How often does random assignment fail? Estimates and recommendations

Matthew H. Goldberg

Yale Program on Climate Change Communication

Yale University

This article is now published in the *Journal of Environmental Psychology*. Please cite as:

Goldberg, M. H. (2019). How often does random assignment fail? Estimates and

recommendations. *Journal of Environmental Psychology*, doi:10.1016/j.jenvp.2019.101351

Abstract

A fundamental goal of the scientific process is to make causal inferences. Random assignment to experimental conditions has been taken to be a gold-standard technique for establishing causality. Despite this, it is unclear how often random assignment fails to eliminate non-trivial differences between experimental conditions. Further, it is unknown *to what extent* larger sample sizes mitigates this issue. Chance differences between experimental conditions may be especially important when investigating topics that are highly sample-dependent, such as climate change and other politicized issues. Three studies examine simulated data (Study 1), three real datasets from original environmental psychology experiments (Study 2), and one nationally-representative dataset (Study 3) and find that differences between conditions that remain after random assignment are surprisingly common for sample sizes typical of social psychological scientific experiments. Methods and practices for identifying and mitigating such differences are discussed, and point to implications that are especially relevant to experiments in social and environmental psychology.

Keywords: random assignment; randomization; confounding; validity

How often does random assignment fail? Estimates and recommendations

How do we best communicate the threat of climate change? Does this education program improve science literacy? Answering questions like these requires causal inference. The most effective method that enables causal inference is random assignment to conditions (Bloom, 2006; Fisher, 1925; Fisher, 1937; Gerber & Green, 2008; Rubin, 1974; Shadish, Cook, & Campbell, 2002). It is well known that random assignment lends greater confidence to causal inferences as sample size gets larger (e.g., Bloom, 2006). However, at commonly used sample sizes in psychological science, it is unclear how often random assignment fails to mitigate differences between conditions that might explain study results. Additionally, even given larger sample sizes, it is unknown *how much larger* is large enough (Deaton & Cartwright, 2018). The aim of this article is to answer these questions using both simulated and real participant data.

Causality

Before answering this question, first it is necessary to define causality and articulate a theoretical framework for it. A *cause* is “that which gives rise to any action, phenomenon, or condition” (Oxford English Dictionary, 2019). Or, in more statistical terms, “causal effects are defined as comparisons of potential outcomes under different treatments on a common set of units” (Rubin, 2005, p. 322).

There are several frameworks through which scholars understand causality in scientific research, but one of the most prominent is the Rubin Causal Model (Rubin, 1974). The model emphasizes what some scholars call the *Fundamental Problem of Causal Inference* (e.g., Holland, 1986): it is impossible to observe the effect of two different treatments on the same participant. Thus, a causal effect is conceptualized as the difference between *potential outcomes*,

where individual participants *could have been* assigned to either the treatment or control condition. In this sense, the average causal effect indicates how much the outcome *would have* changed had the sample been treated (versus not treated). Put simply, although we cannot observe treatment effects for individuals, we can observe the average treatment effect across a sample (Deaton & Cartwright, 2018).

This framework makes two core assumptions: *excludability* and *non-interference* (see Gerber & Green, 2012, pp. 39-45). Excludability is the assumption that the treatment is the sole causal effect on the outcome. Non-interference is the assumption that treatment versus control status of any individual participant is not affected by the status of another participant.

Put simply, “a causal relationship exists if (1) the cause preceded the effect, (2) the cause was related to the effect, and (3) we can find no plausible alternative explanation for the effect other than the cause” (Shadish et al., 2002, p. 6). The first criterion is easily achieved in an experiment by design. The second criterion is easily achieved via data analysis. However, the third criterion is more challenging to meet, as there are essentially infinite potential alternative explanations (i.e., confounds) for any given study’s results, thereby potentially jeopardizing the excludability assumption (Gerber & Green, 2012).

To address the issue of confounding, researchers aim to ensure experimental groups are equal in all respects except for the independent variable (Fisher, 1937; Gerber & Green, 2008; Holland, 1986; Pearl, 2009; Rubin, 1974; Shadish et al., 2002). If experimental conditions are equal on all characteristics except for the independent variable, then only the independent variable can be responsible for differences observed between conditions (Gerber & Green, 2008; Holland, 1986; Shadish et al., 2002).

Fisher (1937) noted the difficulty of creating equal groups: “it would be impossible to present an exhaustive list of such possible differences appropriate to any one kind of experiment, because the uncontrolled causes which may influence the result are always strictly innumerable” (p. 21). To address this issue, Fisher and his contemporaries developed *random assignment*, which ensures that pre-treatment differences are independent of the treatment condition assigned.

Random Assignment and Causality

R. A. Fisher (1925; 1937) developed the foundational concepts of random assignment as a means to aid causal inference. In the context of agricultural research, he developed random assignment and defined it as “using a means which shall ensure that each variety has an equal chance of being test on any particular plot of ground” (Fisher, 1937, p. 56). In the language of social science research, random assignment to conditions is when a random process (e.g., a random number generator, the flip of a coin, choosing from a shuffled deck of cards) is used to assign participants to experimental conditions, giving all participants an equal chance of being assigned to either condition.

Fisher (1937; p. 23) advocated for the use of random assignment to experimental conditions as a method for mitigating the threat to an experiment’s internal validity: “...with satisfactory randomisation, its validity is, indeed, wholly unimpaired” (for a historical account of Fisher’s advocacy for randomization, see Hall, 2007). Since Fisher’s writing, random assignment has been shown to be best-practice of experimental design and causal inference (e.g., Shadish, et al., 2002). For example, in one of the most well-cited texts on causal inference, Shadish and colleagues (2002, p. 248) explain that random assignment is effective because it “ensures that alternative causes are not confounded with a unit’s treatment condition” and “it reduces the plausibility of threats to validity by distributing them randomly over conditions.” In other words,

because alternative causes are randomly distributed across conditions, they become perfectly balanced as sample size approaches infinity (Geber & Green, 2008; Shadish et al., 2002).

Compared to other methods of equating experimental conditions (e.g., matching) a crucial strength of random assignment is that it balances conditions on known *and* unknown variables (Geber & Green, 2008; Shadish et al., 2002). Other methods, such as matching, may equate groups on variables that may be related to the independent and dependent variables, but threats to validity still remain because experimental groups may still systematically differ on unmeasured variables. This is not a problem for random assignment because it renders the assignment of experimental conditions independent of all other variables in the study.

Random Assignment and Sample Size

It is well known that larger sample sizes reduce the probability that random assignment will result in conditions that are unequal (e.g., Bloom, 2006; Shadish et al., 2002). That is, as sample size increases, differences *within* groups increases, but differences *between* groups decreases (Rose, 2001)—making it less likely that a variable other than the experimental manipulation will explain the results.

Beyond the fact that larger samples are less likely to result in chance differences between conditions, it is unclear *how large is large enough*. As Deaton and Cartwright (2018) aptly noted, “Statements about large samples guaranteeing balance are not useful without guidelines about how large is large enough, and such statements cannot be made without knowledge of other causes and how they affect outcomes” (p. 6).

In the present study, instead of comparing other methods to the standard of random assignment (e.g., Shadish, Clark, & Steiner, 2008), the performance of random assignment itself is put to the test—asking how often random assignment fails to eliminate key differences

between conditions in psychological experiments, and what we can do to avoid being misled by such failures. These questions are assessed in the context of environmental psychology experiments, where chance differences between experimental conditions may be particularly consequential, considering the increasing political polarization of the issue (Ballew et al., 2019; Goldberg, van der Linden, Leiserowitz, & Maibach, 2019; McCright & Dunlap, 2011), and the influential role of race/ethnicity and gender in climate change public opinion and issue engagement (e.g., Ballew et al., 2019; McCright & Dunlap, 2011). If potential participants vary widely in their views about climate change along categories of political party, race/ethnicity, and gender, then the differential distribution of these characteristics across different experimental conditions will affect results. This is especially fitting because background knowledge about alternative causes is necessary to understand how likely, and how much, imbalance between conditions will affect results (Deaton & Cartwright, 2018).

When Random Assignment Fails

Random assignment serves many important functions. A primary function is to estimate an unbiased average treatment effect (ATE; Gerber & Green, 2012). In this context, “unbiased” refers to the fact that, over repeated experiments, assignment to conditions will not be systematically related to participants’ scores on a pre-test, demographic characteristics, or any other variable. This is important because an individual experiment may have baseline differences between conditions, but in the long run over many experiments, differences in either direction will cancel each other out.

Random assignment has other important functions. For example, in observational research, researchers need to justify their causal model as well as identify, measure, and control for known confounders (Pearl, 2009). Random assignment reduces (but does not eliminate) the

need for background knowledge about alternative causes (Deaton & Cartwright, 2018) because, over repeated experiments, it balances confounders across conditions whether or not they are known to the researcher (see Gerber & Green, 2012; Shadish et al., 2002).

Additionally, random assignment can be a tool for removing experimenter bias, leaving assignment to conditions to an independent process instead of it being the decision of the experimenter. This is especially relevant for in-person lab experiments where experimenter bias is easily introduced.

One purpose of random assignment that is crucial for causal inference is to ensure that potential confounding variables are evenly distributed across experimental conditions. What does it mean, in this context, for random assignment to “fail?” Random assignment fails to fulfill its function of balancing potential confounders when, after randomization, experimental conditions non-trivially differ on one or more confounding variables—which would raise the concern that this inequality explains observed differences between experimental conditions. Of course, whether a researcher should consider differences between conditions “non-trivial” depends in large part on the context of the research question and the estimated effect size of the treatment.

Indeed, random assignment might have been successfully *executed* such that the *procedure* by which participants were assigned to conditions was random. But, as noted above, random assignment fails to achieve one of its crucial *functions* when non-trivial differences between conditions remain after random assignment to conditions¹.

¹ Some scholars contend that unbalanced conditions after random assignment is “not a failure of randomization” per se (Shadish et al., 2002, p. 303; also see Kenny, 1975, p. 350). That is, balance of potential confounds is not a primary function of random assignment, but rather, as noted in the main text, *in the long run* (i.e., over repeated experiments) assignment to conditions will not be systematically related to differences between conditions. Thus, random assignment serves the function of giving an unbiased estimate of the average treatment effect, even if individual experiments sometimes have baseline differences between experimental conditions.

Many scholars have pointed out this shortcoming of random assignment: differences on confounders or the outcome variable may remain (e.g., Deaton & Cartwright, 2018; Harville, 1975; Krause & Howard, 2003; Rubin, 2008; Seidenfeld, 1981; Student, 1938). Further, recent large-scale registered replication reports support this warning, suggesting that many published findings of significant effects in the psychological scientific literature likely arose by chance (e.g., Open Science Collaboration, 2015), a phenomenon that is increased by small sample sizes (Ioannidis, 2005; Simmons, Nelson, & Simonsohn, 2011).

It is well known, however, that such chance occurrences become less likely as sample sizes get larger (e.g., Bloom, 2006; Gerber & Green, 2008; Shadish et al., 2002). For example, it is possible that in a study of the effect of a pro-climate message on support for a certain policy, the treatment condition will by chance contain more members of a particular political party than the control condition. More relevant to replication failures and high-rates of false positives, it is also possible that, simply by chance, the participants in the treatment condition are already higher on support for that policy than participants in the control condition. It is intuitive that these situations are more likely in a sample of $n = 50$ than $n = 100$. However, beyond the general maxim that larger sample sizes are less likely to contain consequential chance differences, it is yet unclear precisely what size sample is needed to ensure the likelihood of biased random assignment is kept to a negligible level.

Calls for larger sample sizes are a prominent part of the ongoing conversation about statistical power in psychological scientific studies. For example, Fraley and Vazire (2014) assessed empirical studies published between in 2006 and 2010 in six major social-personality

psychology journals and found that the typical sample size in the selected studies was 104². The researchers calculated estimates of statistical power to detect a typical effect size in social-psychological studies ($r = .20$ or $d = .41$; see Richard, Bond, & Stokes-Zoota, 2003) and found that average power, depending on the journal, ranged from 40-77%. Other researchers have estimated average power to be even lower across psychological experiments (35%; Bakker, van Dijk, & Wicherts, 2012). Recent research finds that sample sizes across four top social psychology journals have significantly increased, with an average sample size of 195 in 2018 (Sassenberg & Ditrich, 2019).

It is clear that larger samples are needed. However, even with larger sample sizes, it is still unclear *how often* random assignment leads to non-trivial differences between experimental conditions and *to what extent* larger sample sizes used in psychological research mitigate—or fail to mitigate—this issue.

Insights regarding the research questions investigated in the current article have especially important implications for research in environmental psychology, and even more so for climate change communication. Climate change is a highly politicized and polarized issue in the United States (e.g., Ballew et al., 2019; Ehret, Van Boven, & Sherman, 2018; Goldberg, et al., 2019; McCright & Dunlap, 2011), and therefore the distribution of liberals and conservatives across experimental conditions will be consequential. Further, recent research finds that Latinos are more engaged in the issue of climate change than non-Latino Whites (Ballew et al., 2019), and therefore chance differences between conditions on race/ethnicity may also be consequential

² In an informal investigation into typical sample sizes in environmental psychology, I identified all experiments with random assignment to conditions and a between-subjects manipulation in the October and December issues of the *Journal of Environmental Psychology* and found that the average sample size was 169 (see the OSF project page for all information). Because this was just a descriptive exercise and was not a systematic survey of the environmental psychological literature, this finding should be interpreted with caution.

in research on climate change communication. Finally, differences in familiarity with treatment messages across experimental conditions, among other variables (e.g., political engagement, education), are likely to influence study results (e.g., Goldberg et al., 2019). Because climate change public opinion in the United States is so heterogeneous, there are more ways for differences between experimental conditions to affect results.

Additionally, research in this topic area has great applied importance, therefore raising the need for unbiased effect size estimates, which may be used inform whether and how to approach campaigns to increase public engagement with the issue. Finally, because field experiments are an especially important aspect of testing the ecological validity of applied environmental psychology research, and large sample sizes difficult and expensive to obtain, it is of great applied importance to determine the likelihood of chance differences that may remain after randomization.

Overview

The current studies use simulated data (Study 1), three real datasets from original environmental psychology experiments (Study 2), and one nationally-representative dataset (Study 3) to examine the degree to which differences exist between randomly assigned experimental conditions. It is worth noting why this is an especially informative approach.

First, the use of simulated data and real data bolsters the point that the issues of random differences between conditions are not confined to mere simulations or general points that larger sample sizes minimize differences between conditions. Rather, the issues presented in this article are generalizable to all kinds of experimental data.

Second, it is worth noting why this point is well-illustrated by using these four real datasets in particular. The current research uses data from three original environmental

psychology experiments that examine some form of climate change communication. This is because, as noted above, these experiments seem to be especially vulnerable to issues arising from differences between experimental conditions. For example, any experiment's internal validity will be threatened by differences in baseline values of the dependent variable. However, experiments on climate change communication can still have threats to internal validity even if baseline values of the dependent variable are the same. That is, differences in demographic characteristics, political ideology, or familiarity with the treatment message can also affect results (Goldberg et al., 2019). Thus, these three datasets make clear that internal validity can be threatened in several ways that are especially pertinent to research on highly politicized issues such as climate change. Further, the experiments included represent common paradigms in climate change communication research, and therefore provide useful estimates of threats to internal validity for researchers in the field.

And finally, the current research uses a large nationally-representative dataset, which enables estimates of how likely random differences between experimental conditions are to occur in random samples of the United States population. Inclusion of this dataset shows that, while experiments in environmental psychology might be especially vulnerable to chance differences affecting study results, this issue is not exclusive to environmental psychology, but social-personality psychology more broadly.

Through random subsampling of these datasets, the current studies test the likelihood that experimental conditions will differ on influential variables that are related to the manipulation (e.g., a pre-test of the dependent variable), as well as on variables that likely moderate responses to messages about climate change (i.e., political ideology, education, message familiarity; see Goldberg et al., 2019). Further, the current research tests the degree to which this likelihood is

mitigated by increasing sample sizes. That is, how large is large enough? All data, materials, and analysis code for all studies are available on the Open Science Framework (OSF) project page at <https://osf.io/69vwe/>.

Study 1

Data

Data for this study were simulated using the R statistical software.

Materials and Procedure

To test the current research questions, standardized normal distributions were randomly simulated using the *rnorm* function, with specifications for a mean of zero and a standard deviation of one. First, a population of 10,000 respondents was simulated, and then half were randomly assigned the value of zero (i.e., control condition) and half the value of one (i.e., treatment condition) using the *complete_ra* function (i.e., complete random assignment). Next, random samples were taken with replacement and frequencies were calculated for how often a mean difference of Cohen's $d = .2$, $.3$, and $.4$ occurred by chance. A loop was used to repeat the entire process 1,000 times for each of the following sample sizes: 50, 100, 200, 300, 400, 500, 700, and 1,000. A simulation of a new population and new random assignment to conditions was conducted for each new subsample.

Results and Discussion

Results show that chance differences between conditions are fairly common for sample sizes typical of social psychology experiments (see Figure 1). For a sample of 50 participants, a chance difference of $d = .2$ occurred in 48% of samples, $d = .3$ for 28% of samples, and $d = .4$ for 15% of samples. For a sample of 100 participants, a chance difference of $d = .2$ occurred in 34% of samples, $d = .3$ for 14% of samples, and $d = .4$ for 5% of samples. Even for a sample of 200

participants, exceeding the average sample size in social psychology's top journals in 2018 (Sassenberg & Ditrich, 2019), a chance difference of $d = .2$ occurred in 18% of samples, $d = .3$ for 4% of samples, and $d = .4$ for 1% of samples. Chance differences of $d = .3$ or $.4$ were virtually eliminated ($< 1\%$) for sample sizes 300 and above, but chance differences of $d = .2$ still persisted: occurring in 8% of samples of 300, in 6% of samples of 400, and in 4% of samples of 500.

These results show that commonly used sample sizes in social psychological experiments are susceptible to non-trivial differences between conditions occurring simply by chance. This has especially strong implications for study planning because, even if a researcher accurately estimates their effect size of interest and chooses a sample size that will give their study high statistical power to detect that effect size, chance differences will leave many of these studies significantly underpowered.

Although these results demonstrate the well-known fact that larger sample sizes will reduce the likelihood of chance differences (e.g., Bloom, 2006), it shows what sample size is needed to reduce chance differences to a negligible level. If a researcher desires, for example, to keep chance differences of $d = .2$ in either direction below 5%, a sample size between 400 and 500 would be needed.

Readers might be wondering whether such differences are confined to mere simulations. That is, do similar differences occur by chance in real psychology experiments? Studies 2 and 3 address this question.

Figure 1

Baseline differences between experimental conditions in simulated data

Note. The height of the bars represents the percentage occurrence out of 1,000 subsamples. Each cluster of three bars are the three levels of effect size difference between conditions: Cohen's $d = .2$, $.3$, and $.4$ (Study 1).

Study 2

Method

Data

Three original environmental psychological experiments ($Ns = 765, 776$, and $1,720$) are used to gauge the likelihood of chance differences between randomly assigned conditions. The following criteria were set for inclusion in the study: (a) each dataset was required to have a sample size of at least 700; (b) baseline measurement of influential variables were measured before random assignment to conditions (i.e., pre-test of dependent measures); and (c) the study completion rate was 100% (i.e., no attrition). The sample size minimum was set such that random subsamples with relatively large sample sizes could be drawn from the overall sample and so that subsamples would be sufficiently different from one another. It was necessary to examine baseline variables that were recorded before random assignment to ensure that it was

not possible for the experimental manipulation to affect participants' responses. And finally, it was important to rule out attrition as the source of differences between experimental conditions in order to gauge whether any differences truly arose by chance³. All three samples were recruited from Prime Panels (TurkPrime, 2019) as part of a larger set of studies on climate change communication. Demographic information for each experiment is available below in Table 1. See the OSF project page for the full version of each survey.

Table 1

Demographic information for experimental datasets included in Study 2

Variable	Experiment 1	Experiment 2	Experiment 3
<i>N</i>	765	776	1,720
Age Mean (<i>SD</i>)	39 (15)	45 (16)	44 (17)
Sex			
Male	37%	34%	36%
Female	63%	61%	64%
Education			
No High School	5%	4%	3%
High School	28%	25%	22%
Some College	40%	41%	42%
College Degree	20%	20%	24%
Graduate Degree	7%	10%	9%

Materials and Procedure

The procedure used in Study 1 was nearly identical to that of Study 2. The R statistical software was used to draw random samples with replacement from each dataset and used a loop

³ This point about attrition is more of an issue for researchers assessing chance differences in their own studies than the experiments included in this study. This is because each random subsample in this study included new random assignment to conditions, therefore making it impossible for attrition to be systematically related to assignment to conditions. However, it's important to highlight here that researchers rule out attrition as an explanation for differences between conditions before assessing whether differences occurred by chance.

to repeat this process 1,000 times for sample sizes of 50, 100, 200, 300, 400, 500, 700, and, for the experiment with the largest sample size, 1,000. Although random assignment was conducted in the original experiments, random assignment to conditions was conducted for each new subsample, with half of participants assigned to the treatment condition and half to the control condition. Then the raw mean difference was calculated for what would need to occur on the variable of interest for there to be a difference between the two experimental conditions with a Cohen's d of .2, .3, and .4. For example, for a variable with a standard deviation of 2.005029, a difference of $d = .2$ would be $2.005029 * .2 = .4010058$. Identical to Study 1, to gauge how often random assignment failed to eliminate differences between conditions for a given effect size, the percentage of the 1,000 resamples that contained such effect size differences was computed. This process was repeated for each sample size (50, 100, and so on) and for each variable that was likely to affect the results of the corresponding experiment (see description below).

Experiment 1. Experiment 1 tested the persuasiveness of a radio story in influencing political moderates' and conservatives' beliefs that climate change is human-caused ($N = 765$). Participants were randomly assigned (using Qualtrics' randomizer function) to listen to the treatment message explaining that climate change is not just a "natural cycle" or listen to a control message about the speed of cheetahs. In this dataset, chance differences on a pre-test of the primary dependent measure were investigated: the question asking whether global warming is human-caused. The question was "Assuming global warming IS happening: How much of it do you believe is caused by human activities, natural changes in the environment, or some combination of both?" (1 = *I believe global warming is caused entirely by natural changes in the environment*, 7 = *I believe global warming is caused entirely by human activities*).

Experiment 2. Experiment 2 tested the persuasiveness of a radio story in influencing Christians' general beliefs about climate change ($N = 776$). Using the randomizer function in Qualtrics, participants were randomly assigned to the treatment or control condition. The treatment message included a radio story about an Evangelical Christian that used to be a climate skeptic but eventually changed his mind and is now a climate leader. The control message was the same as in Experiment 1. In this dataset, chance differences were investigated for political ideology, strength of religious identification, and a pre-test of the participant's familiarity with the topic of global warming. Political ideology was measured by asking participants to complete the sentence "In general, I think of myself as..." (1 = *Very liberal*, 5 = *Very conservative*). Strength of religious identification was measured by asking "How important is your religious identity to your sense of who you are?" (1 = *Very unimportant*, 7 = *Very important*). To measure familiarity, participants were given the following prompt "Recently, you may have noticed that global warming has been getting some attention in the news. Global warming refers to the idea that the world's average temperature has been increasing over the past 150 years, may be increasing more in the future, and that the world's climate may change as a result" and then asked "How *familiar* were you with that statement *before* you read it in this survey?" (1 = *Not at all familiar*, 7 = *Very familiar*).

Experiment 3. Experiment 3 tested the persuasiveness of a written message in influencing Christians' beliefs that climate change is a religious issue ($N = 1,720$). Participants were randomly assigned (via the randomizer function in Qualtrics) to read a treatment message or, in the control condition, participants completed a word-sorting task⁴. In this dataset, chance

⁴ This experiment originally had three conditions: one control condition, and two treatment conditions. However, original random assignment to conditions becomes irrelevant in the current analyses because a new round of random assignment is conducted for each random subsample. Nonetheless, original conditions are available in the dataset on the OSF project page for those interested in re-analysis of the original data.

differences were investigated on political ideology, the belief that environmental protection is a religious issue, and participants' self-reported frequency of religious service attendance. Political ideology was measured with the same question as described in Experiment 2. The belief that environmental protection is a religious issue was measured by asking "In your opinion, how much do you think environmental protection is...?" [a religious issue] (1 = *Not at all*, 7 = *Very much*). For religious service attendance, participants were asked "How often do you attend religious services?" (1 = *Never*, 6 = *More than once a week*). For more details on this experiment, see Goldberg et al. (2019) or see the full survey on the OSF project page.

Results and Discussion

Analyses of each of these three studies were tailored to each experiment's likely sources of chance differences that could be introduced via random assignment. In Experiment 1, chance differences between experimental conditions on a pre-test of the primary dependent variable were examined. In Experiment 2, chance differences between experimental conditions on potential moderators were examined. And finally, using a substantially larger dataset, in Experiment 3 both pre-test differences on a key dependent variable as well as two potential moderators were examined. The purpose of examining different types of variables (e.g., pre-test measures of the dependent measure, moderators) was to ensure there was a diverse set of variables that are likely to influence the results of a typical psychological experiment, thereby extending the generalizability of the results.

Experiment 1

The purpose of Experiment 1 was to investigate the efficacy of a radio message in increasing participants' belief that global warming is human-caused. The treatment message explained why global warming is not just a "natural cycle" and was compared to a control

message that explained the biomechanics of cheetahs' speed. To examine the likelihood of chance differences on the primary dependent measure, the extent to which random subsamples varied on a pre-test of belief that global warming is human-caused was examined. Results are displayed in Figure 2. Results show that differences between experimental conditions are common across a range of sample sizes. For a sample size of 50, a mean difference of $d = .2$ was observed in 49% of samples, a difference of $d = .3$ in 32% of samples, and a difference of $d = .4$ in 18% of samples. For a sample size of 100, differences between conditions were still quite common. A $d = .2$ difference was observed in 37% of samples, $d = .3$ in 17% of samples, and $d = .4$ in 6% of samples. For a sample of 200, approximately the average sample size in a recent analysis of social-personality psychology studies from 2018 (Sassenberg & Ditrich, 2019), 21% of samples found a difference of $d = .2$, 6% found a difference of $d = .3$, and 1% found a difference of $d = .4$.

As expected, these biases got smaller as sample size increased, but non-trivial biases remained even for larger sample sizes (see Figure 2). In samples with 500 participants, for example, a $d = .2$ difference between conditions was still observed in 8% of samples. This is especially concerning when researchers are investigating effects of similar size to $d = .2$.

Figure 2

Baseline differences between experimental conditions in the belief that global warming is human-caused

Note. The height of the bars represents the percentage occurrence out of 1,000 subsamples. Each cluster of three bars are the three levels of effect size difference between conditions: Cohen's $d = .2$, $.3$, and $.4$ (Study 2; Experiment 1).

Experiment 2

In Experiment 2, the scope of variables for analysis was broadened, focusing on variables that may moderate the effect of the manipulation. Because this experiment included a manipulation of a message aimed to influence Christians' views on global warming through a message from a fellow Christian, differences between experimental conditions on political ideology and strength of religious identity were examined. Additionally, because non-naïveté has been shown to reduce responsiveness to treatment effects (Chandler, Paolacci, Peer, Mueller, & Ratliff, 2015; Druckman & Leeper, 2012), differences in reported familiarity with the issue of global warming were also analyzed.

Results indicated that differences between experimental conditions on political ideology were common in relatively smaller samples. For a sample size of 50, a difference of $d = .2$ was

observed in 51% of samples, a difference of $d = .3$ was observed in 31% of samples, and a difference of $d = .4$ was observed in 20% of samples. For a sample size of 100, differences of $d = .2$, $.3$, and $.4$ were observed in 35%, 16%, and 7% of samples, respectively. These chance differences introduced via random assignment were still evident in sample sizes of 200, 300, and 400, albeit at progressively lower rates. Ideological differences still were not totally eliminated for samples of 500, where differences of $d = .2$ and $.3$ were evident 8% and 1% of the time, respectively (see Figure 3).

Differences between experimental conditions in strength of religious identity were similar to those of political ideology (Figure 3). It is worth noting, however, that a difference of $d = .2$ in strength of religious identity was still somewhat common for sample sizes of 300 and 400, which were observed in 13% and 12% of samples, respectively.

Differences between conditions in familiarity with the issue of global warming were similar to those of ideology and strength of religious identity (see Figure 3). These results underscore the issue of non-naïveté in psychological research (Chandler, Paolacci, Peer, Mueller, & Ratliff, 2015; Druckman & Leeper, 2012). For example, Goldberg and colleagues (2019) replicated the same experiment across three samples and found that samples that had participants who were more familiar with the message about the scientific consensus on climate change produced smaller effect sizes. In the current study, however, results show that differences in familiarity can be introduced by the process of random assignment, which can shift results in favor of—or against—the hypotheses of interest, depending on which experimental condition happens to have a disproportionate number of participants who are more familiar with the topic.

Figure 3

Baseline differences between experimental conditions

Note. Left panel shows results for ideology, middle panel for strength of religious identification, and right panel for familiarity with the topic of global warming. The height of the bars represents the percentage occurrence out of 1,000 subsamples. Each cluster of three bars are the three levels of effect size difference between conditions: Cohen's $d = .2$, $.3$, and $.4$ (Study 2; Experiment 2).

Experiment 3

Experiment 3 allowed for additional analyses of potential moderators and a pre-test of a key dependent variable, as well as a substantially larger dataset on which to conduct the analyses. All results are displayed in Figure 3. When examining differences on ideology, a pre-test of the belief that environmental protection is a religious issue, and frequency of religious service attendance, results were similar to those reported previously (see Figure 4).

Figure 4

Baseline differences between experimental conditions

Note. Left panel shows results for ideology, middle panel for the belief that environmental protection is a religious issue, and right panel for frequency of religious service attendance. The height of the bars represents the percentage occurrence out of 1,000 subsamples. Each cluster of three bars are the three levels of effect size difference between conditions: Cohen's $d = .2$, $.3$, and $.4$ (Study 2; Experiment 3).

Study 3

The goal of Study 3 was to investigate differences between randomly-assigned conditions in a nationally-representative dataset. This study allowed for random subsampling from a larger population of participants and used data from a study on a broader range of attitudes of the American public. Additionally, it is especially useful to conduct analyses on the ANES dataset because it is nationally-representative and therefore random differences between conditions reflect differences likely to be observed in a random sample of the United States population.

Method

Data

Data were drawn from the 2016 wave of the American National Election Studies (ANES). There were 4,271 total participants. ANES is a survey conducted every election year on the United States electorate, asking questions about public opinion, voting, and political participation. Additional information about the methods and dataset can be found at <https://electionstudies.org> and the dataset itself can be found on the current article's OSF project page.

Materials and Procedure

The materials and procedure used to assess differences that remain after random assignment were identical to those of Study 2. ANES surveys are conducted on the United States electorate, and do not feature experimental manipulations. However, for the purposes of the current study, it serves as a useful proxy for the American population. Thus, using the same procedures as Studies 1 and 2, half of participants' data were randomly assigned to a treatment condition and half to a control condition and how often differences remain at different sample sizes was assessed.

Three variables were assessed in the current study: political ideology, education, and the belief that the government should reduce income inequality. Political ideology was measured by asking “Where would you place yourself on this scale, or haven’t you thought much about this?” (1 = *Extremely liberal*, 7 = *Extremely conservative*). For participants that refused the question, reported “Don’t know,” or reported “Haven’t thought much about this” were coded as missing. Education was measured by asking “What is the highest level of school you have completed or the highest degree you have received?” (1 = *Less than first grade*, 16 = *Doctorate degree*). Participants who gave a different answer, refused, or reported “Don’t know” were coded as missing. To measure beliefs about income inequality, participants were asked to “Please say to what extent you agree or disagree with the following statement: ‘The government should take measures to reduce difference in income levels’” (1 = *Agree strongly*, 5 = *Disagree strongly*). Participants who refused the question or reported “Don’t know” were coded as missing.

Results and Discussion

Results show similar differences between conditions observed in Study 2 (see Figure 5). First, as expected, chance differences between conditions become less likely as sample size increases. However, there is still a relatively high likelihood of non-trivial differences between conditions across a wide range of sample sizes. When the sample size was set to 50, chance differences of $d = .2$ in political ideology occurred 55% of the time, decreasing to 36% of the time for $d = .3$, and 21% of the time for $d = .4$.

Such differences between conditions were also high for sample sizes common in social-personality psychology studies (see Fraley & Vazire, 2014; Sassenberg & Ditrich, 2019). When the sample size was set to 100, chance differences in ideology occurred in 37% of samples for $d = .2$, 19% of samples for $d = .3$, and 8% of samples for $d = .4$. Finally, when sample size was

instead set to 200, chance differences of $d = .2$ in ideology occurred in 23% of samples, differences of $d = .3$ occurred in 8% of samples, and differences of $d = .4$ occurred in 2% of samples. Such biases were mostly mitigated with a sample size of 500, and were virtually eliminated for sample sizes of 700 and 1,000.

Additionally, differences between conditions on other variables were tested, such as education, that often influence outcome variables of interest in psychological research (e.g., van der Linden, Leiserowitz, & Maibach, 2018). Differences in variables that can themselves be dependent measures in psychological research were also tested, such as the degree of belief that the government should reduce income inequality (see Figure 5). Results across these three variables were consistent, and confirm the finding that differences between randomly assigned conditions often remain in sample sizes typically used in psychological research.

Figure 5

Baseline differences between experimental conditions

Note. Left panel shows results for ideology, middle panel for education, and right panel for the belief that the government should reduce income inequality. The height of the bars represents the percentage occurrence out of 1,000 subsamples. Each cluster of three bars are the three levels of effect size difference between conditions: Cohen's $d = .2, .3$, and $.4$ (Study 3).

General Discussion

The current findings demonstrate that even in the context of gold-standard procedures—randomized controlled experiments—non-trivial differences between conditions are still common for a range of typical sample sizes used in psychological research. Results consistently

show that such difference exist in baseline levels of dependent measures as well as in potential moderators.

This is a significant cause for concern, because the purpose—and widely assumed effect—of random assignment is to eliminate these differences in order to enhance internal validity and thus the ability to make causal inferences. Because differences emerged in baseline scores of dependent variables, these differences may explain false-positive or false-negative results. For example, if a researcher conducting an experiment designed to detect a typical effect size observed in social-psychological research ($d = .43$ as reported by Richard et al., 2003), a power analysis using the pwr package in R suggests that the researcher would need approximately 172 participants for a two-group experimental design to achieve 80% power. Our results show that even with a sample size of 200 and a population effect size of $d = .43$, the researcher may observe in their sample an effect size of $d = .2$ larger ($d = .63$) or smaller ($d = .23$) between 17% and 23% of the time simply due to random chance.

Implications for environmental psychology. These findings are important for research in environmental psychology and especially for research on climate change communication, in which experimental results have been shown to be sample-dependent (e.g., Goldberg et al., 2019). Because environmental issues—and climate change in particular—are highly politicized and polarized in the United States (Ballew et al., 2019; Ehret et al., 2018; Goldberg et al., 2019; McCright & Dunlap, 2011), experiments on climate change communication may be especially vulnerable to biases that result from unequal distributions of political ideology, familiarity with the issue, and baseline differences on the dependent variable across experimental conditions.

Unfortunately, it likely would not be sufficient to simply check if treatment effects are moderated by these variables, considering interaction effects are extremely difficult to test with

sufficient statistical power (Gelman, 2018). Thus, as recommended below, researchers of climate change communication should include pre-test measures (on the dependent variable and related variables) whenever possible, test for differences between conditions on those variables and, if non-trivial differences are identified, adjust for them.

These findings may explain some false positive results more generally, or at least inflated effect size estimates, because at a sample size of 200 it is still somewhat common for the treatment condition to *already* have higher scores on the dependent variable of interest. This also may explain false negative results because a researcher may correctly estimate the effect size of interest, but then observe a much smaller effect because of random chance—leading the researcher to incorrectly conclude that a manipulation caused little to no change in the dependent variable. Further, both false positives and false negatives can occur as a result of unequal distribution between conditions of a confounding moderating variable—which this study shows is not uncommon. These findings are especially important for the subfield of environmental psychology where field tests are important but expensive.

Recommendations

Identifying chance differences. Fortunately, there are several ways to identify and mitigate the chance differences reported here. First and foremost is identifying if such chance differences exist in one's data. In correlational research, researchers, editors, and reviewers consistently point out the need to control for confounders and that an experiment would be more effective at doing so. However, in my broad reading of the social-personality psychological literature, I find that while it is common for manuscripts to report that random assignment was performed, it is rare that manuscripts report on whether random assignment effectively reduced differences between conditions on variables likely to influence results. An important

recommendation, then, is for researchers to investigate and report whether random assignment truly “worked.”

One way to do this is to, when possible, assess pre-test scores on the dependent variable. This gives researchers a clear view of the extent to which experimental conditions differ on the primary dependent variables of interest, and also has the added benefit of substantially improving statistical power (Charness, Gneezy, & Kuhn, 2012; Goldberg et al., 2019). Researchers should also measure, test, and report whether random assignment successfully created groups that are equal on key moderators such as demographic variables (e.g., education, age), underlying worldviews (e.g., political ideology), or familiarity with the topic of the experiment. It is important to measure any such variables that may plausibly affect the dependent variable, to ensure that random differences between conditions do not influence the results (Rubin, 2008).

Mitigating chance differences. When researchers discover that random assignment did not successfully eliminate baseline group differences between conditions, there are multiple effective solutions available. If stakes are high, such as in clinical psychological research, or in a field experiment, researchers can re-randomize to see if baseline differences are successfully mitigated (Rubin, 2008; Spratt & Farewell, 1993).

If pre-test differences cannot be evaluated ahead of the experimental treatment, such as in online survey experiments, the solution is similar to de-confounding techniques used in correlational research. For example, researchers can include relevant covariates in their analyses (see Pearl, 2009; Rohrer, 2018). This deconfounds the relationship between the independent and dependent variables by holding the value of each covariate constant (Pearl, 2009). However, in order to increase the precision of estimates of an experimental treatment effect, it is important

that covariates predict the dependent variable (Geber & Green, 2008) and are measured *before* random assignment to experimental conditions (Montgomery, Nyhan, & Torres, 2018).

Although covariate adjustment can be an effective way to reduce differences between experimental conditions, it also gives researchers greater analytical flexibility, which increases the likelihood of Type I errors (Simmons et al., 2011). Thus, if possible to know influential covariates in advance, researchers should pre-register their analysis plan and explicitly state which variables will be included in analyses. If not pre-registered, researchers should be transparent about their decisions that led them to choose covariates on which to adjust.

Another way to prevent or mitigate differences that may arise by chance is matching. Matching is when researchers create groups that are matched on variables that are suspected to be related to the outcome. For example, if sex is expected to have a strong relationship with the outcome, the researcher can compare groups with the same number of males and females in each group. If both experimental groups have the same number of males and females, the relationship between the independent and dependent variables is unconfounded. Even more effective, participants may be matched on a pre-test of the dependent variable (Shadish et al., 2002).

Matching can be effective in aiding causal inference when random assignment is not ethical or possible, but can also help reduce the probability of chance differences when random assignment is used. For example, participants can be matched on a pre-test of the dependent variable and then randomly assigned from within the matched groups (e.g., see Shadish et al., 2002, pp. 304-307 for examples of this procedure).

Another method to mitigate chance differences is stratified analysis. In the above example where sex is expected to be related to the outcome, researchers can split their sample by sex and analyze sex-homogeneous groups separately and then combine the results to determine

the overall treatment effect (Pearl, 1993; Rohrer, 2018). Stratified analysis is more challenging when the relevant covariates have many levels or are continuous. In such cases, researchers can create strata that groups participants that are similar on the variable of interest (Shadish et al., 2002). In such cases, more strata will be more effective in reducing differences due to the corresponding variable, but about 90% of the differences will be removed with five strata (Cochran, 1968).

Although there are helpful methods for reducing differences due to covariate imbalance, none will be as effective as using random assignment with larger sample sizes. This is because all of the above methods require that the researcher know and measure influential covariates, whereas random assignment mitigates chance differences due to known and unknown confounders (Geber & Green, 2008; Shadish et al., 2002). One way to increase sample sizes, and therefore reduce the likelihood of chance differences, is to pool resources across labs (Uhlmann et al., 2019).

An important takeaway is that the problem that the current studies raise regarding random assignment is not an unavoidable limitation nor an insurmountable obstacle. Rather, it is an avoidable limitation that nonetheless affects much of current psychological research.

Implications for statistical power. The evidence reported in the current studies has important implications for study planning, especially when researchers conduct power analyses in order to decide their goal sample size for a given experiment. This is especially important when it is difficult or impossible to record pre-test information. In such a scenario, as is common in psychological research, researchers might not be able to gauge the extent to which experimental groups differ in important ways. This means that there will be many cases where chance differences that remain following random assignment will substantially undermine

statistical power. Because such differences are random by definition, this also means that chance differences can also improve power via inflating the effect size in a given experiment.

Given that the population effect size is unknowable, power analyses heavily rely on informed guesses. Kenny and Judd (2019) argue that power analyses should include an informed guess about the heterogeneity of effect sizes. One way to develop an informed guess is to examine effect size heterogeneity in studies that replicate the same experiment across many samples (e.g., Klein et al., 2018). Researchers can use these large-scale projects to inform their own assessments of the variability of their phenomenon of interest.

Similarly, given that a non-trivial proportion of experiments will have differences that remain after random assignment, researchers should use the results reported here to develop an informed guess about the likelihood of non-trivial differences and adjust their sample sizes accordingly. For example, consider a researcher that estimates that their effect size of interest is $d = .4$. A power analysis suggests that a sample size of about 200 is needed to detect this effect size at 80% power. Researchers can use the results reported here to adjust their power analysis to accommodate a smaller effect size that might occur because of differences introduced during random assignment. Say, for example, this researcher is concerned that differences in political ideology between conditions might affect their results. Results from Study 2 in this article show that a chance difference of $d = .2$ will happen approximately 23% of the time in a sample size of 200. Half of these instances (about 11.5%) will reduce statistical power. If the researcher wanted to protect against a chance difference of $d = .2$, they would need to power their study based on an effect size of $d = .2$ ($d = .4 - .2 = .2$), and therefore need approximately 800 participants to achieve the similar statistical power.

While it is clear that larger samples would substantially improve the precision of effect size estimates, collecting large samples are often unfeasible for many researchers, which may be due to lower resources or labor-intensive data collection. Thus, researchers should consider pooling their resources with other labs (Uhlmann et al., 2019), or perform sequential analyses, where they perform pre-registered data analyses while data collection is still in progress, while ensuring measures are taken to control the Type I error rate (see Lakens, 2014).

Criticisms. One might argue that consequential chance differences between conditions will even out in the long run, and inferential statistics such as the p -value already account for this. Thus, there will be no systematic differences in either direction as the number of experiments increases towards infinity. In theory, this is true. In practice, however, this is highly unlikely. Because decisions to publish scientific articles heavily favor “positive” results (i.e., results that show a significant difference; Rosenthal, 1979; Sterling, 1959), researchers will be more likely to publish false-positive results and leave false-negative results unpublished. Thus, differences that remain after random assignment will be more likely to inflate rather than deflate effect sizes in the published literature. Further, researchers often draw inferences from their individual studies, and therefore the long-run frequency is unhelpful in this situation.

However, recent efforts to improve scientific practices may be promising for reducing nearly exclusive favorability towards positive results. One example is conducting peer-review *before* data are collected—a journal submission type termed “registered reports” (Nosek & Lakens, 2014). For the sake of argument, let’s assume that reforms to scientific publishing practices leads to equal probability of publishing positive and negative results. In such a case, chance differences that occur following random assignment will be equally likely to favor positive or negative results. But this still does not speak to the likelihood that any *particular*

experiment will have chance differences between conditions. This can be highly consequential especially for individual researchers, leading them to conclude that treatment effects are present when they are not (Type I error), or leading them to conclude that treatment effects are non-existent when they are actually present (Type II error). This is especially consequential when funding dedicated to social scientific research is on the decline (Lupia, 2014), and therefore it is less feasible for most researchers to simply conduct more or larger experiments. Further, addressing the issue of chance differences between conditions is often straightforward, yet appears uncommon in the social and environmental psychological literatures. Thus, the fact that chance differences will eventually converge around zero as the number of experiments increases towards infinity is no reason to avoid addressing chance differences that arise in individual experiments.

Conclusion

To conclude, while random assignment is an excellent tool for improving causal inference, it is not without risk (Harville, 1975; Krause & Howard, 2003; Rubin, 2008; Seidenfeld, 1981; Student, 1938). Non-trivial differences between experimental conditions that can directly influence the results of experiments are more likely than one might expect. Thus, researchers should 1) anticipate and measure influential confounders even when conducting randomized experiments and 2) rule out the influence of such confounders either via participant recruitment, statistical methods, or both. These methods should be common practice in social psychological scientific experiments.

References

Bakker, M., van Dijk, A., & Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science*, 7(6), 543-554.

- Ballew, M. T., Goldberg, M. H., Rosenthal, S. A., Cutler, M. J., & Leiserowitz, A. (2018). Climate Change Activism Among Latino and White Americans. *Frontiers in Communication*, 3(58), 1-15.
- Ballew, M. T., Leiserowitz, A., Roser-Renouf, C., Rosenthal, S. A., Kotcher, J. E., Marlon, J. R., ... & Maibach, E. W. (2019). Climate Change in the American Mind: Data, Tools, and Trends. *Environment: Science and Policy for Sustainable Development*, 61(3), 4-18.
- Bloom, H. S. (2006). The core analytics of randomized experiments for social research. *The Sage Handbook of Social Research Methods*, 115-133.
- Charness, G., Gneezy, U., & Kuhn, M. A. (2012). Experimental methods: Between-subject and within-subject design. *Journal of Economic Behavior & Organization*, 81(1), 1-8.
- Chandler, J., Mueller, P., & Paolacci, G. (2014). Nonnaïveté among Amazon Mechanical Turk workers: Consequences and solutions for behavioral researchers. *Behavior Research Methods*, 46(1), 112-130.
- Cochran, W. G. (1968). The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics*, 295-313.
- Deaton, A., & Cartwright, N. (2018). Understanding and misunderstanding randomized controlled trials. *Social Science & Medicine*, 210, 2-21.
- Druckman, J. N., & Leeper, T. J. (2012). Learning more from political communication experiments: Pretreatment and its effects. *American Journal of Political Science*, 56(4), 875-896.
- Ehret, P. J., Van Boven, L., & Sherman, D. K. (2018). Partisan Barriers to Bipartisanship: Understanding Climate Policy Polarization. *Social Psychological and Personality Science*, 9(3), 308-318.

Fisher, R. A. (1925). *Statistical methods for research workers*. Oliver And Boyd; Edinburgh; London.

Fisher, R. A. (1937). *The design of experiments*. Oliver And Boyd; Edinburgh; London.

Fraley, R. C., & Vazire, S. (2014). The N-pact factor: Evaluating the quality of empirical journals with respect to sample size and statistical power. *PloS One*, 9(10), e109019.

Gelman, A. (2018). *You need 16 times the sample size to estimate an interaction than to estimate a main effect*. Accessed April 22, 2019 at <https://statmodeling.stat.columbia.edu/2018/03/15/need-16-times-sample-size-estimate-interaction-estimate-main-effect/>.

Gerber, A. S., & Green, D. P. (2008). Field experiments and natural experiments. In *The Oxford handbook of political science*. Oxford: Oxford University Press.

Gerber, A. S., & Green, D. P. (2012). *Field experiments: Design, analysis, and interpretation*. WW Norton: New York.

Goldberg, M. H., Gustafson, A., Ballew, M. T., Rosenthal, S. A., & Leiserowitz, A. (2019). A Social Identity Approach to Engaging Christians in the Issue of Climate Change. *Science Communication*, 41(4), 442-463.

Goldberg, M., van der Linden, S., Ballew, M. T., Rosenthal, S. A., & Leiserowitz, A. (2019). Convenient but biased? The reliability of convenience samples in research about attitudes toward climate change. *Preprint accessed at <https://osf.io/2h7as/>*.

Goldberg, M. H., van der Linden, S., Ballew, M. T., Rosenthal, S. A., & Leiserowitz, A. (2019). The role of anchoring in judgments about expert consensus. *Journal of Applied Social Psychology*, e0001.

- Goldberg, M. H., van der Linden, S., Leiserowitz, A., & Maibach, E. (2019). Perceived social consensus can reduce ideological biases on climate change. *Environment and Behavior*, doi:10.1177/0013916519853302
- Hall, N. S. (2007). RA Fisher and his advocacy of randomization. *Journal of the History of Biology*, 40(2), 295-325.
- Harville, D. A. (1975). Experimental randomization: Who needs it?. *The American Statistician*, 29(1), 27-31.
- Holland, P. W. (1986). Statistics and causal inference. *Journal of the American statistical Association*, 81(396), 945-960.
- Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS Medicine*, 2(8), e124.
- Kenny, D. A. (1975). A quasi-experimental approach to assessing treatment effects in the nonequivalent control group design. *Psychological Bulletin*, 82(3), 345-362.
- Kenny, D. A., & Judd, C. M. (2019). The unappreciated heterogeneity of effect sizes: Implications for power, precision, planning of research, and replication. *Psychological Methods*, 1-12.
- Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams Jr, R. B., Alper, S., ... & Batra, R. (2018). Many Labs 2: Investigating variation in replicability across samples and settings. *Advances in Methods and Practices in Psychological Science*, 1(4), 443-490.
- Krause, M. S., & Howard, K. I. (2003). What random assignment does and does not do. *Journal of Clinical Psychology*, 59(7), 751-766.
- Lakens, D. (2014). Performing high-powered studies efficiently with sequential analyses. *European Journal of Social Psychology*, 44(7), 701-710.

- Lupia, A. (2014). What is the value of social science? Challenges for researchers and government funders. *PS: Political Science & Politics*, 47(1), 1-7.
- McCright, A. M., & Dunlap, R. E. (2011). Cool dudes: The denial of climate change among conservative white males in the United States. *Global Environmental Change*, 21(4), 1163-1172.
- McCright, A. M., & Dunlap, R. E. (2011). The politicization of climate change and polarization in the American public's views of global warming, 2001–2010. *The Sociological Quarterly*, 52(2), 155-194.
- Montgomery, J. M., Nyhan, B., & Torres, M. (2018). How conditioning on posttreatment variables can ruin your experiment and what to do about it. *American Journal of Political Science*, 62(3), 760-775.
- Nosek, B. A., & Lakens, D. (2014). Registered reports: a method to increase the credibility of published results. *Social Psychology*, 45(3), 137-141.
- Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251), aac4716.
- Oxford English Dictionary. (2019). *Cause*. Retrieved from <https://www.oed.com/view/Entry/29147?rskey=A1A3y8&result=1#eid>.
- Pearl, J. (2009). *Causality*. Cambridge University Press.
- Richard, F. D., Bond Jr, C. F., & Stokes-Zoota, J. J. (2003). One hundred years of social psychology quantitatively described. *Review of General Psychology*, 7(4), 331-363.
- Rohrer, J. M. (2018). Thinking clearly about correlations and causation: Graphical causal models for observational data. *Advances in Methods and Practices in Psychological Science*, 1(1), 27-42.

- Rose, G. (2001). Sick individuals and sick populations. *International Journal of Epidemiology*, 30(3), 427-432.
- Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, 86(3), 638-641.
- Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5), 688-701.
- Rubin, D. B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469), 322-331.
- Rubin, D. B. (2008). Comment: The design and analysis of gold standard randomized experiments. *Journal of the American Statistical Association*, 103(484), 1350-1353.
- Sassenberg, K., & Ditrich, L. (2019). Research in social psychology changed between 2011 and 2016: larger sample sizes, more self-report measures, and more online studies. *Advances in Methods and Practices in Psychological Science*, 107-114.
- Seidenfeld, T. (1981). Levi on the dogma of randomization in experiments. In *Henry E. Kyburg, Jr. & Isaac Levi* (pp. 263-291). Springer, Dordrecht.
- Shadish, W. R., Clark, M. H., & Steiner, P. M. (2008). Can nonrandomized experiments yield accurate answers? A randomized experiment comparing random and nonrandom assignments. *Journal of the American Statistical Association*, 103(484), 1334-1344.
- Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston, MA: Houghton Mifflin.
- Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, 22(11), 1359-1366.

- Sprott, D. A., & Farewell, V. T. (1993). Randomization in experimental science. *Statistical Papers*, 34(1), 89-94.
- Sterling, T. D. (1959). Publication decisions and their possible effects on inferences drawn from tests of significance—or vice versa. *Journal of the American Statistical Association*, 54(285), 30-34.
- Student. (1938). Comparison between balanced and random arrangements of field plots. *Biometrika*, 363-378.
- TurkPrime. (2019, March). Retrieved from turkprime.com.
- Uhlmann, E. L., Ebersole, C. R., Chartier, C. R., Errington, T. M., Kidwell, M. C., Lai, C. K., ... & Nosek, B. A. (2019). Scientific utopia III: Crowdsourcing science. *Perspectives on Psychological Science*, 1745691619850561.
- van der Linden, S., Leiserowitz, A., & Maibach, E. (2018). Scientific agreement can neutralize politicization of facts. *Nature Human Behaviour*, 2(1), 2.