise of online panels has made sampling for survey experiments cheaper and easier than ever. Implementing most of the strategies discussed above for lab experiments is more difficult but not impossible. For example, screening surveys can be used to obtain details on participant characteristics and then only those matching sampling criteria of interest can be invited to complete the experimental study (e.g., Manago et al., 2021).

For online panels, Amazon mTurk is a popular service for obtaining research participants. Researchers can restrict their sample to residents of a specific country for no extra cost. For additional cost, users can apply “premium qualifications” to restrict the sample to participants with certain characteristics (e.g., age, gender, income, education, political affiliation, etc.; see full list at <https://requester.mturk.com/pricing>). Prolific is an increasingly popular tool tailor-made for conducting research online. A benefit of Prolific is the ability to make any number of sample restrictions for no additional cost; in addition, Prolific has a wider range of background characteristics of their participants to choose from (e.g., race-ethnicity, sexual orientation, detailed employment information, etc.; see full list at <https://www.prolific.co/#audience>). Time-Sharing Experiments for the Social Sciences provides researchers with representative samples of the U.S. through a competitive peer-reviewed proposal system (see <https://www.tessexperiments.org/>). Researchers can propose more targeted samples with justification (e.g., Doan et al., 2014).

Countless other options exist for sampling participants for experiments. Our advice is to consider the theoretical motivations of the study first and to let that guide sampling decisions, rather than defaulting to the pure convenience or representative options in all cases.

5.8. Some general guidelines for sample size

In general, because random assignment controls for all characteristics of participants, experiments tend to have relatively high statistical power. Thus, smaller samples for experiments often suffice compared to needed sample sizes for observational studies. Here, we give some general baselines for sample size. These are only meant as heuristics and starting places—conducting a power analysis tailored to a specific study by using means and standard deviations from published work or a pre-test is always the best option.

Consider the simplest experimental design with two conditions (e.g., control and treatment). Consider also two options: a between-subjects design vs a within-subjects design. For each possibility, what is the needed sample size to observe a standardized effect size that is small, medium, or large?²³ Table 1 displays results of power analyses assuming a continuous dependent variable. The within-

Table 1
Needed sample size for various standardized effect sizes for a two-condition experiment.

Effect size	Cohen's <i>d</i>	Needed <i>N</i>	
		Between-Subjects	Within-Subjects
Small	0.2	787	198
Medium	0.5	128	33
Large	0.8	51	14

NOTES: Between-subjects *N* determined for a two-sample *t*-test; within-subjects *N* for a one-sample *t*-test. All power analyses use $\alpha = 0.05$ and $\beta = 0.80$ which are conventional values (Cohen 1988).

²² Having given this caveat the evidence is impressive and includes a wide range of different types of experiments across a wide range of topics.

²³ We are using Cohen's *d* as our standardized measure of effect size; see Section 6.3 below for details on how *d* is calculated.

subjects design requires far smaller sample sizes because each participant provides data for both experimental conditions and because each participant acts as their own control—accounting for idiosyncratic variation across participants.

These are only meant as rough guidelines. Researchers should consider what effect size is meaningful or not in a given context; e.g. a small effect size may be important in one study but irrelevant in another. For the calculations in Table 1, we are also using a conventional significance level of 0.05 which may not be appropriate in all cases (Wasserstein et al. 2019). In addition, the above guidelines are only for the main effect of the experimental condition (i.e., whether the means in control and treatment conditions differ). Often, additional tests are desired and/or comparing effects across subgroups of participants is of interest, which generally require larger sample sizes.

6. Analysis of experiments

Many aspects of experimental design make statistical modeling of experimental data simpler than many observational studies. In addition, the primary tools for analyzing experimental data—e.g., regression and ANOVA based approaches—are the same as those typically used by social scientists for observational studies. However, there are some unique considerations for effective analyses of experiments.

6.1. Testing for covariate balance

The statistical theory that underpins random assignment's ability to set all else equal between experimental condition holds only "in expectation" or as sample size approaches infinity. In finite samples, random chance alone may allow some factors (e.g., certain participant genders or ages) to be over-represented in some experimental conditions. Testing for balance is usually only applicable for experiments that involve between-subjects conditions or a partial within-subjects design; by design a fully within-subjects experiments' conditions are balanced because each participant completes all conditions. The exception is when participants have missing data on only some within-subjects conditions and/or drop out part way through an experiment; in such a case, balance tests are appropriate even with within-subjects designs.

6.1.1. When you should test for balance

Experimental methodologists disagree about the usefulness of balance tests. In particular, Mutz et al. (2019) propose a prominent argument against testing for balance. The argument, in brief, is that only certain characteristics of a participant are measured in any given experiment. We are limited to testing for balance to measured characteristics. Just because we find imbalance on a measured characteristic does not mean there is not imbalance on other unmeasured characteristics (e.g., an aspect of a participant's genetics). We may be tempted to adjust our statistical models for the unbalanced measured characteristics (e.g., add a control for gender if women are over-represented in a certain condition). However, assuming the effects of the characteristics of participants are normally distributed, we should expect there are other unmeasured characteristics that are also unbalanced but we are unable to add to our models. Adding unbalanced measured characteristics to the models can bias our estimates of the treatment effects because we cannot adjust for the countervailing effects of the unmeasured (or unmeasurable) characteristics.

In general, we agree that evidence of imbalance is not reason to discard experimental data. Especially if evidence of imbalance is small in size—i.e. $< |0.2|$ based on the d or h statistics detailed in the next section—we see little reason for concern even if statistically significant imbalance exists on a given factor. However, the exception is for experiments where there are substantial amounts of missing data—e.g., due to dropout, participants refusing to answer some questions, or participants failing attention or manipulation checks. In such cases, it is useful to test whether this missing data mechanism is systematic—e.g., older participants were more likely to not understand a certain condition's manipulation and thus are underrepresented in that condition because they didn't answer the subsequent questions. If age is associated with the outcome, statistically accounting for the imbalanced characteristic is appropriate, but we note that the elegance of random assignment to ensure causal inference is lost and we must then trust the assumptions of the statistical model (e.g., that controlling for age in a regression appropriately removes its confounding effect on the treatment).

6.1.2. Determining balance for individual covariates

Testing balance for individual covariates is straightforward. For example, to test whether the mean age in the control and treatment conditions is significantly different in a between-subjects design you can use a t -test:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\text{var}(x_1)}{n_1} + \frac{\text{var}(x_2)}{n_2}}} = \frac{\overline{age}_{control} - \overline{age}_{treatment}}{\sqrt{\frac{\text{var}(age_{control})}{n_{control}} + \frac{\text{var}(age_{treatment})}{n_{treatment}}}} \quad (2)$$

Where individual t -tests can be performed for all continuous covariates. For binary covariates, a proportion test can be used. For nominal covariates, a χ^2 test of association is appropriate.²⁴ These examples represent minimal tests of balance; it is possible to test for other aspects of distributional equivalence (e.g., whether the variance is the same across conditions; see Austin 2009; Greifer 2022). In our experience, most researchers only provide minimal tests of balance but additional tests are useful in some situations and random

²⁴ Our goal is not to recommend specific tests but instead to illustrate that bivariate tests for balance are straightforward and simply involve choosing the right bivariate test given the measurement level of the covariate and the experimental condition variable.

assignment should balance covariates across all moments of a distribution—not just the first moment (the mean).

Many software packages exist to automate testing for balance, e.g., `balanceplot` in Stata (Mize n.d.) and `cobalt` in R (Greifer 2022). In addition to providing significance tests of imbalance, these packages also standardize the degree of covariate imbalance which helps to understand both the magnitude of imbalance and to compare imbalance across covariates measured on different metrics (e.g., age and income). Standardized imbalance is calculated as (Rosenbaum and Rubin 1985):

$$d = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\text{var}(x_1) + \text{var}(x_2)}{2}}} \quad (3)$$

Which is usually used regardless of the covariate's measurement level. However, special calculations for binary variables or specific categories of nominal variables can be used (Cohen 1988):

$$h = \varphi_1 - \varphi_2 \quad (4)$$

Where φ_k is calculated as:

$$\varphi_k = 2 * \arcsin(\sqrt{p_k}) \quad (5)$$

Where p_k equals the proportion in condition k .

6.1.3. Omnibus tests of imbalance for all covariates

Testing for imbalance on all covariates individually raises the likely possibility of Type 1 errors (false positives). Often, a more informative test is an omnibus test of imbalance that considers all covariates simultaneously. For these tests, it is useful to think of the experimental condition as the dependent variable and all covariates as independent variables; the logic of the test is whether or not considering all covariates simultaneously provides predictive information about which condition a participant was assigned.

For a two-condition experiment, binary logit or probit is appropriate; for a three-plus condition experiment, multinomial logit is appropriate; for experiments with continuous treatments, linear regression can be used. Of interest is the omnibus test for the whole model; the F -test in linear regression and the likelihood-ratio test for logit, probit, or multinomial logit. A null test suggests covariate balance.

6.2. When to use control variables with experimental data

For most studies, control variables are not *needed* with experimental data, though they may be useful in some contexts. As Mutz (2011) states “Random assignment to conditions means that the very notion of ‘control’ variables makes no sense because control is already embedded in the design itself.” (p.127). That is, the point of doing an experiment is to control for characteristics of the participants and thus further controlling for these factors statistically is redundant. In our informal reading of published experimental studies, we see researchers generally either (a) use no control variables at all, or (b) add the same full set of control variables social scientists usually use in observational data (e.g., gender, race, age, SES, political affiliation, etc.). Neither approach matches best practice.

In general, we agree with Mutz (2011) that “... experimental inference is always valid with no covariates whatsoever.” (p.126). And indeed, the average treatment effect (ATE) estimate is an unconditional estimate of differences in means across conditions (Holland 1986):

$$ATE = E(y_{\text{control}}) - E(y_{\text{treatment}}) \quad (6)$$

However, we also agree with Mutz's (2011) recommendation that some well-selected control variables can aid inference. Specifically, controlling for variables that are independently predictive of the outcome should reduce standard error estimates on the treatment variable even if they are uncorrelated with the treatment—as they should be due to random assignment. These potential control variables could be characteristics of the participants (e.g., men are sometimes more critical than women regardless of the experimental condition; see, for example, Mize 2019a) or other treatments in the study. Why does controlling for an orthogonal covariate reduce the standard error estimate on the treatment effect? Consider the formula for the variance estimate of a multiple linear regression coefficient—where $SE(\hat{\beta}_{\text{treatment}}) = \sqrt{\text{Var}(\hat{\beta}_{\text{treatment}})}$ —with covariate z included in the model.

$$\text{Var}(\hat{\beta}_{\text{treatment}}) = \frac{\sigma_e^2}{N\sigma_{\text{treatment}}^2(1 - \rho_{\text{treatment},z}^2)} \quad (7)$$

If random assignment is effective, the correlation between treatment and the covariate ($\rho_{\text{treatment},z}$) should be effectively zero.²⁵

²⁵ Due to random variation this is unlikely to be exactly zero. Thus, controlling for covariates may have some small effect on the estimate of $\hat{\beta}_{\text{treatment}}$; however, it is unlikely to be large enough to impact conclusions. In addition, in the case of multiple covariates the effects on the estimate of $\hat{\beta}_{\text{treatment}}$ should be zero on average and equally likely to minimally increase or reduce the estimate and thus their effects on $\hat{\beta}_{\text{treatment}}$ should average out to zero.

However, if the covariate z explains additional variation in the outcome the error variance of the model (σ_e^2) will be smaller when z is in the model, which will produce a smaller $SE(\hat{\beta}_{treatment})$. Models with and without z should produce the same estimate of $\hat{\beta}_{treatment}$ (with some random deviation), but the model with z should produce a smaller $SE(\hat{\beta}_{treatment})$ estimate, allowing the analyst to better isolate the effect of the treatment from other systematic sources of variation in the outcome.

6.3. Standardized effect size estimates

Depending on field norms, variants of ANOVA (e.g., psychology) or regression (e.g., sociology and political science) tend to be used to test for treatment effects. In either case, it is useful to standardize effect estimates to better contextualize the size of a treatment effect and to facilitate comparison across studies (we usually report effects both in the natural metric of the variables and also in standardized form as both provide useful information). The most popular option for standardized effects is to use [Cohen's \(1988\)](#) guidelines for effect size. A Cohen's d or Cohen's h are first calculated based on Equations (3) and (4) above and then the following rough guidelines are used: anything below 0.2 is considered a trivial effect size, 0.2 and above is a small effect size, 0.5 and above is medium in effect size, and 0.8 and above is a large effect size (above, we used these thresholds for the sample size guidelines discussed in [Section 5.8](#)). While we find these thresholds useful as general guidelines, we encourage researchers to contextualize their effect sizes in the context of their research area and methodology—even small effect sizes are important in some areas and even large effect sizes do not warrant action in other areas.

Cohen's effect sizes as shown in Equations (3) and (4) assume two group comparisons (e.g., control and treatment or two specific conditions from a 3+ condition experiment being compared). Earlier, we encouraged utilization of continuous treatments in some cases. One option to understand the magnitude of the effect in this case is to examine effects over the range of the continuous treatment. Another solution proposed for calculating standardized effects for continuous variables—especially when wishing to compare their effects to that of binary treatments—is to examine a +2 standard deviation change ([Gelman 2008](#)). For example, we can calculate a +2 standard deviation discrete change marginal effect (see [Long 1997](#)) and then standardize:

$$d_{continuous} = \frac{+2 \text{ SD marginal effect}_{treatment}}{\sqrt{\frac{\text{var}(x_1) + \text{var}(x_2)}{2}}} \quad (8)$$

One can then use the 0.2, 0.5, and 0.8 guidelines as above to determine the effect size category.

6.4. Testing for heterogeneous treatment effects

The effect of an experimental manipulation (i.e., the treatment effect) might operate similarly across social locations (e.g., across different groups of participants) or the treatment effect might vary depending on a participant's social location (i.e., a heterogeneous treatment effect). As we discussed in section 5.5.1, much recent methodological work suggests heterogeneous treatment effects are uncommon in social science experiments. However, there are some situations where a heterogeneous treatment effect is strongly suggested by theory; in addition, continued work should evaluate the possibility of heterogeneous treatment effects as new topics and populations are studied. For these reasons, we discuss how to test for such an effect. Consider a linear regression model estimated to determine the effect of a treatment in a two-condition between-subjects experiment:

$$\hat{y} = \hat{\alpha} + \hat{\beta}_{treat} treat_i \quad (9)$$

Where $\hat{\beta}_{treat}$ is the estimated average treatment effect, which is simply the difference in the mean of the treatment from the control group. The treatment effect is homogenous if the value of $\hat{\beta}_{treat}$ is the same across participants with differing characteristics. E.g., if $\hat{\beta}_{treat}$ for women equals $\hat{\beta}_{treat}$ for men then the treatment effect is homogenous across these two groups. This can be tested straightforwardly by incorporating gender and an interaction (product) term in the model:

$$\hat{y} = \hat{\alpha} + \hat{\beta}_{treat} treat_i + \hat{\beta}_{woman} woman_i + \hat{\beta}_{treat*woman} treat_i * woman_i \quad (10)$$

Where the coefficient on the interaction (product) term $\hat{\beta}_{treat*woman}$ provides the test of the heterogeneous treatment effect for a linear model.²⁶ A non-significant interaction term indicates a homogenous treatment effect. In contrast, a significant interaction term suggests the treatment has a different effect for each group; i.e., there is a heterogeneous treatment effect. Continuous covariates, e.g., age, can be tested in the same manner with all usual best practices for testing interaction effects applying ([Brambor et al. 2006](#); [Mize 2019b](#)). We warn that all the usual considerations of statistical significance not necessarily indicating substantive importance apply equally to these tests. E.g., in large samples even trivially small heterogeneous treatment effects could be identified but may not indicate substantively meaningful differences in a treatment's effect.

²⁶ For a nonlinear model the test should be conducted in the predictions using the metric of most interest; e.g., the predicted probabilities for a binary logit or probit model ([Mize 2019a](#)).

6.5. Testing for cultural consensus

In an experimental study, cultural consensus indicates that all participants' ratings within a single experimental condition are similar. There can be cultural consensus in some conditions but not others (e.g., all participants rate men job applicants similarly but ratings of women job applicants vary). If there is cultural consensus across all conditions, there cannot be a heterogeneous treatment effect. However, a lack of cultural consensus *does not* necessarily indicate a heterogeneous treatment effect. In our experience, we see authors and reviewers often mistake a main effect of social location (e.g., age of participant) as indicating a heterogeneous treatment effect. E.g., if we fit a model and obtain estimates:

$$\hat{y} = \hat{\alpha} + \hat{\beta}_{treat} treat_i + \hat{\beta}_{age} age_i + \hat{\beta}_{treat \times age} treat_i * age_i \quad (11)$$

And if the $\hat{\beta}_{treat \times age}$ term is non-significant, we would drop that term from the model and refit the model as:

$$\hat{y} = \hat{\alpha} + \hat{\beta}_{treat} treat_i + \hat{\beta}_{age} age_i \quad (12)$$

If we find a positive significant effect of age, this indicates that older participants rate *all experimental conditions* higher than younger participants. Because $\hat{\beta}_{treat \times age}$ was non-significant in Equation (11), we already know that the treatment effect is homogenous across ages. The significant main effect of age does not harm our inference of $\hat{\beta}_{treat} treat_i$ from Equation (12). Here, age is simply another source of heterogeneity but is uncorrelated with the treatment.²⁷ Such main effects of social location are indeed the reason we conducted an experimental study in the first place: to use random assignment to control for these factors. Our personal experience matches the methodological literature: main effects of social location are common while heterogeneous treatment effects are quite rare (e.g., Coppock et al., 2020).

Sometimes, cultural consensus is of interest in and of itself. For example, multiple theories point to the importance of perceptions of culture and suggest these perceptions should be fairly consensual within a culture (i.e., even if people personally disagree, most people tend to agree on what “most people” think (Correll et al., 2017; Fiske et al., 2002; Heise 2010; Melamed et al., 2019)). For example, Mize and Manago (2018) tested whether any of 12 different demographic factors were associated with reports of cultural stereotypes of gender and sexuality and found broad cultural consensus, supporting stereotype content model predictions (Fiske et al., 2002).

6.6. Accounting for multilevel data structures

Within-subjects designs produce multilevel data structures as each participant responds to multiple experimental conditions. This clustering must be accounted for in any analysis to provide appropriate standard error estimates and significance tests. The methods for accounting for clustering in within-subjects experiments are the same as for any multilevel data: clustered standard errors and multilevel models are both popular solutions (see Luke 2019 for an accessible overview).

7. Continuing debates and future directions

While the experimental method has made considerable progress in recent years, there remain some difficult issues in studying certain topics experimentally. While some of these issues might represent unsolvable problems, we believe most of them provide important future directions for methodological research on experiments and for creative substantive work. We outline a few key areas of current and needed future research below.

7.1. Null findings, replication, and reproducibility

One obstacle faced by experimental researchers is null findings. When testing a hypothesis, which is the primary strength of experiments, there is a chance that the researcher fails to reject the null hypothesis (meaning, they do not find support for the hypothesis). The presence of null findings is not inherently bad. In fact, if a hypothesis is unsupported, this can provide researchers with important information about what is *not* responsible for variation in a dependent variable. In addition, tests of formal theory that return null results provide important information about how the theory should be reconsidered or changed. Within the current publication model, however, null findings are more difficult to publish; this leads to a “file-drawer problem” in which null findings are not published, but instead are filed-away (Malhotra 2021). The lack of published null findings is a problem for science because it fails to move scientific discourse forward. Hypothesis-testing requires both the falsification of and finding support for theories.

Within the social sciences, some disciplines have adopted more tools than others to address the file-drawer problem. In general, the adoption of transparent and reproducible research practices such as pre-registration/registered reports and open-access materials promotes the publication of a wider array of research findings (including null) (Fanelli 2010; Franco et al. 2014; Gerber and Malhotra 2008), replicability of existing research (Christensen et al. 2019; DeHaven 2017; Freese 2007; Freese and Peterson 2017; Logg and Dorison 2021; Simmons et al. 2021; van 't Veer and Giner-Sorolla 2016), and in turn, the development of theory and knowledge. We

²⁷ In this example, age would be a good candidate to add as a control variable to reduce the standard error estimate of $\hat{\beta}_{treat}$ though the estimate of $\hat{\beta}_{treat}$ should be unaffected by having age in the model or not (see Section 6.2).

agree that, in general, pre-registration and open-access materials are good for science (see Boudreau 2021 for recommendations specific to experimental studies). We also echo calls from other researchers for these initiatives to be implemented on a structural level, rather than individual. For example, research would benefit from journals and professional organizations adopting standards for pre-registration and open-access materials (Chambers and Tzavella 2022; Freese 2007; Präg et al. 2022).

7.2. Pairing multiple methods

Just as different methods have different benefits and limitations, different types of experiments also have strengths and weaknesses. Increasingly, researchers are pairing experiments with an additional method in the context of a single project or article. For example, audit studies are popular for their ability to test for effects in a population of interest without changing the context of the decision of interest (e.g., experimentally manipulated job applications sent to real employers; see Gaddis 2018). However, a limitation of audit studies is their “black box” nature where only the outcome is observed but the mechanisms are not. For this reason, audit studies are often paired with another method that can complement their strengths. For example, Rivera and Tilcsik (2016) use in-depth interviews to better understand their audit study results—discussing the intricacies of the hiring process with those making the hiring decisions. Similarly, Quadlin (2018) uses a post-hoc survey experiment to assess the mechanisms underlying the findings observed from an initial audit study. Correll, Benard, and Paik (2007) first conduct a lab experiment to test theoretical mechanisms and then assess the hypothesized effects in an audit study of employers—finding consistent effects. However, without the lab experiment and the ability to ask participants about their evaluation process the mechanisms would have been unclear.

Pairing multiple methods is not limited to audit studies but can be useful for any experimental study. The key is to consider the strengths and limitations of each method and try to pair methods with complementary strengths. In addition, as the Correll, Benard, and Paik (2007) example shows sometimes multiple types of experiments can profitably be used within a single project.

7.3. Cross-national comparative experimental designs

Modern advances in sampling platforms have expanded the possibilities of simultaneously collecting data across multiple populations of interest—e.g., across multiple countries. In addition, the cost of doing so is often no different than the cost of collecting data in a single population. Despite these possibilities, relatively few experiments have used a cross-national comparative design to take advantage of this. We think this is an area ripe for more study.

As a motivating example consider the research question “is there more prejudice against sexual minorities in country A compared to country B?”. While surveys are useful to assess this question, an experiment could assess actual prejudicial behavior and/or be designed to limit socially desirable responses (e.g., Coffman et al. 2016). For example, a survey experiment could be designed to assess sexual minority prejudice and then fielded in both countries simultaneously via one of the platforms discussed in Section 5.7. In addition, many more than two countries could be studied at the same time. Three such exemplary experimental projects using cross-national comparative designs are Pescosolido and colleagues work on the stigma of mental illness using vignette experiments (Krendl and Pescosolido 2020; Olafsdottir and Pescosolido 2011; Pescosolido et al., 2013), Lancee (2019) on hiring discrimination using audit studies, and Gërzhani (2020) examining status and gender using lab experiments.

While we are advocates for more cross-national comparative experimental research, we recognize some difficult issues presented by such work (see Pescosolido and Olafsdottir 2008 for a helpful discussion of many of these issues). For example, translating experimental manipulations and survey instruments into different languages can be difficult, time consuming, and expensive. In addition, ensuring complete equivalence across languages and cultures is challenging. Even when a single language can be used (e.g., English) it is worth considering whether the fluent speakers of that language in a given country represent the population of interest (e.g., are fluent English speakers residing in China of interest or the average Chinese resident?). Finally, issues of sampling discussed in Section 5 need to be considered for each country to ensure sampled participants best test the theoretical question of interest.

7.4. Incorporating biological measures

Traditionally, experiments have been used to assess behavioral outcomes or survey responses. Increasingly, biological measures are being used to assess mechanisms, moderators, and/or outcomes—expanding both the topics available for study experimentally and our understanding of classic questions. For example, Taylor (2016) uses salivary cortisol levels as a stress outcome measure to show that token situations are equally stressful for both advantaged and disadvantaged groups. As another example, Robinson, Clay-Warner, and colleagues (2012; 2015) use measures of facial temperature via infrared cameras to detect emotional experiences. For both examples, the biological measures should provide a more objective measure that is less prone to social desirability or other concerns.

In addition to using biological measures to assess outcomes, social scientists measure biological factors to assess moderators and mediators. For example, Willer et al. (2013) found that men whose masculinity is threatened react with hyper-masculine compensatory responses. A follow-up study showed that this reaction is strongest among men who have high testosterone levels. Finally, Kalkhoff et al. (2020) assess a fundamental mechanism in sociological theories of status using an electroencephalography

(EEG) measure. This provides a new answer to a “black box” question facing a classic sociological theory.

As these examples illustrate, biological measures can be profitably incorporated into experimental studies to better assess outcomes and/or to assess moderators and mediators otherwise difficult to measure.

7.5. Social desirability bias

Above we discussed the important issue of null findings. One common explanation for null findings is social desirability bias. That is, if a well-designed experiment does not support a priori theoretical predictions, then the presumed answer is that participants misrepresented their true beliefs. Several methods have been developed to combat social desirability though our assessment is that no definitive solution exists. Below, we briefly describe three popular methods focused on combatting social desirability and assess pros and cons. To end, we critique the common but potentially misguided explanation of social desirability bias to explain null or surprising effects.

7.5.1. List treatment experiments

In a list treatment experiment (AKA item count technique) a participant is not asked directly about a sensitive issue but is instead given a list of, e.g. five statements, and then asked to report *how many* they agree with. In a control condition, participants do the same but with the sensitive item removed to estimate the baseline agreement with the other four statements. For example, using a list experiment [Coffman et al. \(2016\)](#) find that survey-based reports of LGBT acceptance may be artificially high due to social desirability to report acceptance. Many deviations on the basic list treatment design exist (e.g., [Coffman et al., 2016](#); [Smallpage et al., 2022](#)). Despite some evidence in favor of list experiments, other research suggests list experiments are confusing to participants and induce a non-trivial amount of false positives (e.g., answering yes to a prejudicial belief one does not truly endorse). For example, [Kuhn and Vivyan \(2022\)](#) show that list experiments do reduce “strategic” misreporting but also increase accidental misreporting; whether a list experiment yields more accurate estimates than a traditional design depends on the level of strategic misreporting in the population on that issue.

7.5.2. Randomized response techniques

Randomized response techniques involve participants being randomly assigned to give their true answer or to answer in a directed way (e.g., say no regardless of what you truly think). The logic is that the experimenter will not know whether a given participant’s answer was true or directed—therefore the participant will feel comfortable answering truthfully (when directed to do so). Methodological work suggests the randomized response technique does not always work as intended both because it is confusing to participants and because it does not always inspire trust that answers are truly anonymous ([Coutts and Jann 2011](#)).

7.5.3. Social desirability scales

Social desirability scales attempt to quantify each participant’s propensity to provide socially desirable responses (e.g., [Stöber 2001](#)). This measure can then be interacted with the experimental treatment to assess heterogeneous treatment effects (see Section 6.4 for details). If the interaction effect is null this suggests social desirability does not affect the estimate of the treatment effect. If there is a heterogeneous treatment effect, it is possible to assess the treatment effect at low and high levels of social desirability (e.g., see [Mize 2019a](#) for potential analyses). The key limitation of this approach is that “propensity for socially desirable responses” is a vague concept that is difficult to assess. In addition, it is a measured rather than a manipulated factor so any analysis of its impact does not carry the evidential impact of causality that an experimental manipulation does.

7.5.4. Conclusions on experimental techniques to combat social desirability

Although social desirability bias is a commonly cited concern and explanation for null or surprising treatment effects, we believe the burden of proof lies on the researcher to convince readers that participants were not truthful. Occam’s razor suggests the most likely explanation is: participants in that study did not differ in how they responded to the treatment and/or did not respond in the way the theory would suggest. Especially with the increased use of anonymous online survey experiments, participants have little motivation to misrepresent their true beliefs. Instead, null and surprising findings may represent an important piece of evidence against existing theory and should be reported as such (see [Zigerell 2018](#) for one prominent example of selective reporting of experimental findings due, potentially, to the assumption of social desirability being the cause of unexpected null results).

8. Summary and conclusion

[Fig. 1](#) documents impressive growth in the use of experiments across the social sciences. We are encouraged by the embrace of experiments across fields from a niche status to a more mainstream tool. We also look forward to continued innovation and adoption of the method in areas that have previously not embraced experiments—and indeed one of our goals in this article is to encourage more widespread appreciation and adoption of experiments. Like any tool, experiments are better for some purposes than others, and have associated strengths and weaknesses. Because of the limitations, we hope the experimental method remains a tool and not the only tool used within a discipline, topic area, or theoretical paradigm. We agree with [Cohen \(2003\)](#) that:

“... a series of tests under differing conditions is more informative than a series of tests under the same conditions. Putting it crudely, two tests using different techniques—for example, an experiment and a survey—are likely to be more informative for a theory than two experimental tests. Replication under theoretically equivalent but concretely different circumstances provides a better test than even the most rigorously designed single study.” (p.11)

That is, fields and subdisciplines that limit themselves to the use of a single method limit the types of research questions they can address and also tie their knowledge to the limitations of the method. We believe the methodological pluralism of the social sciences is

one of its strengths and are excited to see experimental methods joining the cannon of widely-used approaches.

Acknowledgments

We thank the special issue Editors, Shawn Bauldry and Weihua An, for inviting us to contribute this article and the anonymous reviewers for their extremely helpful suggestions. We are grateful to Jane Sell, for sharing brilliant insights about theory and helpful comments on earlier drafts of the manuscript. We also thank Scott Feld for the many thoughtful conversations that inspired some of our thinking on sampling. Perhaps most of all, we thank Steve Benard, Jane Sell, Dawn Robinson, and Jody Clay-Warner who taught us experimental methods and inspired our passion for the subject. Finally, we thank Pam Morgan, Erin Straka, Hannah Regan, Chanterla Milner, Alyxis Smith, and Elaine Li for excellent research assistance.

Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.ssresearch.2022.102799>.

References

- Abascal, Maria, 2015. Us and them: black-white relations in the wake of hispanic population growth. *Am. Socio. Rev.* 80 (4), 789–813. <https://doi.org/10.1177/0003122415587313>.
- Anderson, Karl E., Kappas, Arvid, Conney, A.H., H. Leon Bradlow, Fishman, Jack, 1984. The influence of dietary protein and carbohydrate on the principal oxidative biotransformations of estradiol in normal subjects. *J. Clin. Endocrinol. Metab.* 59 (1), 103–107. <https://doi.org/10.1210/jcem-59-1-103>.
- Arechar, Antonio A., Gächter, Simon, Lucas, Molleman, 2018. Conducting interactive experiments online. *Exp. Econ.* 21 (1), 99–131. <https://doi.org/10.1007/s10683-017-9527-2>.
- Aronow, Peter M., Baron, Jonathon, Pinson, Lauren, 2019. A note on dropping experimental subjects who fail a manipulation check. *Polit. Anal.* 27 (4), 572–589.
- Asch, Solomon E., 1956. Studies of independence and conformity: I. A minority of one against a unanimous majority. *Psychol. Monogr.: General and Applied* 70 (9), 1–70. <https://doi.org/10.1037/h0093718>.
- Auspurg, Katrin, Hinz, Thomas, 2014. *Factorial Survey Experiments*. SAGE Publications.
- Auspurg, Katrin, Hinz, Thomas, Sauer, Carsten, 2017. Why should women get less? evidence on the gender pay gap from multifactorial survey experiments. *Am. Socio. Rev.* 82 (1), 179–210. <https://doi.org/10.1177/0003122416683393>.
- Austin, Peter C., 2009. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Stat. Med.* 28 (25), 3083–3107.
- Bail, Christopher A., Argyle, Lisa P., Brown, Taylor W., Bumpus, John P., Chen, Haohan, Hunzaker, M. B., Fallin, Lee, Jaemin, Mann, Marcus, Merhout, Friedolin, Alexander, Volfovsky, 2018. Exposure to opposing views on social media can increase political polarization. *Proc. Natl. Acad. Sci. USA* 115 (37), 9216–9221. <https://doi.org/10.1073/pnas.1804840115>.
- Baldassarri, Delia, Abascal, Maria, 2017. Field experiments across the social sciences. *Annu. Rev. Sociol.* 43, 41–73.
- Bansak, Kirk, Hainmueller, Jens, Hopkins, Daniel J., Yamamoto, Teppei, 2021. *Conjoint Survey Experiments*. Pp. 19–41. In: Druckman, J.N., Green, D.P. (Eds.), *Advances in Experimental Political Science*. Cambridge University Press, Cambridge.
- Benard, S., S J Correll., 2010. Normative discrimination and the motherhood penalty. *Gend. Soc.* 24 (5), 616–646. <https://doi.org/10.1177/0891243210383142>.
- Benard, Stephen, Bianca Manago, Annie Russian, and Youngjoo Cha. Forthcoming. “Mapping the content of asian stereotypes in the United States: intersections with ethnicity, gender, income, and birthplace.” *Soc. Psychol. Q.*
- Boudreau, Cheryl, 2021. Transparency in experimental research. In: Druckman, J.N., Green, D.P. (Eds.), –53 in *Advances In Experimental Political Science*. Cambridge University Press, Cambridge, p. 339.
- Brambor, Thomas, Clark, William Roberts, Golder, Matt, 2006. Understanding interaction models: improving empirical analyses. *Polit. Anal.* 63–82.
- Castilla, E.J., Stephen Benard, 2010. The paradox of meritocracy in organizations. *Adm. Sci. Q.* 55 (4), 543–576.
- Chambers, Christopher D., Loukia, Tzavella, 2022. The past, present and future of registered reports. *Nat. Human Behav.* 6 (1), 29–42. <https://doi.org/10.1038/s41562-021-01193-7>.
- Christensen, Garret, Freese, Jeremy, Miguel, Edward Andrew, 2019. *Transparent and Reproducible Social Science Research: How to Do Open Science*. University of California Press, Oakland, California.
- Clay-Warner, Jody, Robinson, Dawn T., 2015. Infrared thermography as a measure of emotion response. *Emotion Review* 7 (2), 157–162.
- Cleveland, W.S., 1994. *The Elements of Graphing Data*. Hobart Press, Summit, NJ.
- Coffman, Katherine B., Coffman, Lucas C., Keith M. Marzilli Ericson, 2016. The size of the LGBT population and the magnitude of antigay sentiment are substantially underestimated. *Manag. Sci.* <https://doi.org/10.1287/mnsc.2016.2503>.
- Cohen, Bernard P., 1980. The conditional nature of scientific knowledge. In: Freese, L. (Ed.), *Theoretical Methods in Sociology: Seven Essays*. University of Pittsburgh Press, Pittsburgh.
- Cohen, Bernard P., 1989. *Developing Sociological Knowledge: Theory and Method*, second ed. Chicago, IL.
- Cohen, Bernard P., 2003. Creating , testing , and applying social psychological theories. *Soc. Psychol. Q.* 66 (1), 5–16.
- Cohen, Jacob, 1988. *Statistical Power Analysis for the Behavioral Sciences*. Routledge.
- Collett, Jessica L., Ellen Childs., 2011. Minding the gap: meaning, affect, and the potential shortcomings of vignettes. *Soc. Sci. Res.* 40 (2), 513–522. <https://doi.org/10.1016/j.ssresearch.2010.08.008>.
- Coppock, Alexander, 2018. Generalizing from survey experiments conducted on mechanical Turk: a replication approach. *Political Science Research and Methods*. <https://doi.org/10.1017/psrm.2018.10>.
- Coppock, Alexander, Hill, Seth J., Lynn, Vavreck, 2020. The small effects of political advertising are small regardless of context, message, sender, or receiver: evidence from 59 real-time randomized experiments. *Sci. Adv.* 7.
- Coppock, Alexander, Leeper, Thomas J., Mullinix, Kevin J., 2018. Generalizability of heterogeneous treatment effect estimates across samples. *Proc. Natl. Acad. Sci. USA* 115 (49), 12441–12446. <https://doi.org/10.1073/pnas.1808083115>.
- Correll, Shelley J., Stephen Benard, and In Paik, 2007. Getting a job: is there a motherhood penalty? *Am. J. Sociol.* 112 (5), 1297–1339.
- Correll, Shelley J., Ridgeway, Cecilia L., Zuckerman, Ezra W., Jank, Sharon, Jordan-bloch, Sara, Nakagawa, Sandra, 2017. It's the conventional thought that counts : how third-order inference produces status advantage. *Am. Socio. Rev.* <https://doi.org/10.1177/0003122417691503>.
- Coutts, Elisabeth, Ben, Jann, 2011. Sensitive Questions in Online Surveys : Experimental Results for the Randomized Response Technique (RRT) and the Unmatched Count Technique. <https://doi.org/10.1177/0049124110390768>. UCT.

- Crump, Matthew J.C., McDonnell, John V., Gureckis, Todd M., 2013. Evaluating amazon's mechanical Turk as a tool for experimental behavioral research" edited by S. Gilbert. *PLoS One* 8 (3), e57410. <https://doi.org/10.1371/journal.pone.0057410>.
- Deaton, Angus, Cartwright, Nancy, 2018. Understanding and misunderstanding randomized controlled trials. *Soc. Sci. Med.* 210, 2–21. <https://doi.org/10.1016/j.socscimed.2017.12.005>.
- DeHaven, Alexander C., 2017. PreRegistration: A Plan, Not a Prison. Retrieved (<https://www.cos.io/blog/preregistration-plan-not-prison>).
- Doan, Long, Loehr, Annalise, Lisa, R., Miller, 2014. Formal rights and informal privileges for same-sex couples: evidence from a national survey experiment. *Am. Socio. Rev.* 79 (6), 1172–1195. <https://doi.org/10.1177/0003122414555886>.
- Doyle, Joshua, 2021. The effect of cultural trust on cooperation in two behavioral experiments. *Soc. Psychol. Q.* 84 (3), 246–266. <https://doi.org/10.1177/01902752211032134>.
- Dresser, Rebecca, 1992. Wanted single, white male for medical research. *Hastings Cent. Rep.* 22 (1), 24. <https://doi.org/10.2307/3562720>.
- Druckman, James N., Green, Donald P., 2021a. A new era of experimental political science. In: Druckman, J., Green, D.P. (Eds.), 1–16 in *Advances In Experimental Political Science*. Cambridge University Press.
- Druckman, James N., Green, Donald P., 2021b. *Advances in Experimental Political Science*. Cambridge University Press.
- Druckman, James N., Green, Donald P., Kuklinski, James H., Arthur, Lupia, 2006. The growth and development of experimental research in political science. *Am. Polit. Sci. Rev.* 100 (4), 627. <https://doi.org/10.1017/S0003055406062514>.
- Druckman, James N., Kam, Cindy D., 2011. Students as experimental participants: a defense of the 'narrow data base.' pp. 41–57. In: *Handbook of Experimental Political Science*. Cambridge University Press.
- Falk, Armin, Heckman, James J., 2009. Lab experiments are a major source of knowledge in the social sciences. *Science* 326 (5952), 535–538.
- Fanelli, Daniele, 2010. 'Positive' results increase down the hierarchy of the sciences" edited by E. Scalas. *PLoS One* 5 (4), e10068. <https://doi.org/10.1371/journal.pone.0010068>.
- Fiske, Susan T., Cuddy, Amy J.C., Glick, Peter, Xu, Jun, 2002. A model of (often mixed) stereotype content: competence and warmth respectively follow from perceived status and competition. *J. Pers. Soc. Psychol.* 82 (6), 878–902. <https://doi.org/10.1037/0022-3514.82.6.878>.
- Foschi, Martha, 1997. On scope conditions. *Small Group Res.* 28 (4), 535–555. <https://doi.org/10.1177/1046496497284004>.
- Foschi, Martha, 2014. Hypotheses, operationalizations, and manipulation checks. In *Laboratory experiments in the social sciences* 247–268. Elsevier.
- Franco, Annie, Malhotra, Neil, Simonovits, Gabor, 2014. Publication bias in the social sciences: unlocking the file drawer. *Science* 345 (6203), 1502–1505. <https://doi.org/10.1126/science.1255484>.
- Freeman, Jonathan B., Penner, Andrew M., Saperstein, Aliya, Scheutz, Matthias, Ambady, Nalini, 2011. Looking the part: social status cues shape race perception" edited by S. Gilbert. *PLoS One* 6 (9), e25107. <https://doi.org/10.1371/journal.pone.0025107>.
- Freese, Jeremy, 2007. Replication standards for quantitative social science: why not sociology? *Socio. Methods Res.* 36 (2), 153–172. <https://doi.org/10.1177/0049124107306659>.
- Freese, Jeremy, Peterson, David, 2017. Replication in social science. *Annu. Rev. Sociol.* 43, 147–165.
- Freese, L., 1980. The Problem of Cumulative Knowledge." Pp. 13–69 in *Theoretical Methods in Sociology: Seven Essays*. University of Pittsburgh Press, Pittsburgh.
- Gaddis, S. Michael, 2015. Discrimination in the credential society: an audit study of race and college selectivity in the labor market. *Soc. Forces* 93 (4), 1451–1459. <https://doi.org/10.1093/sf/sou111>.
- Gaddis, S. Michael, 2017. How black are lakisha and jamal? Racial perceptions from names used in correspondence audit studies. *Sociological Science* 4, 469–489. <https://doi.org/10.15195/v4.a19>.
- Gaddis, S. Michael, 2018. Audit Studies: behind the Scenes with Theory, Method, and Nuance.
- Gallo, Amy, 2017. A refresher on A/B testing. *Harv. Bus. Rev.* June 28.
- Gelman, Andrew, 2008. Scaling regression inputs by dividing by two standard deviations. *Stat. Med.* 27 (15), 2865–2873.
- Gerber, Alan S., Neil Malhotra., 2008. Publication bias in empirical sociological research: do arbitrary significance levels distort published results? *Socio. Methods Res.* 37 (1), 3–30. <https://doi.org/10.1177/0049124108318973>.
- Gereke, Johanna, Gërkhani, Klarita, 2019. Experimental economics and experimental sociology. In: *Oxford Research Encyclopedia of Economics and Finance*. Oxford University Press.
- Gërkhani, Klarita, 2020. Status ranking and gender inequality: a cross-country experimental comparison. *Res. Soc. Stratif. Mobil.* 65, 100474 <https://doi.org/10.1016/j.rssm.2020.100474>.
- Gërkhani, Klarita, Miller, Luis, 2022. Experimental sociology. In: *Handbook of Sociological Science: Contributions to Rigorous Sociology*. Edwar Elgar, pp. 309–323.
- Giamattei, Marcus, Johann Graf Lambsdorff, 2019. ClassEx — an online tool for lab-in-the-field experiments with smartphones. *Journal of Behavioral and Experimental Finance* 22, 223–231. <https://doi.org/10.1016/j.jbef.2019.04.008>.
- Green, Paul E., Krieger, Abba M., Yoram Wind., 2001. Thirty years of conjoint analysis: reflections and prospects. *Interfaces* 31 (3), S56–S73.
- Greifer, Noah, 2022. Covariate Balance Tables and Plots: A Guide to the Cobalt Package.
- Greiner, D. James, Rubin, Donald B., 2011. Causal effects of perceived. *IMMUTABLE CHARACTERISTICS* 93 (August), 775–785.
- Hainmueller, Jens, Hangartner, Dominik, Yamamoto, Teppei, 2015. Validating vignette and conjoint survey experiments against real-world behavior. *Proc. Natl. Acad. Sci. USA* 112 (8), 2395–2400. <https://doi.org/10.1073/pnas.1416587112>.
- Heise, David R., 2010. Surveying Cultures: Discovering Shared Conceptions and Sentiments.
- Holland, Paul W., 1986. Statistics and causal inference. *J. Am. Stat. Assoc.* 81, 945–960.
- Horiuchi, Yusaku, Markovich, Zachary, Yamamoto, Teppei, 2021. Does conjoint analysis mitigate social desirability bias? *Polit. Anal.* 1–15. <https://doi.org/10.1017/pan.2021.30>.
- Horton, John J., Rand, David G., Zeckhauser, Richard J., 2011. The online laboratory: conducting experiments in a real labor market. *Exp. Econ.* 14 (3), 399–425. <https://doi.org/10.1007/s10683-011-9273-9>.
- Imbens, Guido W., Rubin, Donald B., 2015. *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge University Press.
- Jackson, Michelle, Cox, D.R., 2013. The principles of experimental design and their application in sociology. *Annu. Rev. Sociol.* 39 (1), 27–49. <https://doi.org/10.1146/annurev-soc-071811-145443>.
- Jaschik, Scott, 1990. Report Says NIH Ignores Own Rules on Including Women in its Research, vol. 27. *The Chronicle of Higher Education*, p. 3. June.
- Jones, Jeffrey M., 2022. LGBT Identification in U.S. Ticks up to 7.1%. Gallup.
- Kalkhoff, Will, Melamed, David, Pollock, Josh, Miller, Brennan, Overton, Jon, Pfeiffer, Matthew, 2020. Cracking the black box: capturing the role of expectation states in status processes. *Soc. Psychol. Q.* 83 (1), 26–48.
- Keville, Terri K., 1993. The invisible woman: gender bias in medical research. *Wom. Right Law Rep.* 15 (2–3), 123–142.
- Kincaid, Reilly, 2022. Status, masculinity, and femininity at the intersection of age and gender. *Soc. Sci. Res.* 102695 <https://doi.org/10.1016/j.ssrresearch.2022.102695>.
- Klein, Richard A., Ratliff, Kate A., Vianello, Michelangelo, Adams Jr., Reginald B., Bahník, Štěpán, Bernstein, Michael J., Bocian, Konrad, Brandt, Mark J., Brooks, Beach, Claudia Chloe Brumbaugh, 2014. Investigating variation in replicability: a 'many labs' replication project. *Soc. Psychol.* 45 (3), 142.
- Klein, Richard A., Vianello, Michelangelo, Hasselman, Fred, Adams, Byron G., Reginald B. Adams Jr, Alper, Sinan, Aveyard, Mark, Axt, Jordan R., Babalola, Mayowa T., Bahník, Štěpán, 2018. Many labs 2: investigating variation in replicability across samples and settings. *Advances in Methods and Practices in Psychological Science* 1 (4), 443–490.
- Kotzian, Peter, Stoerber, Thomas, Hoos, Florian, Barbara, E., Weissenberger, 2020. To Be or not to Be in the sample? On using manipulation checks in experimental accounting research. *Account. Res. J.* 33 (3), 469–482.
- Kramer, Adam D.I., Guillory, Jamie E., Hancock, Jeffrey T., 2014. Experimental evidence of massive-scale emotional contagion through social networks. *Proc. Natl. Acad. Sci. USA* 111 (24), 8788–8790. <https://doi.org/10.1073/pnas.1320040111>.

- Kraut, Robert, Olson, Judith, Banaji, Mahzarin, Amy, Bruckman, Cohen, Jeffrey, Couper, Mick, 2004. Psychological research online: report of board of scientific affairs' advisory group on the conduct of research on the internet. *Am. Psychol.* 59 (2), 105–117. <https://doi.org/10.1037/0003-066X.59.2.105>.
- Krendl, Anne C., Pescosolido, Bernice A., 2020. Countries and cultural differences in the stigma of mental illness: the east–west divide. *J. Cross Cult. Psychol.* 51 (2), 149–167. <https://doi.org/10.1177/0022022119901297>.
- Krupnikov, Yanna, Nam, H. Hannah, Style, Hillary, 2021. Convenience samples in political science experiments. In: Druckman, J.N., Green, D.P. (Eds.), –83 in *Advances In Experimental Political Science*. Cambridge University Press, Cambridge, p. 165.
- Kuhn, Patrick M., Nick, Vivyan, 2022. The misreporting trade-off between list experiments and direct questions in practice: partition validation evidence from two countries. *Polit. Anal.* 30 (3), 381–402.
- Lancee, Bram, et al., 2019. The gemm study: a cross-national harmonized field experiment on labour market discrimination: technical report. Available at SSRN 3398191.
- Levay, K.E., Freese, J., Druckman, J.N., 2016. The demographic and political composition of mechanical Turk samples. *Sage Open* 6 (1), 2158244016636433. <https://doi.org/10.1177/2158244016636433>.
- Levine, Mark, Amy Prosser, 2005. Identity and emergency intervention: how social group membership and inclusiveness of group boundaries shape helping behavior. *Personality and Social \ldots*. <https://doi.org/10.1177/0146167204271651>.
- Litman, Leib, Robinson, Jonathan, Abberbock, Tzvi, 2017. TurkPrime.Com: a versatile crowdsourcing data acquisition platform for the behavioral sciences. *Behavior Research Methods* 49 (2), 433–442. <https://doi.org/10.3758/s13428-016-0727-z>.
- Logg, Jennifer M., Dorison, Charles A., 2021. Pre-registration: weighing costs and benefits for researchers. *Organizational Behavior and Human Decision Processes* 167, 18–27. <https://doi.org/10.1016/j.obhdp.2021.05.006>.
- Lohr, Sharon L., 2021. Sampling: Design and Analysis. Chapman and Hall/CRC.
- Long, J. Scott, 1997. Regression Models for Categorical and Limited Dependent Variables, vol. 7. Sage, Thousand Oaks, CA.
- Long, J. Scott, Freese, Jeremy, 2014. Regression Models for Categorical Dependent Variables Using Stata, third ed. Stata Press, College Station, Texas.
- Lucas, Jeffrey W., 2003. Theory-testing, generalization, and the problem of external validity. *Sociological Theory* 21 (3), 236–253. <https://doi.org/10.1111/1467-9558.00187>.
- Luke, Douglas, A., 2019. Multilevel Modeling, vol. 143. Sage publications.
- Malhotra, Neil, 2021. Threats to the scientific credibility of experiments: publication bias and P-hacking. In: Druckman, J.N., Green, D.P. (Eds.), –68 in *Advances In Experimental Political Science*. Cambridge University Press, Cambridge, p. 354.
- Manago, Bianca, 2020. The translation of status characteristics and expectation states theory to survey methods. In: Brenner, P.S. (Ed.), 87–112 in *Understanding Survey Methodology*, vol. 4. Springer.
- Manago, Bianca, Mize, Trenton D., 2022. The status and stigma consequences of mental illness labels, deviant behavior, and fear. *Social Science Research* 105, 102690. <https://doi.org/10.1016/j.ssresearch.2021.102690>.
- Manago, Bianca, Mize, Trenton D., Long, Doan, 2021. Can you really study an army on the internet? Comparing how status tasks perform in the laboratory and online settings. *Sociological Methodology*. <https://doi.org/10.1177/00811750211014242>.
- Markovsky, Barry, 1995. The structure of theories. In: Group Processes: Sociological Analysis. Nelson-Hall, Chicago, IL, pp. 3–24.
- Melamed, David, Harrell, Ashley, Simpson, Brent, 2018. Cooperation, clustering, and assortative mixing in dynamic networks. *Proceedings of the National Academy of Sciences* 115 (5), 951–956. <https://doi.org/10.1073/pnas.1715357115>.
- Melamed, David, Munn, Christopher W., Barry, Leanne, Montgomery, Bradley, Oneya, F., Okuwobi, 2019. Status characteristics, implicit bias, and the production of racial inequality. *American Sociological Review*. <https://doi.org/10.1177/0003122419879101>.
- Melamed, David, Savage, Scott V., 2016. Status, faction sizes, and social influence: testing the theoretical mechanism. *American Journal of Sociology* 122 (1), 201–232. <https://doi.org/10.1086/686943>.
- Mize, Trenton D., 2019a. Doing gender by criticizing leaders: public and private displays of status. *Social Problems* 66 (1). <https://doi.org/10.1093/socpro/spx032>.
- Mize, Trenton D., 2019b. Best practices for estimating, interpreting, and presenting nonlinear interaction effects. *Sociological Science* 6, 81–117. <https://doi.org/10.15195/v6.a4>.
- Mize Trenton D. n.d. “Balanceplot: A Stata Command to Plot Imbalance across Groups”.
- Mize, Trenton D., Bianca Manago., 2018. The stereotype content of sexual orientation. *Social Currents* 5 (5), 458–478. <https://doi.org/10.1177/2329496518761999>.
- Mook, Douglas, G., 1983. In defense of external invalidity. *American Psychologist* 38 (4), 379–387. <https://doi.org/10.1037/0003-066X.38.4.379>.
- Morgan, Stephen L., Christopher Winship., 2014. Counterfactuals and Causal Inference. Cambridge University Press.
- Mullinix, Kevin J., Leeper, Thomas J., Druckman, James N., Jeremy Freese., 2015. The generalizability of survey experiments. *Journal of Experimental Political Science* 2 (2), 109–138. <https://doi.org/10.1017/XPS.2015.19>.
- Mutz, Diana C., 2011. Population-Based Survey Experiments. Princeton University Press.
- Mutz, Diana C., 2021. Improving experimental treatments in political science. *Advances in Experimental Political Science* 219.
- Mutz, Diana C., Pemantle, Robin, Philip Pham., 2019. The perils of balance testing in experimental design: messy analyses of clean data. *American Statistician* 73 (1), 32–42. <https://doi.org/10.1080/00031305.2017.1322143>.
- Olafsdottir, Sigrun, Pescosolido, Bernice A., 2011. Constructing illness: how the public in eight western nations respond to a clinical description of ‘schizophrenia. *Social Science & Medicine* 73 (6), 929–938.
- Oppenheimer, Daniel M., Meyvis, Tom, Davidenko, Nicolas, 2009. Instructional manipulation checks: detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology* 45 (4), 867–872.
- Pager, Devah, Quillian, Lincoln, 2005. Walking the talk? What employers say versus what they do. *American Sociological Review* 70 (3), 355–380. <https://doi.org/10.1177/000312240507000301>.
- Pearl, Judea, 2009. Causality. Cambridge University Press.
- Pedulla, David S., 2014. The positive consequences of negative stereotypes: race, sexual orientation, and the job application process. *Social Psychology Quarterly* 77 (1), 75–94. <https://doi.org/10.1177/0190272513506229>.
- Pescosolido, Bernice A., Medina, Tait R., Martin, Jack K., Scott Long, J., 2013. The ‘backbone’ of stigma: identifying the global core of public prejudice associated with mental illness. *American Journal of Public Health* 103 (5), 853–860.
- Pescosolido, Bernice A., Sigrun Olafsdottir., 2008. The logistics of survey implementation in a comparative study of mental illness: issues and resolutions in translation across cultural boundaries. In: International Conference on Survey Methods in Multinational, Multiregional, and Multicultural Contexts.
- Peterson, Erik, Westwood, Sean J., Iyengar, Shanto, 2021. Beyond attitudes: incorporating measures of behavior in survey experiments. In: Druckman, J.N., Green, D. P. (Eds.), 239–54 in *Advances In Experimental Political Science*. Cambridge University Press, Cambridge.
- Petts, Richard J., Mize, Trenton D., Gayle Kaufman., 2022. Organizational policies, workplace culture, and perceived job commitment of mothers and fathers who take parental leave. *Social Science Research* 103, 102651. <https://doi.org/10.1016/j.ssresearch.2021.102651>.
- Popper, Karl R., 1959. The Logic of Scientific Discovery, second ed. Basic Books, New York.
- Präg, Patrick, Ersanilli, Evelyn, Gugushvili, Alexi, 2022. An invitation to submit. *European Societies* 24 (1), 1–6. <https://doi.org/10.1080/14616696.2022.2029131>.
- Quadlin, Natasha, 2018. The mark of a woman's record: gender and academic performance in hiring. *American Sociological Review* 83 (2), 331–360. <https://doi.org/10.1177/0003122418762291>.
- Reips, Ulf-Dietrich, 2000. The web experiment method. In: Psychological Experiments on the Internet. Elsevier, pp. 89–117.
- Ridgeway, Cecilia L., 2019. Status: Why Is It Everywhere? Why Does It Matter? Russell Sage Foundation.
- Ridgeway, Cecilia L., Correll, Shelley J., 2006. Consensus and the creation of status beliefs. *Social Forces* 85 (1), 431–453. <https://doi.org/10.1353/sof.2006.0139>.
- Rivera, Lauren A., Andras Tilcsik., 2016. Class Advantage, commitment penalty: the gendered effect of social class signals in an elite labor market. *American Sociological Review* 1–68. <https://doi.org/10.1177/0003122416668154>.

- Robinaugh, Donald J., Haslbeck, Jonas M.B., Ryan, Oisín, Fried, Eiko I., Waldorp, Lourens J., 2021. Invisible hands and fine calipers: a call to use formal theory as a toolkit for theory construction. *Perspectives on Psychological Science* 16 (4), 19.
- Robinson, Dawn T., clay-warner, Jody, Moore, christopher D., everett, tiffani, Watts, alexander, Tucker, traci N., Thai, Chi, 2012. Toward an unobtrusive measure of emotion during interaction: thermal imaging techniques. *Advances in Group Processes* 29 (2012), 225–256. [https://doi.org/10.1108/S0882-6145\(2012\)0000029011](https://doi.org/10.1108/S0882-6145(2012)0000029011).
- Rosenbaum, Paul R., Rubin, Donald B., 1985. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician* 39 (1), 33–38.
- Salganik, Matthew J., Peter Sheridan Dodds, Watts, Duncan J., 2006. Experimental study of inequality and unpredictability in an artificial cultural market. *Science* 311 (5762), 854–856. <https://doi.org/10.1126/science.1121066>.
- Salganik, Matthew J., Duncan J. Watts, 2008. Leading the herd astray: an experimental study of self-fulfilling prophecies in an artificial cultural market. *Social Psychology Quarterly* 71 (4), 338–355. <https://doi.org/10.1177/019027250807100404>.
- Salsburg, David., 2001. *The Lady Tasting Tea: How Statistics Revolutionized Science in the Twentieth Century*. Macmillan.
- Sanders, Michael, 2017. Social influences on charitable giving in the workplace. *Journal of Behavioral and Experimental Economics* 66, 129–136. <https://doi.org/10.1016/j.jsocec.2015.12.004>.
- Saperstein, Aliya, Penner, Andrew M., 2014. Beyond the looking glass: exploring fluidity in racial self-identification and interviewer classification. *Sociological Perspectives* 57 (2), 186–207. <https://doi.org/10.1177/0731121414523732>.
- Sell, Jane, 2018. Definitions and the development of theory in social psychology. *Social Psychology Quarterly* 81 (1), 8–22. <https://doi.org/10.1177/0190272518755335>.
- Sell, Jane, Martin, Michael W., 1983. An acultural perspective on experimental social psychology. *Personality and Social Psychology Bulletin* 9 (3), 345–350. <https://doi.org/10.1177/0146167283093003>.
- Shadish, William, Cook, Thomas D., Campbell, Donald T., 2002. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Houghton Mifflin, Boston, MA.
- Simmons, Joseph P., Nelson, Leif, Simonsohn, Uri, 2021. Pre-registration: why and how. *Journal of Consumer Psychology* 31 (1), 151–162. <https://doi.org/10.1002/jcpsy.1208>.
- Simpson, Brent, Harrell, Ashley, Melamed, David, Heiserman, Nicholas, Daniela, V., Negraia, 2018. The roots of reciprocity: gratitude and reputation in generalized exchange systems. *American Sociological Review* 83 (1), 88–110. <https://doi.org/10.1177/0003122417747290>.
- Smallpage, Steven M., Enders, Adam M., Drochon, Hugo, Uscinski, Joseph E., 2022. The impact of social desirability bias on conspiracy belief measurement across cultures. *Political Science Research and Methods* 1–15. <https://doi.org/10.1017/psrm.2022.1>.
- Stepanikova, Irena, 2012. Racial-ethnic biases, time pressure, and medical decisions. *Journal of Health and Social Behavior*. <https://doi.org/10.1177/0022146512445807>.
- Stöber, Joachim, 2001. The social desirability scale-17 (SDS-17). *European Journal of Psychological Assessment* 17 (3), 222–232. <https://doi.org/10.1027/1015-5759.17.3.222>.
- Su, Dan, Peter, M., Steiner, 2020. An evaluation of experimental designs for constructing vignette sets in factorial surveys. *Sociological Methods & Research*, 004912411774642. <https://doi.org/10.1177/0049124117746427>.
- Suri, Siddharth, Watts, Duncan J., 2011. Cooperation and contagion in web-based, networked public goods experiments. *PLoS ONE* 6 (3), e16836. <https://doi.org/10.1371/journal.pone.0016836>.
- Taylor, Catherine J., 2016. ‘Relational by nature’? Men and women do not differ in physiological response to social stressors faced by token women. *American Journal of Sociology* 122 (1), 49–89.
- Thye, Shane, 2014a. Logical and philosophical foundations of experimental research in the social sciences. In: Webster, M., Sell, J. (Eds.), –82 in *Laboratory Experiments in the Social Sciences*. Academic Press, p. 53.
- Thye, Shane R., 2014b. *Logical and Philosophical Foundations of Experimental Research in the Social Sciences*, second ed. Elsevier Inc.
- Tinkler, Justine, Zhao, Jun, Li, Yan, Cecilia, L., Ridgeway, 2019. Honorary Whites ? Asian American Women and the Dominance Penalty. <https://doi.org/10.1177/2378023119836000>.
- Vaisey, Stephen, 2014. The “ Attitudinal Fallacy ” Is a Fallacy : Why We Need Many Methods to Study Culture. <https://doi.org/10.1177/0049124114523395>.
- van ’t Veer, Anna, Elisabeth, Giner-Sorolla, Roger, 2016. Pre-registration in social psychology—a discussion and suggested template. *Journal of Experimental Social Psychology* 67, 2–12. <https://doi.org/10.1016/j.jesp.2016.03.004>.
- Walker, Henry A., Bernard P. Cohen., 1985. Scope statements: imperatives for evaluating theory. *American Sociological Review* 50 (3), 288. <https://doi.org/10.2307/2095540>.
- Wasserstein, Ronald L., Schirm, Allen L., Lazar, Nicole A., 2019. Moving to a world beyond ‘p < 0.05’. *American Statistician* 73 (Suppl. 1), 1–19. <https://doi.org/10.1080/00031305.2019.1583913>.
- Weber, Max, 1922. *Economy and Society: an Outline of Interpretive Sociology*. Univ of California Press.
- Webster, Murray, Kervin, Jon B., 1971. Artificiality in experimental sociology. *Canadian Review of Sociology and Anthropology* 8, 263–273.
- Webster, Murray, Jane Sell, 2007. Theory and experimentation in the social sciences. In: *The SAGE Handbook Of Social Science Methodology*. 1 Oliver’s Yard, 55 City Road. SAGE Publications Ltd, London England EC1Y 1SP United Kingdom, pp. 192–211.
- Webster, Murray, Jane Sell, 2014. *Laboratory Experiments in the Social Sciences*, second ed. Elsevier, London.
- Weinberg, Jill, Freese, Jeremy, McElhattan, David, 2014. Comparing data characteristics and results of an online factorial survey between a population-based and a crowdsourced-recruited sample. *Sociological Science* 1, 292–310. <https://doi.org/10.15195/v1.a19>.
- Weiss, Robert S., 1995. *Learning from Strangers: the Art and Method of Qualitative Interview Studies*. Simon and Schuster.
- Willer, Robb, Ko, Kuwabara, Macy, Michael W., 2009. The false enforcement of unpopular norms. *American Journal of Sociology* 115 (2), 451–490. <https://doi.org/10.1086/599250>.
- Willer, Robb, , Christabel L. Rogalin Rogalin, Conlon, Bridget, Michael, T., Wojnowicz, 2013. Overdoing gender: a test of the masculine overcompensation thesis. *American Journal of Sociology* 980–1022.
- Zelditch Jr., Morris, 1969. Can you really study an army in the laboratory?. In: *A Sociological Reader on Complex Organizations*, pp. 528–539.
- Zelditch, Morris, 2014. Laboratory experiments in sociology. In: Webster, M., Sell, J. (Eds.), 183–97 in *Laboratory Experiments In the Social Sciences*. Elsevier, Amsterdam, Netherlands.
- Zelditch Jr., Morris, 2007. Laboratory experiments in sociology. In: Webster Jr., M., Sell, J. (Eds.), –30 in *Laboratory Experiments In the Social Sciences*. Elsevier, Oxford, England, p. 517.
- Zigerell, L.J., 2018. Black and white discrimination in the United States: evidence from an archive of survey experiment studies. *Research & Politics* 5 (1), 205316801775386. <https://doi.org/10.1177/2053168017753862>.