Convenient but Biased?

The Reliability of Convenience Samples in Research

About Attitudes Toward Climate Change

Matthew H. Goldberg^{a,*}, Sander van der Linden^{b,*}, Matthew T. Ballew^a,
Seth A. Rosenthal^a, and Anthony Leiserowitz^a

^aYale Program on Climate Change Communication
Yale University, 205 Prospect St., New Haven, CT, 06511, United States

^bDepartment of Psychology, School of Biological Sciences
University of Cambridge, Downing St., Cambridge CB2 3EB, United Kingdom

Acknowledgements

This work was supported by the MacArthur Foundation and the TomKat Foundation. Funders were not involved in the study design; the collection, analysis, or interpretation of the data; in writing of the manuscript; or in the decision to submit this article for publication.

Open Practices

All data and materials for this research are available at <https://osf.io/zvdcj/>.

*Corresponding author: Matthew Goldberg, 205 Prospect St., New Haven, CT, 06511. E-mail: matthew.goldberg@yale.edu; Matthew Goldberg and Sander van der Linden contributed equally to this work

Abstract

Ideally, research in the social and behavioral sciences would be conducted using samples that perfectly represent the population of interest. However, it is often not feasible for researchers to collect representative samples, forcing them to rely on convenience samples. While some researchers have found that convenience samples can produce comparable results to population-based samples, others have found evidence of biased results. This suggests that differences in the composition of study samples (e.g., demographic, ideological) can sometimes influence study results, leading to contradictions in the literature. To test this, we conducted four identical experiments across three platforms (Amazon Mechanical Turk, TurkPrime Panels, and Facebook) using two different study designs (mixed factorial and between-subjects). We tested a finding known to be robust: communicating the descriptive norm that most climate scientists are convinced that human-caused climate change is happening leads respondents to update their own beliefs about the issue. The results suggest significant variability in effect sizes depending on the sample source and study design, with effect sizes ranging from near-zero to large. Additionally, we make novel use of rake weighting to adjust for inter-sample differences (demographics, ideology, message familiarity) to determine whether differing results from identical experiments are due to underlying sample differences or to the unreliability of the effects. We discuss the implications these results have for experimental research, especially around politicized issues such as climate change.

Keywords: replication, validity, weighting, attitudes, climate change

Convenient but Biased?

The Reliability of Convenience Samples in Research

About Attitudes Toward Climate Change

In an ideal world, psychology experiments would be conducted with samples that perfectly represent the population to which the results are intended to generalize. However, obtaining representative samples is often difficult and expensive. Thus, researchers often have to rely on convenience samples, such as those collected from undergraduate students, Amazon's Mechanical Turk (MTurk), or social media platforms such as Facebook (e.g., Dixon, 2016; Kerr & Wilson, 2018; Landrum, Hallman, & Jamieson, 2018). This research constraint is becoming more pronounced as funding for the social and behavioral sciences decreases (Lupia, 2014).

Reliance on convenience samples has been pervasive in social psychology research for decades. For example, a content analysis of articles published in 1980 in three top-tier social psychology journals found that, depending on the journal, the percentage of articles that relied on undergraduate student samples ranged from 81% to 89% (Sears, 1986). About two decades later, results from a replication (Henry, 2008) were even more pronounced. Research on prejudice, published in these same journals between 1990 and 2005, relied on student samples between 88% and 98% of the time. A recent content analysis of articles from 2015 indicates that researchers have begun to rely increasingly on convenience samples from MTurk (34%); however, undergraduate student samples have remained popular (44%; Anderson et al., 2018).

What implications does the reliance on convenience samples have for the validity, replicability, and generalizability of experiments? The evidence is mixed. For instance, in a comparison of experiments conducted with both convenience samples (e.g., MTurk, students)

and a representative population-based sample (via the National Science Foundation-funded Time-sharing Experiments for the Social Sciences), the same statistical inference (i.e., statistical significance in the same direction) was made using the convenience sample as the representative sample in 25 of 32 experiments. The researchers concluded that, in general, studies using convenience samples provide comparable results to studies using representative samples¹. Other recent studies have reached similar conclusions (e.g., Berinsky et al., 2012; Weinberg et al., 2014).

Other research, however, suggests that the answer is more complex. For instance, Krupnikov and Levine (2014) found that effects associated with student samples and representative samples were statistically comparable once moderators were considered (e.g., experience with experiments), but this was not true for samples recruited from MTurk. Research on highly politicized topics (e.g., climate change), in particular, has received criticism for relying too much on MTurk samples (Kahan, 2013), especially when making inferences about psychological differences between liberals and conservatives or when using political ideology as a moderator of experimental treatment effects (Clifford, Jewell, & Waggoner, 2015).

Because of the mixed assessment of the generalizability of research findings derived from convenience samples, we suggest that the question of whether convenience samples (i.e., MTurk or students) can serve as proxies for representative samples is not a simple one to answer. The evidence indicates that research results can be influenced by the characteristics of a particular sample. Accordingly, we argue that researchers should investigate the extent to which their research questions, and data they collect to answer them, are susceptible to producing differing

¹The authors did not suggest, however, that convenience samples should be considered to be a substitute for representative samples. Factors such as uncertainties about unmeasured differences between convenience and population samples, lack of variance in potentially influential moderators (e.g., age, education), and the changing nature and composition of convenience samples must also be taken into consideration.

results depending on the characteristics of their samples. To examine this phenomenon, we focus on attitudes about a socially relevant, topical, and highly politicized issue: *global climate change*.

Context and Replication

Widespread failures to reproduce study results across many disciplines have increased scholars' emphasis on replicability. For example, a large-scale effort to replicate findings in psychology found that only 36% of studies reproduced the same substantive findings (Open Science Collaboration, 2015).

Non-replication, however, does not always mean the original findings were spurious (Gilbert et al., 2016). Although an original experiment and its replication can employ identical materials and procedures, it is still possible that some phenomena are sensitive to other contextual differences (e.g., changes over time; differences as a result of culture, location, or population). Indeed, in their re-analysis of the Open Science Collaboration's large-scale study, Van Bavel, Mende-Siedlecki, Brady, and Reinero (2016) found that experiments that were higher in contextual sensitivity were less likely to successfully replicate the original findings. This suggests that non-replication can be due, at least in part, to differences in context sensitivity.

Context and Climate Change Psychology

The social psychological study of climate change attitudes is particularly context sensitive. For example, using multilevel regression and post-stratification (MRP) to estimate average treatment effects at different geographic levels, Zhang et al. (2018) found that a normative message about the scientific consensus on climate change had different impacts depending on geographic location. When evaluating the change in people's perception of the scientific consensus, effect sizes ranged from an estimated increase of 9 percentage points in Charlotte, Virginia to an increase of 25 percentage points in Nicholas, West Virginia. Because

climate change is a highly politicized issue (Dunlap, McCright, & Yarosh, 2016; Leiserowitz et al., 2019) and there are large racial and ethnic differences in concern about climate change (Ballew, Goldberg, Rosenthal, Cutler, & Leiserowitz, 2019; Leiserowitz, Cutler, & Rosenthal, 2017; Schuldt & Pearson, 2016), the results of social science experiments on climate change likely vary substantially depending on the demographic composition of the sample.

The issue of "pre-treatment effects" (i.e., exposure to the treatment information prior to an experiment) is also relevant to the study of attitudes about the scientific consensus on climate change. The effectiveness of a treatment message may be diminished if respondents have already adjusted their attitudes as a result of prior exposure to that message or have already discounted or resisted the message (e.g., Chandler, Paolacci, Peer, Mueller, & Ratliff, 2015; Druckman & Leeper, 2012). In the context of climate change, it is likely that prior exposure to scientific evidence is not uniform across demographic groups. For example, MTurk and Facebook research respondents have been found to be more educated, on average, than members of nationally representative probability samples (Boas, Christenson, & Glick, 2018). If respondents from samples collected via MTurk and Facebook are more educated, it is likely that they have more previous exposure to and knowledge of the scientific consensus on climate change, which could in turn reduce effect sizes of consensus-message experiments using those platforms. In sum, it is important to investigate the differences in pre-treatment effects and sample characteristics among convenience-sample platforms, and to find ways to address these differences.

Experimental Study Designs: Mixed Factorial Versus Between-Subjects

One way to evaluate pre-treatment effects on the outcomes of experiments is to compare results from within-subjects study designs versus between-subjects study designs. In comparison to between-subjects designs (in which dependent variables are measured among members of two

distinct groups only after the experimental manipulation is presented), within-subjects designs (in which the dependent variables are measured among group members before, and then again after, the experimental manipulation) are more sensitive to detecting treatment effects because they are more statistically powerful (Charness, Gneezy, & Kuhn, 2012). In other words, within-subjects designs control for within-group variation and, thus, provide an added layer of clarity about the degree to which participants have "updated" their responses after an intervention. At the same time, however, within-subjects designs are susceptible to potential demand effects (i.e., pretest sensitization; Willson & Putnam, 1988; see also Schwarz, 1999; Wood et al., 2016) in which the repetition of identically worded questions before and after the experimental manipulation can influence responses.

Although more expensive and involved than between-subjects or basic within-subjects designs, within-subjects designs that also include a randomized control group (i.e., mixed factorial designs) are traditionally considered the gold-standard (Campbell, 1957; Huck & McLean, 1975). Mixed factorial designs simultaneously control for both within- and between-subject sources of variation, alleviating some of the problems associated with both between-subjects designs and basic within-subjects designs that do not include a control group. Many scientific consensus message experiments use mixed factorial designs (e.g., Harris et al., 2018; van der Linden et al., 2015; 2018). Accordingly, we compare mixed factorial designs with between-only designs in the current study. This allows us to evaluate the importance of pre-treatment effects as well as the influence of relevant design characteristics.

Correcting for Differences in Sample Characteristics

One common way to correct for known differences between a survey sample and population benchmarks is through weighting. Weighting is “the application of a constant

multiplying factor to the data provided by a respondent to vary the contribution that respondent's results make to the overall estimates" (Dorofeev & Grant, 2006, p. 45). Put simply, respondents from demographic groups that are underrepresented in a sample are weighted more heavily and respondents from overrepresented groups are weighted less heavily in the calculation of the researcher's estimates. However, even when researchers collect samples that are matched on or weighted to key characteristics (e.g., demographics, political party) of the population of interest, it is still possible that the sample differs from the larger population on unknown or unmeasured characteristics (e.g., motivation for participating in the study).

To address these issues, we use a novel strategy by using weighting as a method for comparing two experiments with identical methods, but different samples. To illustrate, consider the following example. A researcher interested in climate change attitudes tests whether a message about the expert consensus on climate change leads people to update their views on the issue. She conducts a simple experiment and finds that, on average, people update their beliefs. To ensure her results are reliable, a colleague replicates the experiment with another sample, but finds that people do not update their beliefs. Although such contradictory results often spark debates about reproducibility, in these cases, it is actually difficult to know whether the original results failed to replicate because they happened by chance or because crucial differences between the two samples led to differences in the results. However, tests of individual moderators are unlikely to sufficiently resolve that question, particularly if two samples differ on more than one or two characteristics at the same time (e.g., age, education, familiarity with the study topic).

In such a scenario, researchers can use weighting to empirically test whether experiments with identical methods, but different sample characteristics, would yield similar results if

underlying sample characteristics were the same (e.g., demographics, political ideology, issue familiarity). In the current paper, we explore three kinds of weights that are relevant to climate change and other highly politicized issues: demographics (sex, age, race, and education), demographics plus ideology, and pre-exposure familiarity with the treatment message. We use these weights to investigate the extent to which each weight—depending on the variables on which the sample is weighted—make the results of two experiments more similar to each other. We suggest that this approach may be promising for researchers who wish to better understand heterogeneity in their experimental results, which may aid theory development (McShane, Tackett, Bockenholt, & Gelman, 2018).

The Current Experiment

The goals of the current experiment are to: (a) evaluate whether treatment effects vary based on sample source; (b) evaluate whether treatment effects vary based on study design (mixed versus between-subjects); and (c) explore whether weighting can explain and mitigate differences across samples. Answering these questions is important for both basic and applied research. For basic research, precise estimates of effect sizes provide crucial information for estimating the appropriate number of participants needed for the desired level of statistical power. Further, treating effect size heterogeneity as something to be understood provides opportunities for theory improvement (McShane et al., 2018). For applied research, the effect size on one's target population is of central interest, and if an effect is heavily sample-dependent, applied researchers need to understand why.

Method

Participants

Participants were recruited from three sources: MTurk, TurkPrime Panels, and Facebook (total $N = 4,216$)². Four samples were collected in total, one each from MTurk and TurkPrime Panels, and two from Facebook. Within each sample, participants were randomly assigned to either a mixed design or a between-subjects design. The total sample sizes, sample sizes for each design, and the demographic breakdown of each sample are displayed in Table 1. No participants were excluded, and all sample sizes were determined before any data analysis. All procedures received institutional review board approval.

Table 1. Sample sizes and demographic composition of each sample.

Variable	MTurk	TurkPrime Panels	Facebook General	Facebook Targeted
<i>N</i>	1,003	1,086	1,043	1,084
<i>n</i> mixed	530	520	426	441
<i>n</i> between	473	566	617	643
Age	<i>M</i> = 34	<i>M</i> = 41	<i>M</i> = 57	<i>M</i> = 37
Sex				
Male	59%	41%	56%	61%
Female	41%	59%	44%	39%
Education				
No High School	1%	4%	1%	3%
High School	10%	25%	10%	33%
Some College	29%	41%	35%	28%
College Degree	46%	23%	30%	20%
Graduate Degree	14%	8%	24%	15%
Ideology				
Liberal	53%	31%	52%	47%
Moderate	20%	42%	20%	31%
Conservative	27%	27%	28%	22%

²The effect size observed in a recent large, nationally representative experiment using the same manipulation and nearly identical materials was large ($d = .88$; van der Linden et al., 2018). We used the pwr package in R to conduct a power analysis to determine the minimum sample size per cell (Champley et al., 2018). The power analysis determined that a minimum of 35 participants per cell are required to detect this effect with 95% power. Because one of the primary goals of this paper is to evaluate variability in effect sizes, we collected sample sizes that were several times larger than what would be required for 95% power (~250 per cell).

Materials and Procedure

All data and materials from these experiments are available on our Open Science Framework (OSF) project page: <https://osf.io/zvdcj/>. In this study, we report all measures, manipulations, and exclusions. Participants were recruited from MTurk, TurkPrime Panels, or Facebook. In all cases, participants opted into the experiment. Participants were compensated for their participation on MTurk and TurkPrime Panels, but not on Facebook. On MTurk and TurkPrime Panels, participants saw a list of studies and chose to participate in our experiment titled: “Topics of Recent Public Interest.” On Facebook, we recruited participants through paid advertisements that had a generic photo of a map of the United States with the caption: “Have 5 minutes to spare? We’d like to hear from you – take this quick survey: Public opinion survey about popular media topics.” All experiments were identically constructed on the Qualtrics survey platform.

Participants began the study by reading that they would be given a recent news topic randomly chosen from 20 possible topics. Participants were then randomly assigned to either the mixed or between-subjects design. Participants in the mixed design answered pre-test questions regarding their beliefs about global warming. Questions included whether participants think global warming is happening, whether it is human-caused, and whether they worry about it. These questions were measured on Likert scales ranging from 1 to 7. To measure participants’ initial beliefs about the scientific consensus on global warming in the mixed design, we asked the following question; “To the best of your knowledge, what percentage of climate scientists have concluded that human-caused global warming is happening? (0% – 100%).” Participants in the mixed design answered these questions both before and after receiving the treatment or control condition whereas participants in the between-subjects design answered these questions only

after receiving the treatment or control condition. Further, participants in the mixed design completed a distractor task between pre- and post-test questions in which they read a brief description of an upcoming Star Wars movie and reported the likelihood that they would see the movie (1 = *Very unlikely*, 5 = *Very likely*). Next, participants were randomly assigned to either a control condition or a treatment condition³. In the control condition, participants completed a neutral word-sorting task in which they dragged-and-dropped words into their corresponding category. For example, “yellow” would be placed in the colors category and “pan” would be placed in the cookware category. In the treatment condition, participants read the following normative statement: “97% of climate scientists have concluded that human-caused global warming is happening.” After their corresponding condition, participants were shown the treatment message and asked “How *familiar* were you with that statement before you read it in this survey?” (1 = *Not at all familiar*, 7 = *Very familiar*). Other questions (e.g., about social norms) were asked for exploratory purposes and can be found on our OSF project page.

Results

For the mixed design, we conducted an ANCOVA with condition (control vs. treatment) predicting post-treatment estimates of scientific consensus while controlling for pre-treatment consensus estimates (e.g., Dugard & Todman, 1995; Oakes & Feldman, 2001; Rausch et al., 2003). For the between-subjects design, we used independent samples *t*-tests.

MTurk

³ For exploratory purposes, we tested a second control condition (a 97% cue unrelated to climate change) and two treatment conditions with minor wording variations. Because there were no differences between control conditions or among treatment conditions, we collapsed across control conditions and across treatment conditions for simplicity and to maintain high statistical power. See Supplementary Information for analyses testing for differences between these conditions.

Analyzing the data from the mixed design, the ANCOVA was significant, $F(1, 527) = 66.53, p < .001, \eta^2_{\text{partial}} = .112$, such that the consensus treatment significantly increased consensus estimates compared to the control group, ($M_{\text{diff}} = 8.62, SE = 1.06, p < .001, 95\% \text{ CI } [6.55, 10.70], d = .51$). For the between-subjects design, the independent samples t -test was also significant in the predicted direction, $t(528) = 4.93, p < .001, 95\% \text{ CI } [4.45, 10.37], d = .30$.

TurkPrime Panels

The ANCOVA for the mixed design was significant, $F(1, 517) = 136.39, p < .001, \eta^2_{\text{partial}} = .209$, indicating that the consensus treatment caused significantly greater estimates of scientific consensus compared to the control group, ($M_{\text{diff}} = 15.03, SE = 1.29, p < .001, 95\% \text{ CI } [12.50, 17.56], d = .71$). Likewise, data from the between-subjects design showed a significant consensus treatment effect, $t(518) = 8.74, p < .001, 95\% \text{ CI } [12.91, 20.39], d = .50$.

Facebook

General sample. An ANCOVA for the mixed design was significant, $F(1, 423) = 26.91, p < .001, \eta^2_{\text{partial}} = .060$, which, as expected, demonstrated that the consensus message led to significantly greater estimates of scientific consensus than did the control condition, ($M_{\text{diff}} = 4.57, SE = .88, p < .001, 95\% \text{ CI } [2.84, 6.30], d = .16$). However, unexpectedly, the analysis of the data from the between-subjects participants design was not significant, $t(424) = 1.59, p = .112, 95\% \text{ CI } [-1.05, 10.00], d = .04$.

Targeted sample. Because of the unexpected finding in the initial Facebook sample, we examined why there was a null finding for that sample, in contrast to strong, significant effects in the other two samples. We hypothesized that participants recruited via Facebook were more familiar with the consensus message and, therefore, less likely to update their estimates because they had already been influenced prior to the experiment (see Druckman & Leeper, 2012). An

ANOVA indicated that familiarity significantly differed across platforms, $F(2, 3129) = 148.79$, $p < .001$, $\eta^2_{\text{partial}} = .087$, and as hypothesized, Facebook participants were significantly more familiar with the consensus message than were MTurk participants, ($M_{\text{diff}} = .90$, $SE = .09$), $p < .001$, 95% CI [.73, 1.07], $d = .47$, or TurkPrime Panels participants, ($M_{\text{diff}} = 1.44$, $SE = .08$), $p < .001$, 95% CI [1.28, 1.61], $d = .75$. Additionally, MTurk participants were significantly more familiar with the consensus message than were TurkPrime Panels participants, ($M_{\text{diff}} = .54$, $SE = .09$), $p < .001$, 95% CI [.37, .71], $d = .23$.

To test the degree to which the lack of a treatment effect in a Facebook between-subjects sample can be explained by differences in familiarity with the consensus message, we recruited an additional Facebook sample that was likely less familiar with the consensus message. To do this, we used Facebook advertising to target and recruit respondents from two populations that are less politically engaged: (1) political moderates and (2) individuals who do not follow any major political figures on Facebook. Confirming the success of this targeting strategy, an independent samples t -test showed that the targeted Facebook sample was significantly less familiar with the consensus message ($M = 4.83$, $SD = 1.89$) than was the Facebook General sample ($M = 5.38$, $SD = 1.80$), $t(2125) = 6.89$, $p < .001$, 95% CI [.40, .71], $d = .30$.

Next, we conducted the same analyses with the targeted Facebook sample as we did with the other samples. The ANCOVA on the mixed design data was significant, $F(1, 438) = 69.27$, $p < .001$, $\eta^2_{\text{partial}} = .137$, such that the consensus treatment led to greater estimates of scientific consensus compared to the control group, ($M_{\text{diff}} = 7.24$, $SE = .87$), $p < .001$, 95% CI [5.53, 8.94], $d = .36$. For the between-subjects design data, an independent samples t -test showed that the consensus treatment versus control difference was also significant in the expected direction, $t(439) = 3.63$, $p < .001$, 95% CI [3.20, 10.77], $d = .35$.

Differences in Baseline Consensus Estimates

To explore whether differences in experimental results might be due to differences in baseline estimates of the scientific consensus, we calculated the pre-test sample means of the estimates of participants in the mixed design for each sampling source. As displayed in Table 2, there were substantial baseline differences between samples. Between-sample baseline differences will be particularly consequential for the between-subjects design, given that their results do not account for these differences.

Table 2. Baseline estimates of the scientific consensus for each sample.

Platform	MTurk	TurkPrime Panels	Facebook General	Facebook Targeted
Pre-test				
Consensus	78.30 (17.73)	69.82 (21.67)	75.47 (29.01)	81.12 (20.18)
Estimate: $M(SD)$				

Internal Meta-analysis

We conducted two random-effects meta-analyses: one to explore heterogeneity in effect-sizes across studies due to study sample (MTurk, TurkPrime Panels, Facebook General, Facebook Targeted) and one to explore heterogeneity in effect-sizes across studies due to study design (mixed vs. between-subjects). Heterogeneity is described by the I^2 statistic, or the percentage of variation across studies that is due to heterogeneity rather than random chance (Higgins & Thompson, 2001). In addition to a frequentist analysis, Bayesian credible intervals for the overall pooled effects from a random-effects model are also reported. They were generated using Markov Chain Monte Carlo simulations (MCMC) with the bmeta package in R (Ding & Baio, 2016). Results are reported in Figures 1 and 2.

Variability in effect sizes due to sample. Random-effects meta-analyses with the four samples were conducted separately for the mixed-design studies and between-subjects studies. For between-subjects studies, there was significant heterogeneity across samples, ($I^2 = 80.4\%$, p

= .002). Further, effect sizes varied considerably, ranging from $d = .04$, 95% CI [-0.12, .21] for the Facebook General sample, to $d = .50$, 95% CI [.33, .67] for the TurkPrime Panels sample. For mixed-design studies, there was also significant heterogeneity across samples, ($I^2 = 71.0\%$, $p = .016$), although significantly less than between-subjects studies (see Figure 1). Effect sizes in the mixed-design studies also varied considerably, but less than in between-subjects studies. Effect sizes for mixed-design studies ranged from $d = .50$, 95% CI [.30, .69] for the Facebook General sample, to $d = .93$, 95% CI [.74, 1.11] for the TurkPrime Panels sample. It is worth noting, for both mixed and between designs, that these effect size ranges have 95% confidence intervals that do not overlap, demonstrating that they differ significantly from each other at least at the .01 alpha level (Cumming, 2009).

Figure 1. Random-effects meta-analysis for sample type.

Note. For mixed designs we used the pre-post difference score as the dependent measure.

Variability in effect sizes due to study design. To understand what proportion of heterogeneity across studies is due to study design, we conducted a random-effects meta-analysis separately for each platform with mixed-design and between-subjects samples. This allowed for us to test whether there was more heterogeneity across mixed-design versus between-subjects studies. There was significant heterogeneity across studies due to study design, which was consistent across all four samples (see Figure 2). In all cases, mixed designs yielded significantly larger effect sizes compared to between-subject designs. For example, the weighted mean effect size for mixed design studies was more than twice that of between-subjects design studies ($d_{\text{mixed}} = .73$, $d_{\text{between}} = .30$). However, across all samples, the weighted pooled consensus-effect was positive and significant, $d = 0.51$, 95% CI [0.31, 0.71], 95% Bayesian credible interval [0.25, 0.77].

Figure 2. Random-effects meta-analysis for study design.

Note. For mixed designs we used the pre-post difference score as the dependent measure.

Potential Remedies for Differences in Sample Characteristics

Because the results of even identical experiments are likely to vary from study to study, it is important to understand whether results differ because of sample characteristics or because the effect of interest is unreliable. For example, in our Facebook General sample, participants reported that they were significantly more familiar with the “97% of climate scientists...” norm before they participated in the study than did participants in the other samples (i.e., they exhibited stronger pre-treatment effects). If the participants who opted in to the study on Facebook had, before the study, already engaged with the consensus message by updating their beliefs to be more in line with the scientific consensus, or perhaps resisted being persuaded, it is substantially less likely that the experimental manipulation would move their estimates of the

scientific consensus. In other words, a study on this sample would, therefore, underestimate the population effect size of the consensus treatment.

Other characteristics, such as education and political ideology, likely play a similar role. To illustrate, we compare the demographic breakdown of the sample with the smallest effect sizes (Facebook General) and the largest effect sizes (TurkPrime Panels). As is evident in Table 1, the Facebook General sample was more educated, $t(2125) = 14.31, p < .001, 95\% \text{ CI } [.52, .69], d = .62$, and more liberal, $t(2127) = -6.74, p < .001, 95\% \text{ CI } [-.47, -.26], d = -.46$, than the TurkPrime Panels sample. It is likely that more educated and liberal participants have already been exposed to and persuaded by the scientific consensus on climate change compared to those who are less educated and more conservative (Zhang et al., 2018). For example, in large nationally representative samples, the perceived consensus estimate typically hovers around 65% (van der Linden et al., 2018), which is lower than in the convenience samples reported here (about 70% – 81%) and elsewhere (Harris et al., 2018; Kerr & Wilson, 2018).

One method to see whether sample differences account for differences in effect sizes is to weight the samples to see whether effect sizes become more similar when known between-sample differences are accounted for. If so, this may help explain potential (non)replication of experimental effects. It can also provide valuable information about whether an experimental treatment has a stronger influence on some people than others.

To test whether adjusting for sample differences via weighting increases the similarity of effect sizes, we used rake weighting (see Kalton & Flores-Cervantes, 2003) to correct for demographic and pre-treatment differences between our samples. Rake weighting is a model-based approach that adjusts survey totals on the assigned variables to match known population totals. We weighted all samples, using our TurkPrime Panels sample as the benchmark. We did

this for two reasons: (a) out of all our samples, TurkPrime was most similar demographically to the general United States population; and (b) TurkPrime had the largest effect size and weighting to that sample provides insight into how much its effect size is driven by differences in sample composition. Generally speaking, weighting to sample benchmarks from previous studies is useful because it is likely that most researchers will not have population benchmarks for some measures on which they wish to weight (e.g., message familiarity), but do have data from previous studies with which to create benchmarks.

We created three different weight variables for each sample using the RAKE command in the SPSS software version 24. The weight variables were benchmarked to: (a) demographics (sex, age, race, and education); (b) those demographics plus political ideology (liberal, moderate, conservative); and (c) consensus message familiarity. The syntax used to create these weights is available on our OSF project page. Unweighted versus weighted results are presented in Table 3.

Table 3. Effect sizes (Cohen's d) and 95% confidence intervals for each sample.

Platform	Design	Weight Characteristics			
		No Weight	Demographics	Demographics + Ideology	Familiarity
		d [95 % CI]	d [95 % CI]	d [95 % CI]	d [95 % CI]
MTurk	Mixed	.51 [.33, .69]	.85 [.66, 1.04]	.91 [.73, 1.09]	.51 [.33, .69]
MTurk	Between	.30 [.12, .49]	.37 [.16, .57]	.37 [.16, .58]	.30 [.12, .49]
TurkPrime Panels	Mixed	.71 [.52, .89]	<i>NA</i>	<i>NA</i>	<i>NA</i>
TurkPrime Panels	Between	.50 [.33, .67]	<i>NA</i>	<i>NA</i>	<i>NA</i>
Facebook General	Mixed	.16 [-.03, .35]	.19 [-.01, .39]	.21 [.00, .41]	.22 [.02, .41]
Facebook General	Between	.04 [-.12, .21]	-.02 [-.20, .15]	.09 [-.08, .27]	.28 [.11, .45]
Facebook Targeted	Mixed	.36 [.17, .56]	.26 [.06, .46]	.21 [-.01, .43]	.38 [.20, .56]
Facebook Targeted	Between	.35 [.19, .51]	.43 [.27, .59]	.41 [.24, .58]	.37 [.22, .52]

Note. Weighted results for TurkPrime Panels are marked as *NA* because weights were created for the other samples using the TurkPrime Panels sample as the benchmark.

Additionally, Table 4 displays differences in effect size between each sample and the effect sizes observed in the TurkPrime Panels sample. For example, the difference between the unweighted MTurk effect size with the between-subjects design and the unweighted TurkPrime

Panels effect size with the same design was $-.20$, indicating that the MTurk effect size using this design had a Cohen's d that was $.20$ smaller than that observed for TurkPrime Panels. When sample weights were used to adjust for demographic differences, the difference between the Cohen's d s decreased to $.13$ (see Table 4). However, when ideology was also considered, the difference remained at $.20$. This indicates that differences in effect sizes between the two samples are less likely due to ideological differences and more likely due to the TurkPrime sample being older, less educated, and/or more female. Additionally, for this same comparison, the familiarity weight made no difference in effect size.

Table 4. Differences in effect size (Cohen's d) between each sample and TurkPrime Panels.

		Weight Characteristics			
		No Weight	Demographics	Demographics + Ideology	Familiarity
MTurk	Mixed	-.20	.14	.20	-.20
MTurk	Between	-.20	-.13	-.13	-.20
Facebook General	Mixed	-.55	-.52	-.50	-.49
Facebook General	Between	-.46	-.52	-.41	-.22
Facebook Targeted	Mixed	-.35	-.45	-.50	-.33
Facebook Targeted	Between	-.15	-.07	-.09	-.13

Note. Negative numbers mean that effect size was lower than the effect size observed for the TurkPrime Panels sample using the same design. Smaller numbers mean effect size was more similar to those observed in the TurkPrime Panels sample.

In contrast, weights that included political ideology and weighting based on pre-treatment consensus message familiarity had a greater effect on the results obtained from the Facebook General sample. The unweighted effect size for the Facebook General sample was $.46$ smaller than that observed for the TurkPrime Panels sample. This difference increased when applying the demographic weights ($.52$), but the difference decreased when weighting based on political ideology in addition to demographics ($.41$). This indicates that the differences between the Facebook General versus the TurkPrime Panels effect size for the between-subjects design are likely due to between-platform ideological differences. The difference in effect size became even

smaller when weighting on message familiarity alone (.22). Together, these results indicate that effect size differences between Facebook General and TurkPrime Panels were more likely due to differences in message familiarity than demographics alone or the combination of demographics and ideology.

To summarize the influence the weights had on effect size differences, we examine how many times the weights made effect sizes more similar to the benchmark sample (TurkPrime Panels). Out of 18 possible cases (3 samples * 2 designs each * 3 weights each), there were 12 cases (67%) in which the difference between effect sizes decreased, 3 (17%) where it increased, 2 (11%) where there was no difference, and 1 (6%) in which the effect size became larger than was observed in TurkPrime Panels (although the absolute difference was the same). Below, we discuss situations in which researchers should expect effect sizes to converge or diverge depending on the characteristics on which their samples differ.

Discussion

The current experiments indicate that the effect of highlighting the scientific consensus on climate change (a descriptive norm) is robust across different sampling platforms and population characteristics; the lone exception was the Facebook General sample using a between-subjects design. However, our internal meta-analysis also indicates that there was significant variability in effect sizes depending on the sample and design. For example, effect sizes observed for TurkPrime Panels were large by social science standards (Cohen, 1992), whereas the effect sizes for the Facebook General sample were indistinguishable from zero. We demonstrate that these differences are due, at least in part, to differences in sample characteristics; the Facebook General sample was significantly more educated and liberal than was the TurkPrime sample, making it less likely that our experimental manipulation in the

Facebook General sample would lead to a statistically detectable effect. Additionally, large differences in familiarity with the consensus treatment message played an important role in explaining effect size differences across samples. This was evident such that samples with greater familiarity with the message had smaller effect sizes. We confirmed this by recruiting a targeted Facebook sample that was less familiar with the treatment message, which demonstrated larger effect sizes.

We further explored the importance of sample differences in baseline beliefs by comparing results from mixed versus between-subjects experimental designs. Mixed design experiments consistently produced larger effect sizes than did those employing a between-subjects design. This supports the finding that experiments that include a pre-test are statistically more powerful than those using a post-only design (Charness, Gneezy, & Kuhn, 2012). This highlights the importance of measuring pre-message attitudes in research on highly politicized topics like climate change, because samples are likely to differ in their baseline beliefs.

Connecting to Previous Research

The current findings are consistent with research concluding that some phenomena are sensitive to context, which may explain non-replication of results (Van Bavel et al., 2016). Further, our findings showing diminished effect sizes for sample participants who are more familiar with the information presented in the treatment support the conclusion that non-naïveté can be a problem in psychology studies (Chandler et al., 2015; Druckman & Leeper, 2012).

Interestingly, our conclusion that effect sizes are heavily sample-dependent (i.e., heterogeneous) differs from the conclusion of the Many Labs 2 project (Klein et al., 2018). In a large-scale replication study, researchers aimed to replicate 28 published effects across approximately 60 samples per effect with participants from 36 different countries and territories

(Klein et al., 2018). The researchers concluded that, while effect sizes varied dramatically depending on the effect being studied, there was little heterogeneity *within* any given effect being studied. Contrary to this conclusion, we believe that for some of the effects Klein et al. tested, their data show substantial heterogeneity. For example, there are several effects where effect sizes ranged from zero to large (Klein et al., 2018; Figure 2). Additionally, I^2 statistics provided in their manuscript show that, for many effects studied, a substantial proportion of the variability in effect size is *not* due to sampling error, which is consistent with the results of the current study. We argue, as recommended by Borenstein, Higgins, Rothstein, and Hedges (2015), that heterogeneity statistics such as I^2 be reported alongside the actual range of effects in order to more fully understand effect size heterogeneity.

Implications for Basic Research

Basic researchers might argue that variability in effect sizes is to be expected and theory-testing is primarily focused on internal validity and the direction of an effect rather than the size of it. We argue that these researchers should be concerned with effect size estimates and heterogeneity (see Cook, Campbell, & Shadish, 2002; McShane et al., 2018). First, precise estimates of effect size are crucial for estimating the sample size needed to achieve the desired statistical power, and therefore important for planning studies designed for theory-testing and replication. Second, variation in effect sizes provides information about potential moderators that should be considered as part of a theory. For example, as McShane and colleagues (2018) argue, effect size heterogeneity should be embraced because it signals a need for theory improvement. Our weighting method is useful for this purpose because it helps disentangle whether effect sizes differ because of sample characteristics or because of unmeasured moderators. If effect sizes are still substantially different from one another after weighting, it suggests that there is a need to

investigate in future studies the factors that may not have been previously measured (e.g., participants' motivation for participating).

Implications for Applied Research

The findings reported here are crucial for applied research, which is concerned with the practical value of study results. To use the current experiments as an example, those aiming to improve public engagement with the issue of climate change need to decide—with limited time and resources—which messages to build into their communication strategy. For applied questions such as this, the effect size on the target population is of primary interest. If that effect size varies dramatically depending on sample characteristics (e.g., education, message familiarity), then this knowledge can enhance the value of an applied approach for example, via geographical targeting of those individuals least likely to be familiar with the message (Zhang et al., 2018). Our weighting method provides a way for researchers to better understand why results from two identical experiments differ in effect sizes, which is a crucial step toward understanding which interventions will be more or less effective across different contexts.

Recommendation for Researchers

A key question remains: where should researchers recruit participants for their experimental social psychological research? The answer depends on what question the researcher is asking and to whom the researcher wishes to generalize the results—two factors that are inextricably linked. Research questions may apply differently to different populations and the effect of an experimental manipulation can vary depending on the underlying characteristics of the sample. As is clear from the current results, the answer to the research question “To what extent does communicating the expert consensus on climate change lead people to update their perception of scientific agreement?” depends, in part, on sample characteristics such as

education, ideology, and message familiarity. This answer is important for both basic and applied reasons. From a basic perspective, the sign of the updating is relevant, especially in light of competing theories about the motivated cognition of scientific information (van der Linden et al., 2018). From an applied perspective, these findings are useful because changes in normative consensus beliefs are known to have important downstream consequences on private attitudes (Lewandowsky et al., 2013; Kerr & Wilson, 2018; van der Linden et al., 2019).

To address issues associated with differences in sample composition, we advise researchers to explicitly define their population of interest and sampling frame, and aim to match the sampling frame to the target population to the greatest extent possible. There is often an implicit assumption that findings broadly generalize to the rest of the population, and that follow-up research determines the boundaries of this generalizability (Simons, Shoda, & Lindsay, 2017). However, these assumptions can lead researchers to commit Type I or Type II errors, depending on how they draw their sample. To use the current experiments as examples, if we had only collected the Facebook General sample using a between-subjects design, we would have concluded that communicating the scientific consensus on climate change has negligible effects on public perception. If we had only collected the TurkPrime Panels sample using a mixed design, we would have concluded that highlighting the scientific consensus leads to large shifts in public perception. In our targeted Facebook sample, we defined our population of interest by sampling participants who were less politically engaged and, therefore, less likely to have already seen the consensus message. Doing so allowed us to examine the boundaries of the scientific consensus effect, rather than starting with the assumption that the experimental manipulation would apply broadly to everyone.

As such, we agree with Simons and colleagues' (2017) proposal that all empirical manuscripts include a “Constraints on Generality” (COG) statement, where authors explicitly identify and justify the target population to which they expect the reported findings to generalize. There are several important incentives for doing so. For instance, a COG statement would help make clear what the appropriate target population is for sampling in a replication study to make the replication more comparable to the original study and to better justify its conclusions.

In addition, given our findings, we recommend that if resources permit, researchers should replicate experiments with different samples or collect different samples in parallel with one another. This allows for an empirical test of the boundary conditions of one's findings, rather than assuming, even implicitly, that the results are universally generalizable (Simons et al., 2017). Further, this is beneficial because it allows researchers to test the extent to which measured sample characteristics (e.g., demographics, prior knowledge) explain differences in effect sizes, aiding theory development (McShane et al., 2018; Slater & Gleason, 2012). In the current paper, we use weighting as a method to help researchers determine whether replication failures can perhaps be explained by differences in such sample characteristics. If a researcher conducts identical experiments with two different samples, they can weight one sample to match the other to empirically test if their effect sizes become more alike. This can provide insight into whether one or a combination of sample characteristics is responsible for different experimental effects. If the effect size of the weighted sample remains the same, this would suggest that the samples likely differ on a variable that has not been measured. If effect sizes converge, this indicates that sample differences influenced the results.

In our test of this strategy, most cases led effect sizes to converge. However, in a few cases, there was either no difference or the effect sizes actually diverged. It may be most useful,

however, to consider the results in the context of how sample characteristics are likely to affect results, rather than to focus primarily on whether weights caused the results to converge or diverge. For example, if one hypothesizes that highly educated participants will have more crystalized attitudes on a topic, and are therefore less likely to change their beliefs, weighting to match a less educated sample should increase the effect size.

As a cautionary note, researchers must be careful about using this technique to draw conclusions about non-replication. Rather, it can be used as a guide for understanding how sample differences affected results. When recruiting new samples, weighting, followed by empirical testing, can help confirm whether results become more similar as samples are made more comparable. This follows the same procedure we used for our targeted Facebook sample with the goal of recruiting participants who were less familiar with our treatment message.

Additionally, sub-sample size must be considered when assessing the feasibility of weighting on any given demographic group. For example, if African-American men make up only 2% of a sample and researchers apply a large weight so they match a population parameter of 20%, it assumes that the attitudes of a tiny sub-sample can be generalized to the full demographic category, which is rarely the case. To guard against prohibitively large weights (i.e., design effects), it is important to consider whether subsample sizes are sufficient to weight on the variables of interest.

The primary limitations to this study are (a) it was limited to the topic of attitudes about climate change and (b) we did not explore the role of cultural factors (e.g., United States vs. Europe) as a source of heterogeneity. The findings observed by Van Bavel and colleagues (2016), however, lead us to suspect that, like research on climate change, research on other highly politicized issues (e.g., vaccines, guns, GMOs) will also be sensitive to context.

Therefore, those results would also likely vary across samples, culture, locations, and populations. Additionally, the Many Labs 2 project observed significant heterogeneity in 39% of effects being studied, indicating that many social psychological phenomena are sensitive to context and sample characteristics (Klein et al., 2018). For individual phenomena, however, it is still necessary to empirically test rather than assume the boundary conditions of one's results.

To conclude, convenience samples can be an inexpensive method to conduct high-quality research. It is important, however, for researchers to critically evaluate how differences in characteristics between samples may affect their results. Testing experimental effects across multiple sample sources is useful not only for evaluating the generalizability of empirical results, but also for formulating theories that predict variability in effect sizes. Doing so can support a more robust and cumulative psychological science.

References

- Anderson, C. A., Allen, J. J., Plante, C., Quigley-McBride, A., Lovett, A., & Rokkum, J. N. (2018). The MTurkification of Social and Personality Psychology. *Personality and Social Psychology Bulletin*, 0146167218798821.
- Ballew, M. T., Goldberg, M. H., Rosenthal, S. A., Cutler, M. J., & Leiserowitz, A. (2019). Climate Change Activism Among Latino and White Americans. *Frontiers in Communication*, 3(58), 1-15.
- Berinsky, A. J., Huber, G. A., & Lenz, G. S. (2012). Evaluating online labor markets for experimental research: Amazon. com's Mechanical Turk. *Political analysis*, 20(3), 351-368.

- Boas, T. C., Christenson, D. P., & Glick, D. M. (2018). Recruiting Large Online Samples in the United States and India: Facebook, Mechanical Turk and Qualtrics. *Political Science Research and Methods*, Forthcoming.
- Borenstein, M., Higgins, J. P., Rothstein, H. R., & Hedges, L. V. (2015). I^2 Is Not an Absolute Measure of Heterogeneity in a Meta-Analysis. *Unpublished manuscript*.
- Campbell, D. T. (1957). Factors relevant to the validity of experiments in social settings. *Psychological bulletin*, 54(4), 297.
- Champley, S., Ekstrom, C., Dalaard, P., Gill, J., Weibelzahl, S., Anandkumar, A., Ford, C., Volcic, R., De Rosario, H. (2018). pwr: Basic Functions for Power Analysis. R package version 1.2-2.
- Chandler, J., Paolacci, G., Peer, E., Mueller, P., & Ratliff, K. A. (2015). Using nonnaive participants can reduce effect sizes. *Psychological Science*, 26(7), 1131-1139.
- 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.
- Clifford, S., Jewell, R. M., & Waggoner, P. D. (2015). Are samples drawn from Mechanical Turk valid for research on political ideology?. *Research & Politics*, 2(4), 2053168015622072.
- Cohen, J. (1992). A power primer. *Psychological Bulletin*, 112(1), 155-159.
- Cook, J., & Lewandowsky, S. (2016). Rational irrationality: Modeling climate change belief polarization using Bayesian networks. *Topics in Cognitive Science*, 8(1), 160-179.
- Cook, T. D., Campbell, D. T., & Shadish, W. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston, MA: Houghton Mifflin.

- Cumming, G. (2009). Inference by eye: reading the overlap of independent confidence intervals. *Statistics in Medicine*, 28(2), 205-220.
- Ding, D., Maibach, E. W., Zhao, X., Roser-Renouf, C., & Leiserowitz, A. (2011). Support for climate policy and societal action are linked to perceptions about scientific agreement. *Nature Climate Change*, 1(9), 462-466.
- Ding, T. & Baio, G. (2016). bmeta: Bayesian Meta-Analysis and Meta-Regression. R package version 0.1.2.
- Dixon, G. (2016). Applying the gateway belief model to genetically modified food perceptions: New insights and additional questions. *Journal of Communication*, 66(6), 888-908.
- Dorofeev, S., & Grant, P. (2006). *Statistics for real-life sample surveys: Non-simple-random samples and weighted data*. Cambridge University Press.
- 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.
- Dugard, P., & Todman, J. (1995). Analysis of pre-test-post-test control group designs in educational research. *Educational Psychology*, 15(2), 181-198.
- Dunlap, R. E., McCright, A. M., & Yarosh, J. H. (2016). The political divide on climate change: Partisan polarization widens in the US. *Environment: Science and Policy for Sustainable Development*, 58(5), 4-23.
- Gilbert, D. T., King, G., Pettigrew, S., & Wilson, T. D. (2016). Comment on “Estimating the reproducibility of psychological science”. *Science*, 351(6277), 1037-1037.

- Harris, A. J., Sildmäe, O., Speekenbrink, M., & Hahn, U. (2018). The potential power of experience in communications of expert consensus levels. *Journal of Risk Research*, 1-17.
- Henrich, J., Heine, S. J., & Norenzayan, A. (2010). Most people are not WEIRD. *Nature*, 466(7302), 29.
- Henry, P. J. (2008). College sophomores in the laboratory redux: Influences of a narrow data base on social psychology's view of the nature of prejudice. *Psychological Inquiry*, 19(2), 49-71.
- Higgins, J. P., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in medicine*, 21(11), 1539-1558.
- Huck, S. W., & McLean, R. A. (1975). Using a repeated measures ANOVA to analyze the data from a pretest-posttest design: A potentially confusing task. *Psychological Bulletin*, 82(4), 511.
- Kahan D. M. (2013). What's a "valid" sample? Problems with Mechanical Turk study samples, part 1. Cultural Cognition Project Blog. Available at: <http://www.culturalcognition.net/blog/2013/7/8/whats-a-valid-sample-problems-with-mechanical-turk-study-sam.html> (accessed 3 May 2015).
- Kahan, D. M., Jenkins-Smith, H., & Braman, D. (2011). Cultural cognition of scientific consensus. *Journal of Risk Research*, 14(2), 147-174.
- Kalton, G., & Flores-Cervantes, I. (2003). Weighting methods. *Journal of Official Statistics*, 19(2), 81.
- Kerr, J. R., & Wilson, M. S. (2018). Changes in perceived scientific consensus shift beliefs about climate change and GM food safety. *PloS one*, 13(7), e0200295.

- 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.
- Krupnikov, Y., & Levine, A. S. (2014). Cross-sample comparisons and external validity. *Journal of Experimental Political Science*, 1(1), 59-80.
- Landrum, A. R., Hallman, W. K., & Jamieson, K. H. (2018). Examining the Impact of Expert Voices: Communicating the Scientific Consensus on Genetically-modified Organisms. *Environmental Communication*, 1-20.
- Leiserowitz, A., Maibach, E., Rosenthal, S., Kotcher, J., Bergquist, P., Ballew, M., Goldberg, M., & Gustafson, A. (2019). *Climate change in the American mind: April 2019*. Yale University and George Mason University. New Haven, CT: Yale Program on Climate Change Communication. doi:10.31219/osf.io/3bwj8
- Leiserowitz, A., Rosenthal, S., & Cutler, M. (2018). *Latinos and Global Warming's Six Americas: May 2017*. Yale University. New Haven, CT: Yale Program on Climate Change Communication.
- Lewandowsky, S., Gignac, G.E., & Vaughan, S. (2013). The pivotal role of perceived scientific consensus in acceptance of science. *Nature Climate Change* 3(4), 399-404.
- 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., & Xiao, C. (2013). Perceived scientific agreement and support for government action on climate change in the USA. *Climatic Change*, 119(2), 511-518.
- McShane, B. B., Tackett, J. L., Bockenholt, U., & Gelman, A. (2018). Large scale replication projects in contemporary psychological research. *The American Statistician*.

- Mullinix, K. J., Leeper, T. J., Druckman, J. N., & Freese, J. (2015). The generalizability of survey experiments. *Journal of Experimental Political Science*, 2(2), 109-138.
- Oakes, J. M., & Feldman, H. A. (2001). Statistical power for nonequivalent pretest-posttest designs: The impact of change-score versus ANCOVA models. *Evaluation Review*, 25(1), 3-28.
- Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251), aac4716.
- Rausch, J. R., Maxwell, S. E., & Kelley, K. (2003). Analytic methods for questions pertaining to a randomized pretest, posttest, follow-up design. *Journal of Clinical Child and Adolescent Psychology*, 32(3), 467-486.
- Schuldt, J. P., & Pearson, A. R. (2016). The role of race and ethnicity in climate change polarization: evidence from a US national survey experiment. *Climatic Change*, 136(3-4), 495-505.
- Schwarz, N. (1999). Self-reports: how the questions shape the answers. *American psychologist*, 54(2), 93.
- Sears, D. O. (1986). College sophomores in the laboratory: Influences of a narrow data base on social psychology's view of human nature. *Journal of personality and social psychology*, 51(3), 515.
- Simons, D. J., Shoda, Y., & Lindsay, D. S. (2017). Constraints on generality (COG): A proposed addition to all empirical papers. *Perspectives on Psychological Science*, 12(6), 1123-1128.
- Slater, M. D., & Gleason, L. S. (2012). Contributing to theory and knowledge in quantitative communication science. *Communication Methods and Measures*, 6(4), 215-236.

- Van Bavel, J. J., Mende-Siedlecki, P., Brady, W. J., & Reinero, D. A. (2016). Contextual sensitivity in scientific reproducibility. *Proceedings of the National Academy of Sciences*, 113(23), 6454-6459.
- van der Linden, S., Leiserowitz, A.A., Feinberg, G.D., & Maibach, E.W. (2015). The scientific consensus on climate change as a gateway belief: Experimental evidence. *PloS One*, 10(2), e0118489.
- van der Linden, S., Leiserowitz, A., & Maibach, E. (2018). Scientific agreement can neutralize politicization of facts. *Nature Human Behaviour*, 2(1), 2-3
- van der Linden, S., Leiserowitz, A., & Maibach, E. (2019). The Gateway Belief Model: A Large-Scale Replication. *Journal of Environmental Psychology*, 62, 49-58.
- Weinberg, J. D., Freese, J., & McElhattan, D. (2014). Comparing Data Characteristics and Results of an Online Factorial Survey between a Population-based and a Crowdsourced-recruited Sample. *Sociological Science*, 1.
- Willson, V. L., & Putnam, R. R. (1982). A meta-analysis of pretest sensitization effects in experimental design. *American Educational Research Journal*, 19(2), 249-258.
- Wood, C., Conner, M., Miles, E., Sandberg, T., Taylor, N., Godin, G., & Sheeran, P. (2016). The impact of asking intention or self-prediction questions on subsequent behavior: a meta-analysis. *Personality and Social Psychology Review*, 20(3), 245-268.
- Zhang, B., van der Linden, S., Mildemberger, M., Marlon, J. R., Howe, P. D., & Leiserowitz, A. (2018). Experimental effects of climate messages vary geographically. *Nature Climate Change*, 8(5), 370-374.