A Review of Multisite Replication Projects in Social Psychology: Is It Viable to Sustain Any Confidence in Social Psychology's Knowledge Base?

Article *in* Perspectives on Psychological Science · November 2022

DOI: 10.1177/17456916221121815

CITATIONS
27

READS
875

3 authors:

Roy Baumeister
Harvard University
713 PUBLICATIONS **155,382** CITATIONS
SEE PROFILE

Dianne Tice
Brigham Young University
87 PUBLICATIONS **28,555** CITATIONS
SEE PROFILE

Brad J Bushman
The Ohio State University
354 PUBLICATIONS **42,254** CITATIONS
SEE PROFILE

Perspectives on Psychological Science, in press

A Review of Multi-site Replication Projects in Social Psychology: Is It Viable to Sustain Any Confidence in Social Psychology's Knowledge Base?

Roy F. Baumeister, *University of Queensland*
Dianne M. Tice, *Brigham Young University*
Brad J. Bushman, *The Ohio State University*

Keywords: replication, replication crisis, manipulation checks, multi-lab replication, multi-site replication

Correspondence

Correspondence concerning this article should be addressed to Brad J. Bushman, School of Communication, The Ohio State University, 3016 Derby Hall, 154 N. Oval Mall, Columbus, OH 43210, USA. Email: bushman.20@osu.edu

Preregistration Statement

The reported research was preregistered and all data and materials are available online: https://osf.io/ar2pg

Abstract

Multi-site (multi-lab/many-lab) replications have emerged as a popular way of verifying prior research findings, but their record in social psychology has prompted distrust of the field and a sense of crisis. We review all 36 multi-site social psychology replications (plus three articles reporting multiple mini-studies). We start by assuming both the original and the multi-site replications were conducted in honest and diligent fashion, despite often yielding different conclusions. Four of the 36 (11%) were clearly successful in terms of providing significant support for the original hypothesis, and five others (14%) had mixed results. The remaining 27 (75%) were failures. Multiple explanations for the generally poor record of replications are considered and relevant evidence assessed, including: original hypothesis was wrong; hypothesis not tested because of operational failure; low engagement of participants; and bias toward failure. There was evidence for each of these, with low engagement emerging as a widespread problem (reflected in high rates of discarded data and weak manipulation checks). The few procedures with actual interpersonal interaction fared much better than others. We discuss implications in relation to manipulation checks, effect sizes, and impact on the field and offer recommendations for improving future multi-site projects.

A Review of Multi-site Replication Projects in Social Psychology:
Is It Viable to Sustain Any Confidence in Social Psychology's Knowledge Base?

Concern over the replicability of psychological science has risen in recent years, especially in social psychology. A combination of events led to a general sense of a "replication crisis." Anecdotal reports of failure stimulated a large-scale attempt to replicate 100 previously published findings (Open Science Collaboration, 2015). Less than half were significant, with social psychology faring worse than cognitive psychology. Even among successes, effect sizes were considerably smaller than the original effects. In an unrelated development, several high-profile cases of researcher fraud were found in social psychology (e.g., Diederik Stapel), and the fact that fraud happens at all makes one suspicious of the field's knowledge base. Fraudulent findings are of course highly unlikely to replicate. Yet another unrelated development, Bem's (2011) findings of precognition, published in social psychology's leading journal (*Journal of Personality and Social Psychology*), seemed *a priori* implausible to many readers, raising further doubt about the validity of the knowledge base.

The importance of replication has led to many researchers to embrace multi-site replications as a presumably rigorous method for verifying the validity of scientific findings (e.g., Nosek et al., 2022). The large samples that come with conducting essentially the same experiment simultaneously in multiple labs should increase the statistical power to detect even small effects. Yet the results have continued to be quite disappointing. The present article reviews the published results of these multi-site replication attempts in social psychology, in the hope of understanding why these attempts fail so frequently.

The idea for our article arose during the 2020 Society of Personality and Social Psychology (SPSP) conference. Researchers there presented the results of a multi-site attempt to replicate the impact of mortality salience on worldview defense (Klein et al., 2021). Specifically, being reminded of one's mortality makes one assert one's cultural worldview more strongly and resist criticisms of it. This has been a major line of work derived from Terror Management Theory (Greenberg et al., 1986), and there are many published replications already of mortality salience effects. Yet the multi-site attempt failed to find significant support for it, despite the large sample. One of us has been engaged in friendly feuding with the Terror Management theorists for over three decades (e.g., Baumeister & Tice, 1990; Greenberg et al., 1990). Despite not being a supporter of their theory, his own laboratory has replicated the mortality salience effects over a dozen times, in multiple publications. Based on both his own data and on the many published studies finding similar effects, he has confidence in the reality of that causal pattern (though disputing the broader implications). Why, then, would a multi-laboratory replication attempt of what appears to be a true effect fail, even with input by the original investigators?

Our assumption is that the amount of fraud in scientific work, while obviously not zero, is quite small. We proceed on the assumption that both the original researchers and the replicators are honest, competent scientists sincerely trying to conduct good research. Therefore, we are inclined to accept both the original results and the replications, even though the modal pattern is that the original results are significant, whereas the replications are nonsignificant. Indeed, some replication articles report nonsignificant trends in the opposite direction from the original result. In that case, it is not merely that the replication finding is too weak to be significant even in a very large sample. Rather, there is no sign whatsoever of the effect. Thus, we have two sets of researchers testing the same hypothesis and coming to opposite conclusions. Again, this is what usually happens, at least in social psychology. What can one make of this

discrepancy? What are its implications for building a broad theoretical understanding of the phenomena? And for future empirical projects?

Features of High-Quality Replications

We begin with Fiedler's (2017) definition of strong research: "Strong research must rely on sufficiently large samples that allow for powerful statistical tests of precisely predicted relationships, minimizing false positives and failures to replicate" (p. 46). While emphasizing the value of replication, Fiedler notes that this does not mean something will work every time. He points out that "a negative outcome may reflect the overshadowing impact of an uncontrolled cause" (p. 57). In particular, concern over false positive findings may distract from a potentially larger problem, namely false negatives. False negative findings may impair scientific progress more than false positive findings, insofar as further research will expose false positive findings whereas false negative findings may discourage further work (Fiedler et al, 2012). Indeed, false positive findings may not only be exposed by subsequent work, but it also seems that scientists are remarkably accurate at predicting in advance which findings will replicate (see Camerer et al., 2018). Small samples as traditionally used in social psychology will cause both false positives and false negatives. The larger, aggregated samples in multi-site replications may reduce false negatives caused by small samples — but it is possible that other features of the multi-site procedure will offset that advantage. The damaging effects of false negative findings from multi-site replications is likely to be considerably worse than false negative findings from single studies because the research community may regard the null result as firm evidence that the phenomenon does not exist, thereby stifling further research and preventing self-correction. Importantly, most efforts to reduce false positive findings will increase false negative findings (Fiedler et al., 2012).

What would be best practices for conducting a multi-site replication? None of the present authors has conducted such a project, but based on our extensive reading of the literature we can list several important features — as well as several debatable ones. It deserves mention that having many laboratories and many participants is an apparent strength, but if the replication is conducted in a flawed manner, those features quickly lose their value. They may even be damaging insofar as the apparent strength of so much data lends credibility to misleading findings.

The clear strengths would include preregistration, including specifying the primary analyses — but also a willingness to conduct and report further, exploratory analyses, as long as the distinction is explicit. A large aggregate sample is desirable. Given that replications nearly always find reductions in effect size from the original, the sample should be extra-large.

Excluding large amounts of data can seriously compromise the quality of replication. Excluded data are especially problematic if they are not evenly distributed across conditions and for the same reasons (e.g., Cook & Campbell, 1979). It is important to report whether the excluded participants were evenly distributed across conditions, because otherwise the different conditions sample somewhat different populations (a problem that increases with more exclusions). If exclusion rates are high, analyses should be reported both ways (i.e., both full sample and after exclusions). As a commendable example, McCarthy et al. (2021) reported their pre-registered analyses but also, after noting the high proportion of discarded data, reported multiple analyses with different (as well as no) exclusion criteria.

Whether procedures should be precisely the same for all participants is debatable. Here we appreciate the practice of using multiple different procedures to test the same hypothesis

(e.g., Vohs et al., 2021). To be sure, when different procedures yield different results, interpretation is compromised — but it is nevertheless scientifically important to know this.

Manipulation checks are essential, especially considering the high rate of failure in multi-site projects. A manipulation check should do more than verify that the instructions were correctly perceived: It should confirm that the requisite psychological states were successfully manipulated (Fiedler et al., 2021). Ideally, and especially in cases of failed replications, some evidence that the dependent variable was sufficiently sensitive to detect group differences would also be provided. Attention checks could also be useful, especially for computer-administered studies. In case of failure accompanied by weak manipulation checks, further analyses would be appropriate to see whether any support for the hypothesis can be found among participants whose manipulation check data were strongest — with the caveat that such analyses introduce selective sampling and hence potential confounds.

It may seem that we are holding a double standard, given that we fault replications for not having manipulation checks when the original studies might also have omitted them. We have three responses to that objection. First, establishing the validity of a manipulation is a property of good research in general, and that applies equally to original studies and replications. Second, the multi-site replications are intended to raise the quality of the field's knowledge base, so improvements such as adding manipulation checks are consistent with that. Third, within the purview of our efforts to understand this literature, some double standard is perhaps unfortunately necessary. The modal case we shall report involves a significant original finding but a nonsignificant multi-site replication. The need for the original authors to validate their manipulation is thus less — because, after all, it worked. With the failed replication, in contrast, it is vital to know whether the manipulation was effective. A significant manipulation check combined with a null result on the dependent variable is evidence that the original hypothesis is wrong. But as many authors of these replications have acknowledged (e.g, Buttrick et al., 2020), if the procedures fail to manipulate the independent variable, then no conclusions about the hypothesis can be drawn.

To be sure, manipulation checks do not solve all problems. Gruijters (2021) has pointed out that many social psychology procedures manipulate more than just the focal variable, thus enabling alternative explanations. In the past, it has been largely the job of editors and reviewers to raise these issues, and authors may have been asked to collect additional data to rule out specific alternative explanations. Fully validating a manipulation would ideally involve both establishing that the focal independent variable was successfully and substantially manipulated — and that other relevant variables were not. For example, affect is often a source of alternative explanations, and a great many studies include affect measures to show that their manipulation did not create emotional side effects; hence the remarkably low rate of significant findings for emotion mediation across half a century of social psychology research (DeWall et al., 2016). Gruijters also makes the valuable point that even a significant manipulation check does not necessarily indicate that the manipulation was strong enough to produce a significant effect on the dependent variable. Manipulation checks should presumably have large effect sizes, in order that the treatment groups are meaningfully different. Indeed, the effect size of a manipulation check may be more meaningful than the effect size on the dependent variable because the former is a helpful indication of how valid the study was.

Fidelity to the original study seems central to the very idea of replication. Ellefson and Oppenheimer (2022) have argued for the importance of such fidelity. However, given that no replication can be truly exact, some departures from precisely duplicating original procedures

may be desirable, such as updating materials to reflect current social and historical factors. In case of failed replication, clear discussion of departures from the original procedure is needed. However, successful replications gain credibility with increasing discrepancies from the original procedure because these indicate greater generality and robustness.

Given that the purpose of experimentation is generalizable knowledge, exact fidelity is a mixed blessing — depending on outcome. A successful replication is more informative and persuasive to the extent that it was a purely conceptual replication with different operational definitions of key variables from the original study. In contrast, replication failure becomes increasingly ambiguous with each departure from the original study, and so it is most informative when fidelity is maximal.

A particular fidelity problem arises when the original study did not include a manipulation check, so that adding one reduces the replication's fidelity. Manipulation checks are generally a positive feature indicating good quality research, but sometimes they are impractical. Nevertheless, given the high failure rate of multi-site replications, manipulation checks are crucial for establishing whether the experiment provided a fair test of the hypothesis. If the original study worked as predicted, and if reviewers did not find potential confounds that suggested alternative explanations, it seems reasonable to assume, at least provisionally, that the manipulation did what the original authors proposed. When a replication fails, however, it is immensely valuable to have clear evidence as to whether the manipulation was effective.

It is important to realize that many (original) social psychology publications have fallen short of best practices (especially as we now understand best practices). Many classic social psychology studies had no manipulation checks, used small samples, discarded participants without reporting full-sample analyses, and the like. More broadly, all social science research is imperfect. Replication failures may indicate problems with original studies, replication studies, or both.

Possible Explanations for Failed Replications

Experiments typically test causal hypotheses of the form "*X* causes *Y*," with *X* being the independent variable and *Y* being the dependent variable. As we shall show, the modal pattern in social psychology has been for original research reports to indicate significant support for the causal relationship but for the multi-site replication to fail to find any such support. Because the two findings contradict each other, and a causal relationship cannot simultaneously exist and not exist, the challenge is to understand what could account for the difference. Fiedler (2017) also points out that such so-called "minimal hypotheses," such as that *X* causes an increase or decrease in *Y*, are insufficient for scientific theory building. It is therefore useful to explore why *X* may cause *Y* sometimes but not other times.

A first possible explanation would be that there are boundary conditions: the hypothesized causation occurs under some conditions and not others. Relatively few hypotheses are put forward with a theoretical elucidation of when this effect will versus will not be observed. In this case, both the original and the replication finding may be correct evidence about reality. The two projects merely lie on opposite sides of some crucial boundary condition. The challenge for advancing theory is to delineate what the boundary condition(s) may be, and for subsequent work to show that in the same study the finding can be replicated and not replicated by manipulating that boundary condition. We concur that future theorizing in psychology would benefit from articulating likely boundary conditions (Fiedler, 2017. Without theoretical specification, it is mostly not possible to evaluate whether boundary conditions

account for replication failures in the present sample of multi-lab replications. At most, they can be post hoc speculations.

A second possibility would be that the original finding was wrong, and the replication correctly finds that *X* does not cause *Y*. In the present sample of studies, the best evidence for this would be that the replication provides significant and substantial evidence of a successful manipulation check, but the predicted differences on the dependent variable were not significant. (They could be significant in the opposite direction, which would also speak very strongly against the original hypothesis.) Further evidence that social psychology's failed replications show that original findings were spurious would be that findings that have been obtained only once or a few times would fare worse (in our sample) than those which have been replicated many times.

A third possibility would be that the replication failed to provide a valid test of the hypothesis. We call this *operational failure*. One sign of this would be that the manipulation check is not significant. Indeed, for a strong test of a hypothesis, the manipulation check would have to be not only significant but also with a large effect size, indicating that the treatment groups were substantially different on the independent variable. Manipulation checks should ideally verify not simply that the manipulation was perceived but that it produced the intended difference in psychological states (Fiedler et al., 2021). A priming manipulation, for example, could be checked simply by verifying that the participant unscrambled the assigned words to make coherent phrases using the target word — but could also be checked by showing that it made the concept mentally accessible.

Other signs of operational failures would include high rates of discarding data. Social psychologists have long accepted that it is fair to discard some data points for valid reasons, such as from the occasional participant who misunderstood the instructions or was suspicious. But journals traditionally would not publish studies from which substantial numbers were excluded. Differential rates of discarding in different conditions produces selective attrition (Cook & Campbell, 1979), making it difficult to know whether observed differences were due to the manipulated independent variable (as hypothesized) or were confounded by differential sampling (e.g., if only one condition deletes inattentive participants at a high rate). Differential attrition may be particularly common with online procedures (Zhou & Fishbach, 2016). High rates of attrition may also signal that participants were not fully engaged with the experiment and instead were simply trying to get the study over with as quickly as possible. Older social psychologists recall training in how to ensure each participant was motivationally and emotionally engaged in the procedure (called *experimental realism*; Berkowitz & Donnerstein, 1982).

Exactly why engagement would be lower among participants in a multi-site replication than in the original study is not entirely clear. Research on social loafing has suggested that people are generally less engaged when performing anonymously as part of a large group than when individually identified (e.g., Karau & Williams, 1993) — but participants were often told explicitly that their identities would be submerged in a large group. Most consent forms explicitly state that data will be analyzed at the group level (i.e., the data from all individual participants within a group will be combined). It is also plausible that the streamlining of procedures for mass administration reduces the engagement, such as by minimizing social contact between the participant and the experimenter.

Low participant engagement may be a serious problem for many studies, including replications, particularly in this era of online data collection. The global pandemic lockdown alerted many instructors to the fact that teaching is more effective in person than via computer-

mediated communication such as Zoom (e.g., Alpert et al., 2016; Halloran et al., 2021; Kofoed et al., 2021; Riegg Cellini & Grueso, 2021). Overall, online learning has been shown to be less engaging (e.g., Kop, Fournier, & Mak, 2011) and lead to less learning (e.g., Heppen, Sorensen, Allensworth, Walters, Rickles, Stachel & Michelman, 2017) than traditional in person education. Even in a best-case scenario such as The Netherlands (described by its authors as best case because it had a short lockdown, equitable school funding, and excellent rates of high quality internet access), a study of around 350,000 students found that computer-mediated distance learning led to lower academic outcomes than traditional education (Engzell, Frey, and Verhagen, 2021). Although He et al. (2021) found that online learning was comparable to in person learning for health science students, this was only the case for synchronous learning, whereas most if not all online psychology experiments are asynchronous. A more suitable analogy may be the switch to asynchronous, computer-mediated learning during the COVID pandemic. A recent Brookings study found that students' learning and achievement levels dropped between a tenth and a quarter of a standard deviation between 2019 and 2021 (Kuhfeld et al., 2022). Lower levels of engagement and performance in asynchronous online learning may be analogous to lower quality data due to less engagement in asynchronous online psychology experiments. Hence, we coded studies for whether live personal interaction was involved.

Possible Sources of Bias in Present Authors

Because discussions of sensitive issues may be clouded by bias, we wish to lay out the broad assumptions and goals for our approach. Each of the present authors has long cultivated a positive attitude about social psychology in general. Our positive outlook — tempered, to be sure, with judicious skepticism about specific issues — is clear in a textbook coauthored by two of us (Baumeister & Bushman, 2021). It is therefore fair to suspect our analysis of bias in favor of the existing literature, and indeed we think our admiration for the general body of work by social psychologists over the past half century fueled our curiosity about why the multi-site replications fail so often. We think readers seeking a perspective with the opposite bias (i.e., preferring to think that most social psychology research has yielded no useful knowledge) will have no difficulty finding sources to make that case.

As one highly relevant example, the term *p*-hacking has come into use recently to disparage conducting multiple analyses and then reporting the best results. The term encompasses some practices that were never acceptable, such as analyzing frequently during the data collection process and stopping the collection whenever the .05 significance line was successfully crossed. It also encompasses practices that were once considered acceptable, such as if the initial analysis yielded a .06 *p*-level and so the researcher tried using gender as covariate or transforming the data, which sometimes nudged the finding across the .05 line. Skeptics have taken that to propose that the literature is rife with false positive findings, and that view deserves to be respected. A more charitable view, however, is that with small samples, many true effects would reach only marginal levels of significance due to low statistical power, and such formerly acceptable *p*-hacking would sometimes enable the finding to reach significance and get published — thus essentially rescuing a true effect from a false negative result (Type II Error). The two perspectives differ as to how easily or frequently a bogus significant finding (Type I Error) could be manufactured with a truly nonexistent effect. Our view is that such occurrences would be rare and would hardly ever replicate, so when findings have been obtained repeatedly, we suspect that there is a genuine phenomenon there — even though the evidence for it was inflated.

We agree with the skeptics, and with the great bulk of the replication literature, indicating that effect sizes in original reports of lab findings are inflated, and even the formerly acceptable forms of *p*-hacking would likely contribute to this. Nevertheless, in contrast to the doom-and-gloom tone of much writing about replication problems, we seek to explore the perspective that many of the phenomena are genuine, even if the original reports included inflated effect sizes.

Two of us were also heavy contributors to the ego depletion literature, which has a mixed record in multi-site replications, thereby rendering our own possible bias as something potentially substantial yet hard to specify. We do have confidence in our own work, which we always sought to conduct carefully and consistently with prevailing best practices (which, admittedly, have changed substantially over the decades). An early multi-site replication reported a failure to find evidence of ego depletion, though subsequent reanalyses did indicate support. Crucially, however, a recent multi-site replication found strong and unequivocal support for ego depletion, thus vindicating our work (Dang et al., 2021). Indeed, if one accepts all the social psychological multi-site findings uncritically, ego depletion is one of only four clear successes we uncovered. Indeed, the rarity of success in social psychology multi-site revisions enabled us to make the case elsewhere that ego depletion is at present social psychology's best replicated finding (Baumeister & Tice, 2022). Whether we would therefore be biased in favor of or against the multi-site approach is therefore unclear. At minimum, we would be likely to notice factors that have contributed to weak replication results in general. Our conscious goal is to advance understanding of the field, and, by definition, we cannot know what our unconscious biases are.

Our attempt to understand the pattern of widespread failure to replicate is also hampered by the literature itself. No one would suggest that the multi-site replications are a representative sample of all the work done in social psychology. Indeed, around a third of the multi-site replications have focused on priming, which is one important phenomenon in social psychology — but just one of many. Meanwhile, many hypotheses and studies have not been tested in multi-site replications.

METHOD

Literature Search

We searched the data base PsycINFO through 2022 (July 28) using the terms "*multi-site*" OR *multisite* OR "*multi-lab*" OR *multilab* OR *replicat** OR "*many labs*." The wildcard * character retrieves terms with letter(s) after * (e.g., *replicate*, *replicates*, *replicated*, *replication*). The search retrieved 7,361 articles. These articles were then searched to identify multi-site replications of social psychology experiments (broadly defined, e.g., including applied areas such as law, environment, health, consumer behavior, industrial-organizational behavior). The search excluded dissertations, articles not written in English, and articles that were not peer-reviewed. Methodology was restricted to "experimental replication." Based on a reviewer suggestion, we used the same terms in the journal *Collabra: Psychology* and found an additional 54 articles. Of these, one was relevant. Our search resulted in 36 relevant replications, which are marked with an asterisk (*) in the Reference section. Some articles seemed at best borderline social psychology, as we shall note. All multi-site replication studies included in our analysis were registered in advance.

Three additional articles are described in a narrative fashion (Klein et al., 2014, 2018; Schweinsberg et al., 2016[1]). These reported multi-site replications of multiple different studies

[1] We thank an anonymous reviewer for bringing the Schweinsberg et al., (2016) article to our attention. Investigation revealed that our literature search did not retrieve this study because it

rather than a single study. These indeed qualified as multi-site replications, but the experiments are perhaps best described as mini-studies, given that they were chosen for brevity, rapidity, and convenience. Importantly, these mini-studies are not statistically independent because the same participants completed several mini-studies. The assumption of independence is a foundation for many statistical tests, including the ones we report in this paper.

Coded Variables

For each multi-site replication, we coded: (1) publication year of multi-site replication study, (2) publication year of original study, (3) whether an author on the original study was also an author on the multi-site replication study, (4) the number of authors on the multi-site replication study, (5) the number of labs contributing data in the multi-site replication study, (6) the total sample size in in the multi-site replication study, (7) the sample size analyzed in the multi-site replication study, (8) the total sample size in the original study, (9) the sample size analyzed in the original study, (10) whether the multi-site replication study used exactly the same procedure as the original study, (11) social interaction in the multi-site replication study (live ongoing interactions were coded as 2; computer mediated, computer mediated simulated, and observing people interacting were coded 1; solitary responding with or without computer, completing tasks alone in classrooms with other students, computer administered hypothetical vignettes were coded 0), (12) social interaction in the original study (coded the same as social interaction in the multi-site replication study), (13) manipulation check in multi-site replication study (*significant*, *nonsignificant*, *mixed=significant for some measures but nonsignificant for other measures*, or *none*), and (14) manipulation check in original study (*significant*, *nonsignificant*, *mixed*, or *none*).

We were going to code how many times the original study been replicated before and by whom, but that was difficult to determine in a precise manner. We were also going to code whether a journal was a flagship journal as a rough indicator of journal quality, but we decided not to code this variable because some journals were difficult to classify.

We considered a multi-site replication to be successful if it yielded significant results on the main dependent measure(s) in the same direction as the original study, thereby upholding the original finding (coded 2). We coded the results as mixed/partial success if multiple analyses are reported, some of which indicate significant confirmation of the original (primary) finding while other analyses (e.g., using different protocols) on the same dependent variable were not significant (coded 1). (Note, these analyses had to be across the multiple labs; a significant result in one lab was not sufficient to code as partial success.) There were some borderline cases in which most analyses failed but a lone result emerged from an analysis that lacked credibility. Our codings are presented in Table S1 in the Supplemental Materials, and anyone who wishes to re-code and reanalyze is welcome to do so. We considered a multi-site replication to be unsuccessful if it yielded nonsignificant results on all measures and/or results in the opposite direction as the original (coded 0).

Each variable was coded by two independent coders. Percent agreement ranged from 94% to 100% (M = 98.5%; Median = 100%). Disagreements were resolved via discussion among the three authors.

RESULTS

was not classified as an "experimental replication." The purpose of our article was to examine multi-site replications of social psychology experiments.

We found 36 articles reporting multi-site replications in social psychology. We considered four of them (11%) to be largely successful (despite smaller effect sizes than the original studies) and five additional ones (14%) to be partial or mixed successes[2]. Counting the mixed successes with the successes (a stretch) still leaves 75% as failures. We emphasize that generalizing to all social psychology is unwarranted, because the sample of effects chosen for replication is not at all a randomly created, representative sample of the field's knowledge as a whole. (In particular, field studies, labor-intensive studies, and studies with special populations seem underrepresented among replications.) Nevertheless, the 75% failure rate is the most generous description of the entire population of published multi-site social psychology replications we could locate as of June 2022.

As already noted, we also report separately on three other articles (Klein et al., 2014, 2018; Schweinsberg et al., 2016), each of which reported multiple different replications of different studies. The studies they sought to replicate were not all in social psychology and involved brief, one-shot measures, typically of social judgment and/or hypothetical reactions to imaginary situations, with large multi-lab samples. Thus, Klein et al. (2014) had each participant at each session perform 13 different brief experiments, some of which involved just a few seconds to answer a single question. Klein et al. (2018) replicated 28 different brief experiments; each protocol was administered to approximately half of 125 samples that comprised 15,305 participants from 36 different countries and territories. Schweinsberg et al. (2016) had 25 research groups replicate between three and 10 unpublished findings regarding moral judgments of hypothetical or imaginary events. We decided not to cover these on the same level as the studies devoted to replicating a single finding for multiple reasons, including the necessarily briefer presentation of results and resultant lack of information about manipulation checks, number of participants discarded from analyses, statistical non-independence of results, and the like. Nevertheless, they do qualify as multi-site replications and we describe them separately.

Effect Sizes

Almost all multi-site replications of social psychology effects find results with considerably smaller effect sizes than the original study. This is obviously true for studies reporting nonsignificant results, but it is true even for the few that report positive, significant results in line with the original finding (see section on successful replications, below). It is highly likely that effect sizes reported throughout the social psychology literature are frequently inflated due to publication bias and *p*-hacking. The shrinking of effect sizes was also noted in medical research (Hunniford et al., in press), who attributed it to superior methodological rigor, including larger samples, in multi-site replications than in single-site original studies. Fiedler and Prager (2018) proposed that regression toward the mean (i.e., the statistical tendency for extreme observations or results to be followed by others closer to the mean) almost guarantees that replications (whether multi-lab or single-lab) will produce smaller effects than original studies. We have also proposed that smaller effects would be consistent with lesser engagement of participants.

[2] There are always borderline cases. The five mixed ones reported significant and supportive findings in some analyses but not in others. Two additional ones (Bouwmeester et al., 2017; McCarthy et al. 2018b reported at least one significant finding consistent with the original result but were still coded as failures. Bouwmeester noted that analysis lacked legitimacy because of extensive and differential attriation. McCarthy et al. had two measures of their dependent variable, both of which produced but tiny significant effects but in opposite directions.

The very notion of a true effect size for a social psychology laboratory manipulation of a situational variable, such as interpersonal rejection, may be elusive if not meaningless. In some fields, it may be possible to establish true effect sizes. Insofar as social psychologists wish to accept true effect sizes, the pervasive shrinkage during replications poses a challenge. Original researchers may often have been both good and lucky (because unlucky ones failed to publish), so they report effect sizes that will inevitably shrink in replications (Fiedler & Prager, 2018). Fortunately, numerical effect sizes are almost never stipulated in psychological theorizing, so the inability to establish correct estimates may not be a major problem for advancing theory.

Nevertheless, it is also entirely possible that multi-site replications may be biased toward underestimating effects. Presumably the reasons for that would overlap heavily with the reasons for the broad tendency for multi-site replications to fail, such as streamlining procedures for efficiency, low engagement of participants, and procedures that were better suited to the original than the replication sample.

Manipulation Checks and Operation Failure

Still, the findings may not be as dismal as first seemed. Manipulation checks were missing from half (50%) of replications, as well as nearly all of the mini-studies. In these cases, it is impossible to interpret whether the null findings indicate falsification of the hypothesis or mere operational failure. The hypothesis was not tested if the experimental treatments failed to create the intended difference on the independent variable. It was also not tested if the dependent measure was insensitive.

It is also important to note that there are two kinds of manipulation checks. One verifies that the manipulation was correctly perceived, enacted, and administered. The other verifies that the resulting psychological states differed as intended. As a commendable example, Ghelfi et al. (2020) administered beverages of different tastes to manipulate disgust. Their manipulation checks verified both that the bitter beverage tasted more bitter than the others *and* that it caused more disgust. But very few replications have been that thorough.

True Failures to Replicate

This and the next few sections will summarize the different types of findings. Readers who wish to read the narrative descriptions of all 36 multi-site replications can access them in Supplemental Materials.

We start with the bad news. Here we summarize the nine true failures to replicate, indicating a falsification of the hypothesis (see Table S2 in Supplemental Materials for full list). This requires a significant and presumably sizeable difference on the manipulation check but a null (nonsignificant) finding on the dependent variable. (Ideally, one should also show that the dependent variable was sensitive, such as by providing a significant difference as a result of another manipulation, even of a different variable.) As one example, Williams and Bargh (2008) found that briefly holding hotpack in one's hand caused people to select a reward for a friend rather for themselves, as compared to briefly holding a coldpack. Lynott et al. (2014) attempted a replication in three sites. The manipulation check was large and significant in that the hotpack was rated as much warmer than the coldpack. Verifying that the notion of hot/warm or cold was unconsciously activated in participants' minds is difficult, but given that participants did rate it correctly from memory suggests it was. None of the three sites replicated the effect, and indeed one site found a significant effect in the opposite direction. The combined results approached significance, but in the direction opposite to the original finding.

In this category, we also included several multi-site replications that did not report a manipulation check as such, but for which it seemed fair to assume that the independent variable

was successfully manipulated. As an example, Dijksterhuis and van Knippenberg (1998) reported that priming participants with the concept of 'professor' caused them to perform better on a trivia test than priming participants with the concept of 'soccer hooligan'. O'Donnell et al. (2018) reported 23 direct replications in a multi-site project. They found no evidence of improved performance based on priming with "professor", nor of moderation by gender. There was no manipulation check, but given that participants had to write a paragraph imagining their life as either professor or soccer hooligan, it seems reasonable to assume that those different roles were activated in their thinking.

Operational Failures

Operational failures refer to replication attempts in which the independent variable was not effectively manipulated. They therefore do not constitute falsifications of the hypothesis, because they were unable to provide a test of it. Nevertheless, they do raise other concerns, such as the generalizability of the methods. Table S3 in Supplemental Materials lists the six multi-site replications in this category.

For example, inducing people to disbelieve in free will caused them to become more prone to antisocial behavior (cheating on a test to claim extra reward money), as shown by Vohs and Schooler (2008). An early attempt by Embley et al. (2015) at replication failed, and one possibility was that the manipulation (reading a difficult passage from a science book) was ineffective because it was hard to understand. (The original study may have sampled a highly intelligent and conscientious population.) Buttrick et al. (2020) sought to replicate the original effect in five sites, using both the original manipulation and a revised one that supposedly would be easier to understand. The effect on cheating was in the predicted direction but weak and nonsignificant. However, it appeared that neither the original nor the revised manipulation was able to alter people's beliefs in free will. Thus, the study was unable to test the hypothesis. A follow-up analysis found the correlation between the manipulation check and the dependent variable was not significant, unlike several other cases. Thus, no support for the original hypothesis could be salvaged. Nevertheless, as Buttrick et al. themselves point out, this investigation permits no conclusions about the hypothesis that disbelief in free will causes cheating, because the manipulation failed.

Ambiguous Failures to Replicate

We designate as ambiguous those failures for which it is unclear whether operational failure or hypothesis falsification occurred. Half of the replications did not collect or report manipulation check data, so it is impossible to ascertain whether the failures among them indicate falsification of the hypothesis, failure to test it properly because of operational failure, or both. Table S4 in Supplemental Materials lists these seven replications.

As an example, Payne et al. (2008) found that priming people with the importance of responding accurately and honestly led to higher correlations between implicit and explicit attitudes, as compared to priming them with the awareness that everyone is biased and it is necessary to be on guard. A replication in Italy found a significant effect also, though much reduced in effect size (Vianello, 2015). Ebersole et al. (2020) did a direct replication in four US and four Italian laboratories. The American labs found a significant difference in the opposite direction, whereas the Italian labs found no difference. There were no manipulation checks. It could have been an operational failure, as authors note, but the failure could possibly reflect historical change, given that race relations in America deteriorated between the time the original study was conducted (2008) and the time the replications were conducted (2016/2017).

Mixed Successes

We turn now to the minority of multi-site replications that provide at least some support for the original finding. Table S5 in Supplemental Materials lists the five partly successful replications. Establishing precise boundaries for this category was more difficult than for any others because many studies reported many analyses. We defined this category as studies providing significant support for the original finding with some analyses but falling short of significance with other analyses. These analyses had to be across the full data set (thus, a significant finding from one of the many laboratories was not enough), and the successful one had to be sufficient to enable some disinterested scholars to regard it as a successful replication.

Classification of the mixed successes can be explained as follows. For Ghelfi et al. (2020), the most important comparison was between the bitter and neutral beverage conditions, and that comparison was significant, although the sweet beverage condition's results were radically different from the original and complicated the theoretical point. Findings by Moran et al. (2021) and Vohs et al. (2021) were significant or not depending on how many participants were excluded from analyses. Skorb et al. (2020) tried three different protocols, of which only one provided a significant result, and that was after excluding over half the sample. Baranski et al. (2020) replicated a finding about imaginary perpetrators but not the complementary finding for imaginary victims — thus contradicting the overarching theoretical point despite replicating one of the findings.

In classifying these as mixed or partial successes, emphasis must be on 'mixed or partial.' In general, the preregistered analyses yielded the poorest results. Authors of original articles may take some comfort and justification from the fact that significant support was found at all. Critics and skeptics may claim that that is akin to grasping at straws and that the results mainly point toward failure.

True/Full Successes

Only four multi-site replications have provided unqualified support for the original hypothesis. These are listed in Table S6 in Supplemental Materials. The biggest success we found was a study on eyewitness identification. Garry et al. (2008) had two participants think they were watching the same videotaped crime. In fact, they saw different versions with different details. They discussed the film after watching it, and their different recollections influenced each other, so that participants ended up claiming to recall details they never actually observed. Earlier studies had often found similar effects, with both similar and different methods. Ito et al. (2019) conducted a replication in 10 different countries and replicated the effect fully. Most unusually, their effect size was comparable to (even slightly larger than) the original effect size. Although this study was framed and published as applied cognitive psychology, it does have a social interaction as a key independent variable and does qualify as applied social psychology.

Three Multi-Lab, Multi-Mini-Study Articles

Two articles by Klein et al. (2014, 2018) each reported multi-lab replications of several different effects. These were typically very brief studies, administered to participants one after another, almost exclusively asking how the participant thought about something, with neither social interaction nor emotion. These fared better than the other studies. Many are more properly characterized as cognitive psychology studies, marketing studies, or as judgment and decision-making studies than as social psychology studies, and replications in those other fields generally fare better than in social psychology. In general, manipulation checks were not reported, so all failures qualify as ambiguous as to whether hypothesis falsification or as operational failure.

The successful replications from Klein et al. (2018) are as follows. An (imaginary) man who accidentally hurts a baby is blamed more than a baby who accidentally hurts a grown man.

A gift giver is rated as more generous despite giving a cheaper gift if the gift was at the high end of the range of possible prices for that particular product (as compared to a more expensive gift but at the low end of the range for that product). Students think someone who has not done the assigned reading is more likely to be called on by the instructor, as compared to someone who has done the reading. When people make a binary (hypothetical) choice, they overestimate how many others would make the same choice (i.e., the false consensus effect, replicated twice). In the trolley problem, people approve of someone who steers the train to kill one person instead of five, more than they approve of someone who pushes a fat man off a bridge to block the train (again killing one but saving five). An imagined executive who expresses indifference to how his policy will affect the environment is rated as intentionally harming the environment but not as intentionally benefiting it. Correspondence bias (a.k.a. fundamental attribution error) was replicated: People tended to think an essay writer believed what he advocated in the essay, even if they were told the writer had been assigned to write on that side of an issue. The last may be partly confounded, given that the essay argued its point rather than signaling detachment, and there was some evidence that the more strongly the essay argued its point, the more participants assumed the writer believed what he said.

As to the failures: Klein et al. (2018) failed to show a sequence effect, in which rating relationship satisfaction first and then life satisfaction altered the second rating based on the first, whereas rating life satisfaction first had no effect on subsequent rating of relationship satisfaction. (This was significant in the opposite direction from the original.) The finding that imaginary jurors awarded custody more to the parent with extreme good and bad traits, rather than the one with neutral traits, was significant in opposite direction. The finding that hand-copying a description of an immoral action made people rate hand cleansers more favorably was not replicated. Likewise, they failed to replicate the finding that people formed higher opinions of a leader based on an organizational chart that had a longer line between the leader and his team (compared to a shorter line). Another finding that was significant in the opposite direction was that the choice between a kiss from a favorite movie star and a $50 cash payment depended on whether the outcome was guaranteed or merely a slight chance. The attempt to replicate an anchoring effect (from the number on a product designation to an estimate of what percent of sales would occur within the USA) found a null result.

Three priming studies were included in Klein et al. (2018). Two were not significant. Priming orderliness had originally made people express more eagerness to pursue personal goals, but this failed to replicate. Priming heat had originally led to greater belief in and concern about global warming, but this also failed to replicate. On a more positive note, priming the consumer mindset — by referring to people as consumers rather than as individuals — caused participants to predict that such people would generally fail to conserve water. This finding is notable given that it is the only one of many attempted multi-lab replications of priming that yielded a significant result. Unfortunately, there is possibly a confound here, given that the word "consumer" suggests the person may consume water (rather than conserve it), an implication that the word "individual" does not evoke.

An earlier project likewise selected studies for brevity, simple (two-cell) design, and ease of administration, so that multiple studies could be administered in the same session (Klein et al., 2014). Most of these were judgment and decision making and wording/framing studies, but some could be regarded as social psychology studies (and that distinction is fuzzy). In particular, whereas anchoring and adjustment failed to replicate in Klein et al. (2018), the 2014 paper reported four successful replications. This effect is often covered in social psychology textbooks,

including in the present authors' textbook (Baumeister & Bushman, 2021) — but it is essentially a common mistake people make when estimating a number, so there is nothing very social about it. In terms of more purely social psychology: Imagining contact with Muslims reduced anti-Muslim prejudice. Another was that an ambiguous quotation was rated more favorably when attributed to a liked than a disliked historical figure. Two priming studies (flag and currency) failed to replicate.

A methodologically innovative project by Schweinsberg et al. (2016) obtained multi-site replications of 10 findings that one laboratory had accumulated but not yet published. These were again mini-studies, such as asking whether a corporate executive or a vandal was more likely to burn in hell, or whether a manager who is mean to all subordinates is worse than a manager who is only mean to subordinates who belong to racial minorities. This project differed from most other multi-site replications not only in: (1) replicating unpublished findings, (2) using quick mini-studies, (3) including the original authors, and (4) including replicating original studies that had failed to yield significant support, but also in: (5) selecting the replicating labs by invitation only, indeed selected/invited by the original authors based on the expectation that the replicating labs would have similar subject populations and would be likely to get similar effects without needing to modify the procedures (other than translating into the local language). The authors note that this last feature removed the adversarial element that other replication projects have. All the studies involved moral judgments associated with the same general theory (a person-centered account of moral judgment).

Schweinsberg et al.'s (2016) findings stand out as highly successful replications, and so the methods they adopted may offer valuable guides to future attempts. Eight of the 10 findings were successfully replicated, and a ninth confirmed the original hypothesis much better than the original (nonsignificant) finding. (Note, a significant replication combined with a nonsignificant original finding is precisely the opposite of the modal pattern we have found.) The tenth was the main failure, testing the hypothesis that executives of charitable organizations are held to higher moral standards than for-profit corporate executives.

Discarding Data

Disturbingly high rates of data exclusion were found in multiple replications. The proportion of data discarded from the main analyses in the replication studies ranged from 0% to 39% (M = 10.5%, Median = 5.6%). In terms of raw numbers, the studies discarded an average of 318 participants (range 0 to 2,484 participants). Supplementary analyses sometimes excluded even more. In comparison, the proportion of data discarded from original studies ranged from 0% to 13% (M = 1.1%, Median = 0%).

Moreover, in the multi-site replications, the excluded data were often quite unequally distributed across conditions, which raises questions about whether non-excluded data are unbiased. Usually this was the result of pre-registered exclusion criteria — but ones that were presumably specified with the expectation that only a few outliers or inattentive participants would be discarded, not a sizeable fraction of the sample. We review these here, noting also whether results were different between the full and truncated samples:

Baranski et al. (2020) noted that nearly half (45%) their sample failed the attention check. They reported that the focal interaction was significant with the full sample but not the truncated sample, although the patterns were similar and both departed from the original finding.

The Bouwmeester et al. (2017) project required participants to respond either before or after 10 seconds, and many failed to comply. Two-thirds of participants in the fast-response condition failed to respond by the deadline (as compared with only 7.5% in the slow-response

condition who responded early). Differential attrition is of course a serious threat to internal validity (e.g., Cook & Campbell, 1979). Excluding them yielded a significant result supporting the original finding, while retaining them yielded a nonsignificant result in the opposite direction. Other exclusions (e.g., for failing to understand, or having previously participated in similar studies) produced yet different results. In particular, excluding non-comprehending participants yielded a significant result. Non-comprehension itself suggests inadequate engagement: A participant who does not understand what is going on cannot usually provide data relevant to a fair test of the hypothesis. The analysis using all three of their preregistered exclusion criteria reduced their sample from 3,596 to 792 participants – and yielded their most significant result. But excluding 78% of the sample is highly worrisome.

Dang et al. (2021) presented analyses based on both full sample and after excluding a sizable number (19%) who appeared to respond randomly. Ego depletion was significantly supported in both analyses. Random responding certainly indicates low engagement. Excluding all those participants increased the effect size by 60%. In that sense, the exclusion criteria do appear to have had some valid basis (and strengthened the result) but did not change the positive conclusion.

The O'Donnell et al. (2018) project excluded at multiple levels: 40 laboratories participated in the project, but nearly half (17 labs) of these were discarded entirely, mainly for not recruiting enough male participants. Additional participants were excluded laboratory by laboratory. In some analyses, over half the remaining participants were excluded for apparent suspicion. Such analyses, retaining only a small proportion of the total, did find a significant effect consistent with the initial hypothesis. Including data from all 40 laboratories yielded a nonsignificant effect. Thus, in this study, the exclusions increased successful replication.

McCarthy et al. (2018a) recruited 1,246 but reported analyses on 1,069, thus excluding 14.2%. They did not report analyses on the full sample, possibly because most of the excluded data were not usable.

McCarthy et al. (2018b) excluded four of 26 laboratories entirely (for failing to recruit at least 100 participants in each condition — which is thus not a sign of low engagement among participants) and additional ones from each laboratory, dropping their initial full sample from 7,343 to 5,610 participants, thus excluding 23.6% of their data. Restoring the four additional laboratories yielded almost identical results.

McCarthy et al. (2021) ran both a close replication and a conceptual replication protocol. Their pre-registered criteria dictated deleting 34% and 35.6% of participants, respectively, and they excluded significantly more in the hostile priming condition than in the neutral control condition, which threatens the internal validity of their findings. They reported multiple analyses restoring some or all of the excluded data, and the main finding was nonsignificant in all analyses.

The study of unconscious conditioning by Moran et al. (2021) had to exclude participants who were aware of the manipulation, which was assessed in various ways for different analyses. The most rigorous exclusion eliminated 49% of the sample. The analysis with the largest *N* eliminated only 9%, and it was the only one that yielded a significant result. Thus, again, more exclusion failed to improve the findings and even seems to have weakened them.

Panero et al. (2016) recruited 1,302 participants and excluded 510, or 39% of the sample, for various reasons. They report only analyses based on the reduced final sample.

Sanchez et al. (2017) deleted 7.9% of their sample for various reasons. They reported that results remain identical in alternate analyses retaining them all.

Skorb et al. (2020) reported both high and uneven rates of excluding participants. In their three protocols, 57%, 64%, and 0% were excluded. Their sample thus dropped from the initial 1,018 to 663 participants. They reported that results maintained the same patterns in alternate analyses of the full sample. Inspection of their online tables indicates, however, that the effects dropped out of the significant range. Thus, in this case, excluding suspicious and noncompliant participants improved the results slightly — including making one significant result possible.

Verschuere et al. (2018) tested 7,158 participants and for their primary analyses included only 4,674 participants, thus discarding 35% of the sample. They do not report a full sample analysis but restoring some participants yielded no improvement.

Vohs et al. (2021) discarded 1,068 participants (30% of their sample). The replication was significant with the full sample but not after deleting the 1,068 participants. This high rate contrasts with a pre-registered single-lab replication of the same phenomenon (ego depletion), which discarded 9% of the sample (Garrison et al., 2019). Schmeichel[3] was an author on both studies and it is reasonable to assume that the criteria were similar. Discarding 9% is worrisome, but hardly as bad as 30%. The high rate is consistent with the low-engagement hypothesis. And, again, discarding participants weakened the results.

To summarize: excluding a substantial number of participants sometimes improved results (Bouwmeester et al., 2017; Dang et al., 2021; O'Donnell et al., 2018), sometimes made them substantially worse (Baranski et al., 2020; Moran et al., 2021; Vohs et al., 2021), and sometimes made no difference (McCarthy et al. 2018b; McCarthy et al., 2021; Sanchez et al., 2017; Skorb et al., 2020; Verschuere et al., 2018). Apart from Dang et al. (2021), whose results were significant in both analyses, and McCarthy et al. (2021), whose results were nonsignificant either way, many of the exclusions were decisive in terms of whether the replication's omnibus finding was significant or not. The purpose of excluding participants is presumably to improve the test of the hypothesis and, if the hypothesis is correct, to increase the chances of a significant result. In these studies, this usually fails or even backfires.

Analysis of Coded Variables

The sample of 36 articles is small and not representative, so any statistical work with them should be considered exploratory and descriptive (of the full population of published multi-site replications in social psychology). Nevertheless, we conducted a series of such analyses. These analyses and predictions were preregistered:
https://osf.io/t7u3x/?view_only=3596c6bbaa2447a0b3f65b59617d9c0f

Significance tests for manipulation checks and replication effects were treated as rank order data (0 = *nonsignificant*, 1 = *mixed*, 2 = *significant*) and were therefore analyzed using nonparametric procedures. For social interaction, live ongoing interactions were coded as 2, and computer mediated, computer mediated simulated, and observing people interacting were coded

[3] Schmeichel (personal communication) noted an intriguing problem with preregistered exclusion criteria. The Garrison et al. (2019) paper had preregistered to exclude any participant whose error rate was more than three standard deviations from the mean, as is typical. However, ego depletion increases errors, and a chi-square test revealed that there were indeed significantly more participants from the ego depletion than the control condition who were designated for exclusion on that basis. Dropping them from Study 2 dropped the overall finding below the .05 significance threshold. In retrospect, those participants should be considered relevant to the hypothesis.

1. All others (e.g., solitary responding with or without computer, completing tasks alone in classrooms with other students, computer administered hypothetical vignettes) were coded 0.

A meta-analysis by Richard et al. (2003) found an average correlation of .21 from over 25,000 social psychology studies involving over 8 million participants conducted over the past 100 years). Therefore, we used $r_s = .21$ as a benchmark for a correlation to be considered meaningful in our analysis, regardless of whether it was statistically significant at the .05 significance level.

As predicted, multi-lab studies with successful manipulation checks were more successful at replicating original results than ones with failed checks. Although the Spearman rank-order correlation was nonsignificant (given low statistical power), it was quite large, r_s (N = 18) = .377, p = .123. Similar results were obtained when manipulation checks coded 0 and 1 were combined and compared to manipulation checks coded 2, r_s (N=18) = .381, p = .118. Replications without manipulation checks (N = 18), a full 50% of replications, were excluded from these analyses. We also coded whether original studies included manipulation checks. One-third of original studies did, and all these studies found significant effects for their manipulation checks.

Discarding more data did not increase the success rate. This is surprising, because the purpose of discarding data is presumably to strengthen the test of the hypothesis. If anything, the trends indicated that higher numbers of discarded participants and higher rates of discarding both correlated with *lower* success at replication (though not significantly). This is consistent with what we reported in the previous section on discarding data, namely that discarding has not generally improved replication outcomes and often seems to backfire.

We also conducted some exploratory analyses. The most remarkable finding concerned whether the procedure included actual social interaction. This correlated positively with replication success at r_s (N = 36) = .701, p < .001. Similar results were obtained when social interactions coded 1 and 2 were combined and compared to no social interaction at all, r_s (N=36) = .695, p < .001. Thus, the important point appears to be whether the study included any social interaction. Overall, 80.6% of replications contained no social interaction. None of the replication failures featured live social interaction.

In addition, we coded whether original studies featured social interaction, which was also highly correlated with whether a replication was successful, r_s (N = 36) = .557, p < .001. Similar results were obtained when social interactions coded 1 and 2 were combined and compared to no social interaction at all, r_s (N=36) = .562, p < .001. Perhaps one reason so many social psychology multi-site replications fail is that the authors attempt to replicate original studies that featured no social interaction. Overall, 75% of original studies that authors tried to replicate contained no social interaction at all.

Year of publication was positively related to successful replications, r_s (N = 36) = .403, p = .015, which could signal that the field is improving at conducting these projects over time. Alternatively, this could mean that at the beginning of the replication efforts, about 10 years ago, effects were chosen because they were thought to be unreliable.

Replications that included the original authors fared better than those that did not, ϕ (N = 36) = .327, p = .147. Exact replications also were more successful than modified replications, ϕ (N = 36) = .279, p = .246. It did not matter how many authors or labs were included on the multi-lab replication.

DISCUSSION

We have reported on all the multi-site social psychology replication attempts that have been published (as of 28 July 2022). Their record is admittedly dispiriting. Effect sizes are uniformly much smaller than the original studies. The majority are nonsignificant, with only four yielding clear significant support for the original hypothesis and a few others presenting mixed results. The few successful ones often followed up previous attempts that had failed to replicate, which suggests it often may take some learning to conduct an effective and successful replication.

Whom should one believe? When an original, significant finding encounters a multi-site replication that is nonsignificant, scientists must decide whether to continue believing the hypothesis. Cautious skepticism should be maintained in both directions. The multi-site replications have several clear advantages in terms of credibility, including large samples, preregistration, and often greater transparency. Our review has also noted some issues with them, however, including often high rates of discarded data, weak or absent manipulation checks, and a general impression of low engagement among participants.

The failed replications certainly cast some doubt on the published research literature. Due to publication bias, early null results are often not reported. In social psychology especially, the effect sizes in the published literature appear to be generally inflated. (To be sure, that does not guarantee that the multi-site replication's effect size is true or correct.) Some *p*-hacking practices are likely to have inflated effect sizes, indeed helping weak real effects attain significance with underpowered samples. The more damning view, that *p*-hacking and publication bias have created a surfeit of false positive findings — ostensibly significant evidence for nonexistent effects — seems less plausible to us, especially with findings that have been replicated several times. Nevertheless, we acknowledge that other social psychologists believe that the research literature is full of bogus evidence purporting to support hypotheses that are objectively false (e,g., Schimmack, 2020). Such a dismal outlook cannot be refuted from this review.

More precisely, the pessimistic view that the social psychology literature is full of false findings can neither be validated nor refuted based on the currently available multi-site replications. It is true that most of the multi-site replications have yielded nonsignificant findings and only four of 36 (11.1%) yielded clear, significant support (and three of those with reduced effect sizes). Crucially, however, these are hardly a representative sample of hypotheses and findings in social psychology. The selection of what to replicate and which procedures to use requires some attention from the field as a whole.

Moreover, there are indeed many successful replications apart from the limited set of multi-site projects. It is worth noting that several classic social psychology studies have been replicated multiple times in individual labs, despite the lack of multi-site testing. For example, the original Solomon Asch (1955) conformity study was replicated by Asch himself (1956) as well as others (e.g., Allen & Crutchfield, 1963), including in several other countries besides the United States, such as Bosnia and Herzegovina (Ušto et al., 2019) Japan (Takano & Sogon, 2008), Kuwait (Amir, 1984), Portugal (Neto, 1995), and The Netherlands (Vlaander, & Van Rooijen, 1985). Similarly, the Stanley Milgram (1963) obedience study has been replicated in many different contexts by Milgram himself (1974), as well as others (Burger, 2009), including in other countries such as in Poland (Doliński et al., 2017), in a French "real" TV game show (Beauvois et al., 2012; Bègue et al., 2015), and in a virtual reality environment (Dambrun & Valentiné, 2010). All of these successful replications involved social interaction. It is however impossible to know how many failed replications there have been.

Even the multi-site findings are not as uniformly discouraging as they initially seem. As we noted, many of the nonsignificant results are accompanied by null results on a manipulation check. We have called these operational failures and emphasize that insofar as the manipulation fails, the study does not constitute a test of the hypothesis. The original theoretical point can only be disconfirmed when a study provides a significant and presumably substantial difference on the manipulation check while also providing a null result on the dependent measure. Ideally, the sensitivity of the dependent variable should also be verified, by showing that it can detect some (real) differences under the same conditions. While there are some true failures to replicate that meet these criteria, most of those covered here do not. Indeed, there were several cases in which an ostensibly failed multi-site replication yielded significant support for the original hypothesis, if one corrects for the weak manipulation check and hence bases the analysis on the minority of participants for whom the manipulation was successful (e.g., Cheung et al., 2016; Dang, 2016).

Our review suggests it would be mistaken to regard the multi-site method as the best, most objective test of a hypothesis. For now, the multi-site method appears to be a fairly weak way of verifying hypotheses, possibly biased toward false negative findings. Perhaps for the present it would be appropriate to treat multi-site replications similarly to original findings: Significant positive results are precious and informative, whereas nonsignificant results are often ambiguous (especially without evidence that the replication was an effective test, such as by strong effects on manipulation checks).

Explaining Failures

Here we revisit the multiple possible explanations for failure outlined in the introduction. There was some evidence for each type. The formidable assortment of failed multi-site replications does not all fit neatly into a single explanation.

Hypothesis was wrong. If the original finding was spurious, the theoretical point can be contradicted by a multi-site replication. Several studies met the criteria of reporting both a significant and large difference on the manipulation check and a null result on the dependent variable. Only such a combination can justify the conclusion that the hypothesis was effectively tested and falsified by the multi-site replication. The following were the pure falsified hypotheses: that priming warm/cold with a handheld pack would cause more prosocial behavior (Lynott et al., 2014); that thinking about comfort foods reduces feelings of loneliness (Ong et al, 2015); that holding a red item makes a woman seem sexier (Pollett et al., 2019); and that trigger warnings increase feelings of vulnerability (Bellet et al., 2020). Wagenmakers et al. (2016) also provided a seemingly true failure to replicate the finding that holding a pen in lips vs. mouth, thereby evoking facial feedback of smile or frown, would alter ratings of how funny some cartoons are – but subsequent work has reaffirmed the hypothesis by showing that the Wagenmakers video manipulation check counteracted the manipulation (Noah et al., 2018).

The possibility that manipulation checks could affect the dependent variable measures (for another example, see Kühnen, 2010) presents a challenge for all psychological research. The challenge is exacerbated with multi-site replications. Given the high failure rate of such projects, it is vital to know whether the manipulation itself failed. Adding a manipulation check when the original procedure lacked one is desirable but may alter the manipulation decisively, as in the Wagenmakers case. We recommend that multi-site projects administer manipulation checks to only half or two-thirds of the sample, thereby making it possible to test whether the check altered responses on the main dependent measure.

There was another set of studies that we judged did not need manipulation checks because the manipulation was unmissable. These hypotheses therefore also count as having been

falsified: writing about one's imaginary life as a professor (vs. soccer hooligan) would improve mental performance on a trivia test (O'Donnell et al., 2018); telling people to respond fast (vs. slow) on a social goods dilemma would make them behave more prosocially (though note that the minority who did respond within the time did replicate the original finding) (Bouwmeester et al., 2017); that priming people with action words would cause improved cognitive performance on Scholastic Aptitude Test (SAT) items (Chartier et al., 2020); and that people believe more in "climate change" than "global warming" (Soutter et al., 2020).

The Elaboration Likelihood Model's (ELM) finding that persuasion would reflect a three-way interaction among personal involvement, endorser identity, and argument strength was not replicated by Kerr et al. (2015), and one sample did have significant manipulation check differences on all three independent variables[4] – but the others did not show that the crucial manipulation of argument strength was successful, and their manipulation checks for the other two variables were weak. Thus, it was at best a feeble test. Our view is that the positive results for the Elaboration Likelihood Model by Ebersole et al. (2017) should take precedence.

Our theory proposed another sign that would indicate that replication failures falsify the original hypothesis. Specifically, hypotheses that had many previous replications would presumably be more successful than hypotheses that had only been found once or twice. Later, we discuss the implications of multi-site replications of hypotheses that had already garnered extensive previous support. For now, the point is that multi-site replication failures do not uniformly indicate that the original hypothesis was wrong, but in some cases this conclusion is justified.

Operational Failures: Manipulation failure. A failed manipulation check means the hypothesis was not tested. These do raise concerns about why the manipulation did not work as intended, but the findings are not sufficient to overturn the original findings and reject the hypothesis.

There were multiple clear instances of manipulation failure: Cheung et al (2016), Buttrick et al. (2020), IJzerman et al (2020), Hagger et al (2016), Corker et al (2020); and DeJong et al. (2009). These six replications constitute 17% of the total sample. Moreover, the true number of operational failures is almost certainly higher, given that many failed replications did not report manipulation checks. If the ratio is the same as with studies that did report manipulation checks, then about half those ambiguous ones will also be operational failures.

All of this suggests another dimension to the so-called replication crisis. The procedures that manipulated independent variables successfully in the original study failed in replication to produce the intended difference between the experimental and control groups. To illustrate: If a high-anxiety treatment condition reports the same anxiety level as the control group, the experiment cannot demonstrate anything about anxiety. But it raises the important question of why that manipulation did make the original sample, yet not the replication sample, differentially anxious.

Operational Failure: Low engagement among participants. Another form of operational failure involves low engagement by participants, so that they are not emotionally or

[4] The endorser manipulation involved celebrities who were only recognized as such by a majority of the United Kingdom (UK) sample, not the other samples, so in that sense the manipulation was confirmed in the UK sample. However, liking for the celebrities was not higher in the UK sample than liking for ordinary citizen endorsers. One might argue that that is sufficient to invalidate the study as an effective test of the hypothesis, insofar as liking is needed.

motivationally engaged in the situation. We proposed two signs that participants in replication studies are not heavily engaged and therefore fail to exhibit the hypothesized responses. One was high rates of excluding data based on some quality deficit, such as not following instructions. The other was much weaker effect sizes on the manipulation check (which might be a sign of other problems, but would nevertheless follow from low engagement). There was abundant evidence of both problems.

As covered in the Results section, many of the multi-site replications discarded large amounts of data, sometimes over half. Some also discarded more from one condition than another (and others failed to report such a breakdown). Typically, these followed from pre-registered criteria. In two cases, excluding large amounts of data yielded the only significant positive finding — but more commonly, the exclusion of large amounts of data yielded no benefit or even made the findings weaker.

Insofar as low engagement contributes to the replication failures, it would be desirable to acknowledge this and take steps to correct it. Manipulation checks are typically used to evaluate the sample and procedure, but perhaps one could use them to select subsamples that responded best to the manipulation. For example, one could select half the participants in the experimental condition whose manipulation checks indicated the strongest reaction and analyze them specifically. To be sure, such selection introduces possible confounds, such as individual difference dimensions. Nevertheless, a replication can claim some support for the original finding if the effect on the dependent variable is notably stronger among participants whose manipulation checks indicate that they responded as intended (as opposed to participants whose manipulation check data indicate little or no effect).

We have already noted that effect sizes (including for manipulation checks) are weaker, often much weaker, in multi-site replications (indeed in other replications too) than in original social psychology findings. There may be multiple reasons for this, including (we assume) inflation of effect sizes in original studies based on publication bias and *p*-hacking, and regression to the mean. But weak effect sizes are also consistent with low engagement.

The low engagement findings point to a common and destructive misconception in the field. It is tempting to assume that a multi-site replication is a strong test of the hypothesis, because of the large sample, which should bolster statistical significance. But quite possibly the putative strength stemming from the large sample is nullified and counteracted by the low engagement. Multi-site investigations seem to be quite weak rather than strong tests, and the low engagement of participants seems a prominent reason for that. Moreover, engagement is not a dichotomous variable, and it may be misleading to assume that the 25% discarded participants were not engaged while the remaining 75% were highly engaged. Rather, high rates of discarding may indicate that the whole sample was infected with low engagement, even if some participants did take it all seriously and respond earnestly.

The low engagement may be partly due to social psychology's shift away from the highly involving personal experiences cultivated by early researchers and toward collecting data from participants sitting alone at computers (see section on live interaction, below). Oppenheimer, Meyvis, and Davidenko (2009) highlighted the need for attention checks in online studies, reporting that over a third (46% and 35%) of participants in their two studies failed them. This could be interpreted as indicating that the rest are fully involved, so that the 35% or 46% merely supply large amounts of error variance to dilute the results. Alternatively, if one assumes a continuum of motivational engagement, the 35 or 46% are not engaged while many of the others (who manage to pass the attention check) may be still only barely complying — so that most of

the sample is compromised by low engagement. Oppenheimer et al. reported another (pilot) study in which they reduced the failure rate to 14% among highly motivated participants, which again confirms the importance of subjective motivation and engagement. In contrast, many online samples may lack such high motivation. Webb and Tangney (in press) reported multiple checks for quality responding with an MTurk sample, and each check eliminated substantially more data — fitting the view that the full sample was infected with low engagement, which can only be found in different ways.

Low engagement will only affect some social psychology findings. As noteworthy examples, anchoring-adjustment and false consensus fared well among the mini-study replications (Klein et al., 2014, 2018). Simple mental mistakes may replicate well (or even better) among people who are not highly engaged. Likewise, Schweinsberg et al. (2016) had very good success replicating moral judgments of hypothetical situations, which obviously did not require personal involvement in any fashion (and indeed they discarded no data at all). In contrast, the early forms of social psychology often depended on creating highly involving experiences, and participant disengagement could be fatal to attempts to replicate those.

Lack of Social Interaction. Over the years, social psychology experiments have shifted from elaborately staged, highly involving live interactions to reliance on solitary individuals sitting at computers making ratings (see Baumeister et al., 2007). Hauser et al. (2018) note that online materials "are typically relatively uninvolving" (p. 998). We went back and examined the procedures for actual social interaction and found a strong relationship to replication success.

The single most successful multi-site replication we found (Ito et al., 2019) was the only one to feature live, unscripted social interaction. Perhaps not coincidentally, it was the only one outside the mini-studies to replicate the original effect size. Three others featured live, ongoing interaction with the experimenter, and they also were either complete or mixed successes. One of the other full successes involved having participants observe a supposedly unscripted social interaction and tracked their eye movements during it. Two studies used computers but simulated live, ongoing interactions, and these had some albeit weak findings consistent with the original (Bouwmeester et al., 2017; Skorb et al., 2020). Thus, the few studies featuring genuine interpersonal interaction replicated reasonably well.

In sharp contrast, the failures contained all the studies conducted by having solitary participants making ratings on computer, or in a few cases on paper, as well as the studies run in large rooms with participants again responding in solitary fashion rather than interacting. None of the failed replications featured live social interaction[5].

Thus, there is a substantial difference in replication success as a function of whether the procedure included human-to-human social interaction or was conducted as a mostly solitary procedure. This likely overlaps substantially with the point about low engagement, insofar as participants become more involved when dealing with another human being than merely sitting at a computer making solitary ratings and at best imagining a social event (e.g., Hauser et al., 2018). The preference for the latter (solitary) sort of procedure may reflect the constraints under which many multi-lab replications operate. The need for large samples makes labor-intensive interpersonal interactions costly. But the costs of relying on computers as the essential medium for social psychology research may include impairments in replicability.

[5] Skorb et al. (2020) could be considered a failure, given that the one significant finding was heavily compromised by discarding over half the data.

The opposite result would have been plausible. After all, it easier to standardize a computer-administered protocol than a study containing live, semi-unscripted interaction. Method-based error variance should seemingly be higher in the live interaction studies — yet they replicated better, not worse.

We had initially suspected that replicability suffered when researchers switched from original studies containing live interaction to featuring computer-administered procedures. Although there were a few such cases, in general the original and the multi-site replication were the same in terms of whether there was live interaction. Any difference due to live social interaction operated mainly at the level of selection of studies to replicate rather than different procedures for testing the same hypothesis. We recommend that in selecting future topics for multi-site replication, social psychologists give some priority to findings involving live interpersonal interaction.

To be sure, the mini-studies in the Klein et al. (2014, 2018; also Schweinsberg et al., 2016) papers were administered by computer and had a notably higher success rate than the full studies. Again, however, these focused on phenomena that do not depend on personal engagement, such as making moral judgments about hypothetical vignettes, or estimating numbers. If social psychology desires replicable successes without having to include live interactions, it may profit by focusing mainly on studying how people think about things they have no personal reason to care about.

A speculative solution is that live social interaction engages the individual more fully than computer-administered questionnaires. That would explain the benefits of live interaction for social psychology studies, as well as the higher replicability of judgement and decision making and other findings that do not depend on high engagement, such as the estimation error in anchoring-adjustment.

Editorial Bias Favoring Failure. It is uncomfortable to discuss possible systemic bias in the editorial system, though critics have long (and quite plausibly) asserted that the published literature can be misleading because of editorial bias in favor of significant findings. The contribution of any such bias to the replication debate is difficult to gauge, especially with objective data, so any discussion here is impressionistic. Anecdotal evidence is consistent with this interpretation (Schmeichel, personal communication.) Vohs et al. (2021) found both significant and nonsignificant results depending on the exclusion of over a thousand participants. The editor directed them to feature the nonsignificant results in the published article, based on the preregistered analyses, while consigning the significant results to the Supplementary Online materials. This contrasts with the usual and best practice, which is to report results both ways when there are many exclusions. It is impossible to know whether this reflects a general pattern, but some evidence would be found insofar as studies with mixed results are reported mainly as failed replications (thus emphasizing the nonsignificant rather than the significant findings).

No strong conclusions can be drawn regarding possible editorial bias, but we think it worth mentioning, in part because publication bias is widely assumed to contribute to the inflated effect sizes among original findings. Galiani et al. (2017) surveyed journal editors in economics and found that they preferred to publish failed rather than successful replications. Such a preference would be understandable, insofar as journal editors seek to preserve their much sought-after journal space for new information that advances the field. A successful replication provides no new knowledge, in an important sense, because it merely confirms what has already been found. In contrast, a failed replication suggests that currently held beliefs should be

questioned, revised, or even discarded. Hence an editor might plausibly believe that a failed replication is a more important and newsworthy contribution than a successful one.

We note that a multi-site replication offers a rare opportunity to conduct a meta-analysis with zero publication bias. That assumes that the journal has agreed in advance (i.e., before data collection) to publish the paper, whatever the outcome. All studies in the project can then be included in the meta-analysis. To be sure, the typically high rate of data exclusion does compromise the validity to some degree. But the meta-analysis of the multi-site replications includes all studies that were part of it, regardless of outcome.

Explaining Success

Although our focus has been on the failures of multi-site replications, it is worth considering the successes together. As already reported, the big multi-site replications focusing on a single theory or hypothesis yielded four successes, as follows. Eyewitnesses who confer among themselves before testifying (in simulated trial situation) influence each other's testimony. Prior exercise of self-regulation leads to impaired performance on a subsequent, different test of self-regulation (i.e., ego depletion). The personality trait of need for control interacts with quality of persuasive argument to influence attitude change (i.e., Elaboration Likelihood model). Last, people look at someone whom they expect to speak.

These are a motley group, but one thing they have in common is being heavily cognitive. (Ego depletion is not necessarily a cognitive phenomenon, but the successful multi-site replication by Dang et al., 2021, used highly cognitive procedures.) Cognitive effects have tended to replicate better than more purely social ones (Open Science Collaboration, 2015; Wilson & Wixted, 2018), and so relying more on cognitive procedures may increase replicability, as these results suggest.

The idea that focusing on cognitive processes will improve replicability gains credence from the mini-studies, which had a higher success rate than the experiments devoted to a single hypothesis or theory. These assessed quick thought reactions and reveal common biases and mental mistakes. Crucially, they do not rely on emotional or motivational engagement by participants. For example, the finding that people blame a (hypothetical) man who accidentally hurts a baby more than they blame a baby who accidentally hurts a grown man (replicated by Klein et al.) probably does not require deep emotional involvement or careful thought. The same goes for a false consensus effect, in which people estimate that many others would share their opinions. This line of thought suggests an alternative way forward, which is for social psychology to dispense with studying phenomena that engage people's motivations and limit research to quick thought-reaction procedures. Such an approach (which does appear to be the trend in the field at present) may have the benefit of improving replicability in multi-site online procedures, though some would object that there are hidden costs in neglecting to study more highly involving, behavioral phenomena.

Focusing more on socially cognitive processes would be a departure from the roots of social psychology, which had a strong behavioral focus despite the cognitive thrust of some early dominant theories (e.g., cognitive dissonance[6], attribution). Nevertheless, it seems well suited to the current preference for online methods, easily administered procedures, and large samples. Focusing mainly on how people think about the social world and how they think they would react to hypothetical situations offers a fertile ground for further research, and if it combines with the promise of improvements in replicability, it could be a good way to go.

[6] Purists will note that so-called cognitive dissonance was in fact more a motivational than a cognitive theory and was studied as such for at least its first decade or two.

To be sure, the dismal record of social priming in multi-site replications sends a cautionary message about shifting toward an ever-more-cognitive social psychology. Priming, may seem at first blush to be a cognitive phenomenon — though accumulating evidence suggests that it relies heavily on motivation (Weingarten et al., 2016). There have been over a dozen multi-site attempts to replicate priming effects, all of which failed. (One mini-study did find a significant result, but that unfortunately seems confounded.) Priming strikes us as the primary upcoming battleground for the theoretical implications of multi-site replications. There is a large volume of published findings in support of priming, but the multi-site record provides no encouragement for believing that the phenomenon is real.

More broadly, these multi-site replications have created a paradoxical pattern in the research and publication process. Authors continue to write articles and to cite previously published studies. It is apparently fine to cite findings that have obtained only once or twice. Meanwhile, however, findings that have been obtained dozens of times lose credibility once they are tested and dismissed in multi-site replications. New authors must therefore build their theoretical case on shaky grounds. Most scientists would presumably agree that findings that have been obtained dozens of times and by multiple laboratories have more credibility than those that have been found only once or twice. Social psychology may be moving toward the opposite assumption.

CONCLUDING REMARKS

The multi-site replication once held appeal as a definitive method for verifying the truth of social psychology's findings and theories. Thus far, it has failed to confirm multiple well-replicated findings and brought the field's credibility into serious doubt. We freely concede that it is possible to take this record as a sign that social psychology, as practiced for the past half century, has been an exercise in futility marked by dubious theories built around false positive findings. Nevertheless, we think it is also possible to maintain a more positive attitude about social psychology's research literature. The multi-site replication, perhaps especially as administered by impersonal procedures and characterized by low participant engagement, could be a weak and flawed method for verifying social psychology findings. Its drawbacks may be especially problematic for social psychology, and multi-site replications do seem to succeed better in other fields, though such a comparison is beyond the scope of our review. The disappearance of live interpersonal interaction from social psychology's methods may be a particularly costly loss: We found that studies with actual human interaction replicated much better than ones relying on solitary responses. The most effective social psychology multi-site replications involve actual social interaction. Thus, as an alternative to making social psychology more cognitive, there may be utility in revisiting the more social side of social psychology.

Moreover, the purpose and value of the laboratory experiment, as developed in the early years of social psychology, can be reconsidered. Our impression is that such early researchers did not look upon their findings as establishing infinitely replicable laws of nature but as showing that certain causal effects could be obtained under optimal conditions. Even cognitive dissonance, which dominated the field's thinking during its formative years, was not always found. The appropriate conclusion based on significant social psychology experimental findings could perhaps be characterized as "sometimes this happens." Indeed, some areas of research have already begun acknowledging this explicitly. Moral licensing patterns, for example, have been found repeatedly – but so have significant findings in the opposite direction (Mullen & Monin, 2016).

Although "sometimes this happens" may be disappointing as compared to establishing universal laws, perhaps the field should accept this with both humility and pride. It is valuable to demonstrate regular patterns in behavior, even if they will not be found all the time across diverse circumstances and populations. Historians, for example, seek to establish what happened but do not expect to end up with a robust grand theory of history that will explain all the past and predict all the future. Social psychologists might likewise be content with showing that independent variable *X* sometimes causes dependent variable *Y*, without expecting that to occur under all circumstances. This would map the field's future agenda as containing at least two steps: (1) show that causal patterns occur sometimes, and (2) identifying boundary conditions to indicate when it does and does not occur. Such an approach may enable social psychologists to build on the positive achievements and contributions of earlier generations, while also weeding out the false positive findings and shelving those that occur only under rare circumstances. Multi-site replications may have an important and even constructive role in that sort of future. That could at least be a more productive and constructive way of viewing social psychology's research activities than the current one of producing provocative, important findings and then discarding them based on failed multi-site replications.

References

Note. Multi-lab replications included in the analyses are marked with an asterisk (*).

Albarracín, D., Handley, I. M., Noguchi, K., McCulloch, K. C., Li, H., Leeper, J., Brown, R. D., Earl, A., & Hart, W. P. (2008). Increasing and decreasing motor and cognitive output: A model of general action and inaction goals. *Journal of Personality and Social Psychology*, *95*(3), 510–523. doi:10.1037/a0012833

Allen, V. L., & Crutchfield, R. S. (1963). Generalization of experimentally reinforced conformity. *Journal of Abnormal and Social Psychology*, *67*(4), 326–333. doi:10.1037/h0042074

Alpert, W. T., Couch, K. A., & Harmon, O. R. (2016). A randomized assessment of online learning. *American Economic Review*, *106*(5), 378-382. doi:10.1257/aer.p20161057

Amir, T. (1984). The Asch conformity effect: A study in Kuwait. *Social Behavior and Personality: An International Journal*, *12*(2), 187–190. doi:10.2224/sbp.1984.12.2.187

Asch, S. E. (1955, November). Opinions and social pressure. *Scientific American, 193*(5), 31–35. doi:10.1038/scientificamerican1155-31

Asch, S. E. (1956). Studies of independence and conformity: I A minority of one against a unanimous majority. *Psychological Monographs: General and Applied*, *70*(9), 1–70. doi:10.1037/h0093718

*Baranski, E., Baskin, E., Coary, S., Ebersole, C. R., Krueger, L. E., Lazarević, L. B., Miller, J. K., Orlić, A., Penner, M. R., Purić, D., Rife, S. C., Vaughn, L. A., Wichman, A. L., & Žeželj, I. (2020). Many Labs 5: Registered replication of Shnabel and Nadler (2008), Study 4. *Advances in Methods and Practices in Psychological Science*, *3*(3), 405–417. doi:10.1177/2515245920917334

Bargh, J.A., & Melnikoff, D. (2019). Does physical warmth prime social warmth? *Social Psychology*, *50*(3), 207-210. doi:10.1027/1864-9335/a000387

Baumeister, R. F., & Bushman, B. J. (2021). *Social psychology and human nature* (5th ed.). Belmont, CA: Wadsworth.

Baumeister, R. F., Vohs, K. D., & Funder, D. C. (2007). Psychology as the science of self-reports and finger movements: Whatever happened to actual behavior? *Perspectives on Psychological Science, 2*, 396-403. doi:10.1111/j.1745-6916.2007.00051.x

Baumeister, R. F., & Tice, D. M. (1990). Anxiety and social exclusion. *Journal of Social and Clinical Psychology*, *9*(2), 165–195. doi:10.1521/jscp.1990.9.2.165

Baumeister, R.F., & Tice, D.M. (2022). Ego depletion is the best replicated finding in all of social psychology. *Scholarly Journal of Psychology and Behavioral Sciences*, *6*. MS ID 000234. Doi: 10.32474/SJPBS.2021.06.000234.

Beauvois, J.-L., Courbet, D., & Oberlé, D. (2012). The prescriptive power of the television host A transposition of Milgram's obedience paradigm to the context of TV game show. *European Review of Applied Psychology / Revue Européenne de Psychologie Appliquée*, *62*(3), 111–119. doi:10.1016/j.erap.2012.02.001

Bègue, L., Beauvois J-L., Courbet, D., Oberlé, D., Lepage, J., & Duke, A. A. (2015). Personality predicts obedience in a Milgram paradigm. *Journal of Personality*, *83*(3), 299–306. doi:10.1111/jopy.12104

Bellet, B. W., Jones, P. J., & McNally, R. J. (2018). Trigger warning: Empirical evidence ahead. *Journal of Behavior Therapy and Experimental Psychiatry*, *61*, 134–141. doi:10.1016/j.jbtep.2018.07.002

*Bellet, B. W., Jones, P. J., Meyersburg, C. A., Brenneman, M. M., Morehead, K. E., & McNally, R. J. (2020). Trigger warnings and resilience in college students: A preregistered replication and extension. *Journal of Experimental Psychology: Applied*, *26*(4), 717–723. doi:10.1037/xap0000270.supp (Supplemental)

Bem, D. J. (2011). Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology*, 100(3), 407–425. doi:10.1037/a0021524

Berkowitz, L., & Donnerstein, E. (1982). External validity is more than skin deep. *American Psychologist*, *37*(3), 245–257. doi:10.1037/0003-066X.37.3.245

*Bouwmeester, S., Verkoeijen, P. P. J. L., Aczel, B., Barbosa, F., Bègue, L., Brañas-Garza, P., Chmura, T. G. H., Cornelissen, G., Døssing, F. S., Espín, A. M., Evans, A. M., Ferreira-Santos, F., Fiedler, S., Flegr, J., Ghaffari, M., Glöckner, A., Goeschl, T., Guo, L., Hauser, O. P., … Wollbrant, C. E. (2017). Registered replication report: Rand, Greene, and Nowak (2012). *Perspectives on Psychological Science*, *12*(3), 527–542. doi:10.1177/1745691617693624

Bryan, C. J., Yeager, D. S., & O'Brien, J. M. (2019). Replicator degrees of freedom allow publication of misleading failures to replicate. PNAS Proceedings of the National Academy of Sciences of the United States of America, 116(51), 25535–25545. doi:10.1073/pnas.1910951116

Burger, J. M. (2009). Replicating Milgram: Would people still obey today? *American Psychologist*, *64*(1), 1–11. doi:10.1037/a0010932

*Buttrick, N. R., Aczel, B., Aeschbach, L. F., Bakos, B. E., Brühlmann, F., Claypool, H. M., Hüffmeier, J., Kovacs, M., Schuepfer, K., Szecsi, P., Szuts, A., Szöke, O., Thomae, M., Torka, A.-K., Walker, R. J., & Wood, M. J. (2020). Many Labs 5: Registered replication of Vohs and Schooler (2008), Experiment 1. *Advances in Methods and Practices in Psychological Science*, *3*(3), 429–438. doi:10.1177/2515245920917931

Camerer, C.F., Dreber, A., Holzmeister, F. *et al.* (2018). Evaluating the replicability of social science experiments in *Nature* and *Science* between 2010 and 2015. *Nature Human Behaviour*, **2,** 637–644. doi:10.1038/s41562-018-0399-z

*Chartier, C. R., Arnal, J. D., Arrow, H., Bloxsom, N. G., Bonfiglio, D. B. V., Brumbaugh, C. C., Corker, K. S., Ebersole, C. R., Garinther, A., Giessner, S. R., Hughes, S., Inzlicht, M., Lin, H., Mercier, B., Metzger, M., Rangel, D., Saunders, B., Schmidt, K., Storage, D., & Tocco, C. (2020). Many Labs 5: Registered replication of Albarracín et al (2008), Experiment 5. *Advances in Methods and Practices in Psychological Science*, *3*(3), 332–339. doi:10.1177/2515245920945963

*Cheung, I., Campbell, L., & LeBel, E. P. (2016). Registered replication report: Study 1 from Finkel, Rusbult, Kumashiro, & Hannon (2002). *Perspectives on Psychological Science*, *11*(5), 750–764. doi:10.1177/1745691616664694

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation* (pp. 37-94). Boston, MA: Houghton Mifflin.

*Corker, K. S., Arnal, J. D., Bonfiglio, D. B. V., Curran, P. G., Chartier, C. R., Chopik, W. J., Guadagno, R. E., Kimbrough, A. M., Schmidt, K., & Wiggins, B. J. (2020). Many Labs 5: Registered replication of Albarracín et al (2008), Experiment 7. *Advances in Methods and Practices in Psychological Science*, *3*(3), 340–352. doi:10.1177/2515245920925750

Crosby, J. R., Monin, B., & Richardson, D. (2008). Where do we look during potentially offensive behavior? *Psychological Science*, *19*(3), 226–228. doi:10.1111/j.1467-9280.2008.02072.x

Dambrun, M., & Valentiné, E. (2010). Reopening the study of extreme social behaviors: Obedience to authority within an immersive video environment. *European Journal of Social Psychology*, *40*(5), 760–773. doi:10.1002/ejsp.646

Dang, J. (2016). Commentary: A multilab preregistered replication of the ego depletion effect. *Frontiers in Psychology*, *7*, 1155. doi:10.3389/fpsyg.2016.01155.

*Dang, J., Barker, P., Baumert, A., Bentvelzen, M., Berkman, E., Buchholz, N., Buczny, J., Chen, Z., De Cristofaro, V., de Vries, L., Dewitte, S., Giacomantonio, M., Gong, R., Homan, M., Imhoff, R., Ismail, I., Jia, L., Kubiak, T., Lange, F., … Zinkernagel, A. (2021). A multilab replication of the ego depletion effect. *Social Psychological and Personality Science*, *12*(1), 14–24. doi:10.1177/1948550619887702

Danziger, S., Levav, J, & Avnaim-Pessoa, L. (2011). Extraneous factors in judicial decisions. PNAS, 108, 6889–6892. doi:10.1073/pnas.1018033108

DeJong, W., Schneider, S. K., Towvim, L. G., Murphy, M. J., Doerr, E. E., Simonsen, N. R., Mason, K. E., & Scribner, R. A. (2006). A Multisite Randomized Trial of Social Norms Marketing Campaigns to Reduce College Student Drinking. *Journal of Studies on Alcohol*, *67*(6), 868–879. doi:10.15288/jsa.2006.67.868

*DeJong, W., Schneider, S. K., Towvim, L. G., Murphy, M. J., Doerr, E. E., Simonsen, N. R., Mason, K. E., & Scribner, R. A. (2009). A multisite randomized trial of social norms marketing campaigns to reduce college student drinking: A replication failure. *Substance Abuse*, *30*(2), 127–140. doi:10.1080/08897070902802059

DeWall, C.N., Baumeister, R.F., Chester, D.S., & Bushman, B.J. (2016). How often does currently felt emotion predict social behavior and judgment? A meta-analytic test of two theories. *Emotion Review*, *8*, 136-163. doi:10.1177/1754073915572690

Dijksterhuis, A., & van Knippenberg, A. (1998). The relation between perception and behavior, or how to win a game of Trivial Pursuit. *Journal of Personality and Social Psychology*, *74*(4), 865–877. doi:10.1037/0022-3514.74.4.865

Doliński, D., Grzyb, T., Folwarczny, M., Grzybała, P., Krzyszycha, K., Martynowska, K., & Trojanowski, J. (2017). Would you deliver an electric shock in 2015? Obedience in the experimental paradigm developed by Stanley Milgram in the 50 years following the original studies. *Social Psychological and Personality Science*, *8*(8), 927–933. doi:10.1177/1948550617693060

*Ebersole, C. R., Alaei, R., Atherton, O. E., Bernstein, M. J., Brown, M., Chartier, C. R., Chung, L. Y., Hermann, A. D., Joy-Gaba, J. A., Line, M. J., Rule, N. O., Sacco, D. F., Vaughn, L. A., & Nosek, B. A. (2017). Observe, hypothesize, test, repeat: Luttrell, Petty and Xu (2017) demonstrate good science. *Journal of Experimental Social Psychology*, *69*, 184–186. doi:10.1016/j.jesp.2016.12.005

*Ebersole, C. R., Andrighetto, L., Casini, E., Chiorri, C., Dalla Rosa, A., Domaneschi, F., Ferguson, I. R., Fryberger, E., Giacomantonio, M., Grahe, J. E., Joy-Gaba, J. A., Langford, E. V., Nichols, A. L., Panno, A., Parks, K. P., Preti, E., Richetin, J., & Vianello, M. (2020). Many Labs 5: Registered replication of Payne, Burkley, and Stokes (2008), Study 4. *Advances in Methods and Practices in Psychological Science*, *3*(3), 387–393. doi:10.1177/2515245919885609

*Eerland, A., Sherrill, A. M., Magliano, J. P., & Zwaan, R. A. (2016). Registered replication report: Hart & Albarracín (2011). *Perspectives on Psychological Science*, *11*(1), 158–171. doi:10.1177/1745691615605826

Ellefson, M.R., & Oppenheimer, D.M. (2022). Is replication possible without fidelity? *Psychological Methods*, advance online publication. doi:10.1037/met0000473

Engzell, P., Frey, A., & Verhagen, M. D. (2021). Learning loss due to school closures during the COVID-19 pandemic. *Proceedings of the National Academy of Sciences*, *118*(17), e2022376118. doi.org/10.1073/pnas.2022376118

Embley, J., Johnson, L. G., & Giner-Sorolla, R. (2015). Reproducibility project: Replication report – Replication of Study 1 by Vohs & Schooler (2008). Retrieved from https://osf.io/uwt5f/

Eskine, K. J., Kacinik, N. A., & Prinz, J. J. (2011). A bad taste in the mouth: Gustatory disgust influences moral judgment. *Psychological Science*, *22*(3), 295–299. doi:10.1177/0956797611398497

Fay, A.J., & Maner, J.K. (2020). Interactive effects of tactile warmth and ambient temperature on the search for social affiliation. *Social Psychology*, *51*, 1990204. doi:10.1027/1864-9335/a000407

Fiedler, K. (2017). What constitutes strong psychological science? The (neglected) role of diagnosticity and a priori theorizing. *Perspectives on Psychological Science*, *12*, 46-61. doi:10.1177/1745691616654458

Fiedler, K., Kutzner, F., & Krueger, J.I. (2012). The long way from alpha-error control to validity proper: Problems with a short-sighted false-positive debate. *Perspectives on Psychological Science*, *7*, 661-669. doi:10.1177/1745691612462587

Fiedler, K., McCaughey, L., & Prager, J. (2021). Quo vadis, methodology? The key role of manipulation checks for validity control and quality of science. *Perspectives on Psychological Science*, *16*, 816-826. doi:10.1177/1745691620970602

Fiedler, K., & Prager, J. (2018). The regression trap and other pitfalls of replication science — illustrated by the rport of the Open Science Collaboration. *Basic and Applied Social Psychology*, *40*, 115-124. doi:10.1080/01973533.2017.1421953

Finkel, E. J., Rusbult, C. E., Kumashiro, M., & Hannon, P. A. (2002). Dealing with betrayal in close relationships: Does commitment promote forgiveness? *Journal of Personality and Social Psychology*, *82*(6), 956–974. doi:10.1037/0022-3514.82.6.956

Förster, J., Liberman, N., & Kuschel, S. (2008). The effect of global versus local processing styles on assimilation versus contrast in social judgment. *Journal of Personality and Social Psychology*, *94*(4), 579–599. doi:10.1037/0022-3514.94.4.579

Friese, M., Loschelder, D. D., Gieseler, K., Frankenbach, J., & Inzlicht, M. (2019). Is ego depletion real? An analysis of arguments. *Personality and Social Psychology Review*, *23*(2), 107–131. doi:10.1177/1088868318762183

Galiani, S., Gertler, P., & Romero, M. (2017). *Incentives for replication in economics*. Retrieved from https://www.nber.org/system/files/working_papers/w23576/w23576.pdf

Garrison, K. E., Finley, A. J., & Schmeichel, B. J. (2019). Ego depletion reduces attention control: Evidence from two high-powered preregistered experiments. *Personality and Social Psychology Bulletin*, *45*(5), 728–739. doi:10.1177/0146167218796473

Garry, M., French, L., Kinzett, T., & Mori, K. (2008). Eyewitness memory following a discussion: Using the MORI technique with a Western sample. *Applied Cognitive Psychology*, *22*(4), 431–439. doi:10.1002/acp.1376

Gervais, W. M., & Norenzayan, A. (2012). Analytic thinking promotes religious disbelief. *Science*, *336*(6080), 493–496. doi:10.1126/science.1215647

*Ghelfi, E., Christopherson, C. D., Urry, H. L., Lenne, R. L., Legate, N., Ann Fischer, M., Wagemans, F. M. A., Wiggins, B., Barrett, T., Bornstein, M., de Haan, B., Guberman, J., Issa, N., Kim, J., Na, E., O'Brien, J., Paulk, A., Peck, T., Sashihara, M., … Sullivan, D. (2020). Reexamining the effect of gustatory disgust on moral judgment: A multilab direct replication of Eskine, Kacinik, and Prinz (2011). *Advances in Methods and Practices in Psychological Science*, *3*(1), 3–23. doi:10.1177/2515245919881152

Gilbert, D. T., King, G., Pettigrew, S., & Wilson, T. D. (2016). Comment on "Estimating the reproducibility of psychological science." Science, *351*(6277). doi:10.1126/science.aad7243

Greenberg, J., Pyszczynski, T., & Solomon, S. (1986). The causes and consequences of a need for self-esteem: A terror management theory. In R. F. Baumeister (Ed.), *Public and private self* (pp. 189-212). New York: Springer-Verlag.

Greenberg, J., Pyszczynski, T., & Solomon, S. (1990). Anxiety concerning social exclusion: Innate response or one consequence of the need for terror management? *Journal of Social and Clinical Psychology, 9*, 202-213. doi:10.1521/jscp.1990.9.2.202

Gruijters, S. L. K. (2021). Making inferential leaps: Manipulation checks and the road towards strong inference. *Journal of Experimental Social Psychology*, *98*, 104251.

*Hagger, M. S., Chatzisarantis, N. L. D., Alberts, H., Anggono, C. O., Batailler, C., Birt, A. R., Brand, R., Brandt, M. J., Brewer, G., Bruyneel, S., Calvillo, D. P., Campbell, W. K., Cannon, P. R., Carlucci, M., Carruth, N. P., Cheung, T., Crowell, A., De Ridder, D. T. D., Dewitte, S., … Zwienenberg, M. (2016). A multilab preregistered replication of the ego-depletion effect. *Perspectives on Psychological Science*, *11*(4), 546–573. doi:10.1177/1745691616652873

Halloran, C., Jack, R., Okun, J. C., & Oster, E. (2021). Pandemic schooling mode and student test scores: evidence from US states. *National Bureau of Economic Research*. doi:10.3386/w29497

Hart, W., & Albarracín, D. (2011). Learning about what others were doing: Verb aspect and attributions of mundane and criminal intent for past actions. *Psychological Science*, *22*(2), 261–266. doi:10.1177/0956797610395393

Hauser, D.J., Ellsworth, P.C., & Gonzalez, R. (2018). Are manipulation checks necessary? *Frontiers in Psychology: Personality and Social Psychology*, *9*, 998. doi:10.3389/fpsyg.2018.00998

He L, Yang N, Xu L, Ping F, Li W, Sun Q, Li Y, Zhu H, Zhang H. (2021). Synchronous distance education vs traditional education for health science students: A systematic review and meta-analysis. *Medical Education 55*(3):293-308. doi: 10.1111/medu.14364

Heppen, J.B., Sorensen, N., Allensworth, E., Walters, K., Rickles, J., Taylor, S. S., & Michelman, V. (2017). The struggle to pass algebra: Online versus face-to-face credit recovery for at-risk urban students. *Journal of Research on Educational Effectiveness*, *10*(2), 272 - 296. https://doi.org/10.1080/19345747.2016.1168500

Hunniford, V.T., Grudniewicz, A., Fergusson, D.A., Montroy, J., Grigor, E., Lansdell, C., & Lalu, M.M. (2022). A systematic assessment of preclinical multilaboratory studies and a comparison to single laboratory studies. Unpublished manuscript, Ottawa Hospital Research Institute, Ottawa, Canada.

Hurley, P.J. (2015). Ego depletion: Applications and implications for auditing research. *Journal of Accounting Literature, 35*, 47-76. doi:10.1016/j.acclit.2015.10.001

*Ijzerman, H., Ropovik, I., Ebersole, C. R., Tidwell, N. D., Markiewicz, Ł., de Lima, T. J. S., Wolf, D., Novak, S. A., Collins, W. M., Menon, M., de Souza, L. E. C., Sawicki, P., Boucher, L., Białek, M., Idzikowska, K., Razza, T. S., Kraus, S., Weissgerber, S. C., Baník, G., … Day, C. R. (2020). Many Labs 5: Registered

replication of Förster, Liberman, and Kuschel's (2008) Study 1. *Advances in Methods and Practices in Psychological Science*, *3*(3), 366–376. doi:10.1177/2515245920916513

*Ito, H., Barzykowski, K., Grzesik, M., Gülgöz, S., Gürdere, C., Janssen, S. M. J., Khor, J., Rowthorn, H., Wade, K. A., Luna, K., Albuquerque, P. B., Kumar, D., Singh, A. D., Cecconello, W. W., Cadavid, S., Laird, N. C., Baldassari, M. J., Lindsay, D. S., & Mori, K. (2019). Eyewitness memory distortion following co-witness discussion: A replication of Garry, French, Kinzett, and Mori (2008) in ten countries. *Journal of Applied Research in Memory and Cognition*, *8*(1), 68–77. doi:10.1016/j.jarmac.2018.09.004

Karau, S. J., & Williams, K. D. (1993). Social loafing: A meta-analytic review and theoretical integration. *Journal of Personality and Social Psychology*, *65*(4), 681–706. doi:10.1037/0022-3514.65.4.681

Keller, T., & Kiss, H.J. (2021). Do exhausted primary school students cheat more? A randomized field experiment. *PLOS One, 16 (12),* e0260141. doi:10.1371/journal.pone.0260141

*Kerr, G., Schultz, D. E., Kitchen, P. J., Mulhern, F. J., & Beede, P. (2015). Does traditional advertising theory apply to the digital world? A replication analysis questions the relevance of the elaboration likelihood model. *Journal of Advertising Research*, *55*(4), 390–400. doi:10.2501/JAR-2015-001

Kidd, D. C., & Castano, E. (2013). Reading literary fiction improves theory of mind. *Science*, *342*(6156), 377–380. doi:10.1126/science.1239918

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Jr., Bahník, Š., Bernstein, M. J., Bocian, K., Brandt, M. J., Brooks, B., Brumbaugh, C. C., Cemalcilar, Z., Chandler, J., Cheong, W., Davis, W. E., Devos, T., Eisner, M., Frankowska, N., Furrow, D., Galliani, E. M., … Nosek, B. A. (2014). Investigating variation in replicability: A "many labs" replication project. *Social Psychology*, *45*(3), 142–152. doi:10.1027/1864-9335/a000178

Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams, R. B., Jr., Alper, S., Aveyard, M., Axt, J. R., Babalola, M. T., Bahník, Š., Batra, R., Berkics, M., Bernstein, M. J., Berry, D. R., Bialobrzeska, O., Binan, E. D., Bocian, K., Brandt, M. J., Busching, R., … et al. (2018). Many labs 2: Investigating variation in replicability across samples and settings. *Advances in Methods and Practices in Psychological Science*, *1*(4), 443–490. doi:10.1177/2515245918810225

*Klein, R. A., Cook, C. L., Ebersole, C. R., Vitiello, C., Nosek, B. A., Chartier, C. R., Christopherson, C. D., Clay, S., Collisson, B., Crawford, J., Cromar, R., Vidamuerte, D., Gardiner, G., Gosnell, C., Grahe, J., Hall, C., Joy-Gaba, J., Legg, A. M., Levitan, C., Mancini, A., Manfredi, D., Miller, J. M., Nave, G., Redford, L., Schlitz, I., Schmidt, K., Skorinko, J., Storage, D., Swanson, T., van swol, L., Vaughn, L. A., & Ratliff, K. (2021). *Many labs 4: Failure to replicate mortality salience effect with and without original author involvement*. Retrieved from https://psyarxiv.com/vef2c/

Kofoed, M. S., Gilmore, D., Gebhart, L., & Moschitto, R. (2021). Zooming to class?: Experimental evidence on college students' online learning during COVID-19. Retrieved from https://docs.iza.org/dp14356.pdf

Kop, R., Fournier, H., & Mak, J. S. F. (2011). "A pedagogy of abundance or a pedagogy to support human beings? Participant support on massive open online courses", *International Review of Research in Open and Distance Learning*, vol. 12, pp. 74-93.

Kühnen, U. (2010). Manipulation checks as manipulation: Another look at the ease-of-retrieval heuristic. *Personality and Social Psychology Bulletin, 36*, 47-58. doi:10.1177/0146167209346746.

Kuhfeld, M., Soland, J., Lewis, K., & Morton, E. (2022: March 3). The pandemic has had devastating impacts on learning. Brown Center Chalkboard, https://www.brookings.edu/blog/brown-center-chalkboard/2022/03/03/the-pandemic-has-had-devastating-impacts-on-learning-what-will-it-take-to-help-students-catch-up/.

Lin, H. (2014). Red-colored products enhance the attractiveness of women. *Displays*, *35*(4), 202–205. doi:10.1016/j.displa.2014.05.009

Luttrell, A., Petty, R. E., & Xu, M. (2017). Replicating and fixing failed replications: The case of need for cognition and argument quality. *Journal of Experimental Social Psychology*, *69*, 178–183. doi:10.1016/j.jesp.2016.09.006

*Lynott, D., Corker, K. S., Wortman, J., Connell, L., Donnellan, M. B., Lucas, R. E., & O'Brien, K. (2014). Replication of "Experiencing physical warmth promotes interpersonal warmth" by Williams and Bargh (2008). *Social Psychology*, *45*(3), 216–222. doi:10.1027/1864-9335/a000187

*Mathur, M. B., Bart-Plange, D.-J., Aczel, B., Bernstein, M. H., Ciunci, A. M., Ebersole, C. R., Falcão, F., Ashbaugh, K., Hilliard, R. A., Jern, A., Kellier, D. J., Kessinger, G., Kolb, V. S., Kovacs, M., Lage, C. A., Langford, E. V., Lins, S., Manfredi, D., Meyet, V., … Frank, M. C. (2020). Many Labs 5: Registered multisite replication of the tempting-fate effects in Risen and Gilovich (2008). *Advances in Methods and Practices in Psychological Science*, *3*(3), 394–404. doi:10.1177/2515245918785165

Mazar, N., Amir, O., & Ariely, D. (2008). The dishonesty of honest people: A theory of self-concept maintenance. *Journal of Marketing Research*, *45*(6), 633–644. doi:10.1509/jmkr.45.6.633.

*McCarthy, R., Gervais, W., Aczel, B., Al-Kire, R. L., Aveyard, M., Marcella Baraldo, S., Baruh, L., Basch, C., Baumert, A., Behler, A., Bettencourt, A., Bitar, A., Bouxom, H., Buck, A., Cemalcilar, Z., Chekroun, P., Chen, J. M., del Fresno- Díaz, Á., Ducham, A., … Zogmaister, C. (2021). A Multi-Site Collaborative Study of the Hostile Priming Effect. *Collabra: Psychology*, *7*(1). doi:10.1525/collabra.18738

*McCarthy, R. J., Hartnett, J. L., Heider, J. D., Scherer, C. R., Wood, S. E., Nichols, A. L., Edlund, J. E., & Walker, W. R. (2018). An investigation of abstract construal on impression formation: A multi-lab replication of McCarthy and Skowronski (2011). *International Review of Social Psychology*, *31*(1). doi:10.5334/irsp.133

McCarthy, R. J., & Skowronski, J. J. (2011). You're getting warmer: Level of construal affects the impact of central traits on impression formation. *Journal of Experimental Social Psychology, 47,* 1304–1307. doi:10.1016/j.jesp.2011.05.017

*McCarthy, R. J., Skowronski, J. J., Verschuere, B., Meijer, E. H., Jim, A., Hoogesteyn, K., Orthey, R., Acar, O. A., Aczel, B., Bakos, B. E., Barbosa, F., Baskin, E., Bègue, L., Ben-Shakhar, G., Birt, A. R., Blatz, L., Charman, S. D., Claesen, A., Clay, S. L., … Yıldız, E. (2018). Registered replication report on Srull and Wyer (1979). *Advances in Methods and Practices in Psychological Science*, *1*(3), 321–336. doi:10.1177/2515245918777487

Milgram, S. (1963). Behavioral study of obedience. *Journal of Abnormal and Social Psychology*, *67*(4), 371–378. doi:10.1037/h0040525

Milgram, S. (1974). *Obedience to authority*. New York, NY: Harper

*Moran, T., Hughes, S., Hussey, I., Vadillo, M. A., Olson, M. A., Aust, F., Bading, K., Balas, R., Benedict, T., Corneille, O., Douglas, S. B., Ferguson, M. J., Fritzlen, K. A., Gast, A., Gawronski, B., Giménez-Fernández, T., Hanusz, K., Heycke, T., Högden, F., … De Houwer, J. (2021). Incidental attitude formation via the surveillance task: A preregistered replication of the Olson and Fazio (2001) study. *Psychological Science*, *32*(1), 120–131. doi:10.1177/0956797620968526

Mullen, E., & Monin, B. (2016). Consistency versus licensing effects of past moral behavior. *Annual Review of Psychology*, *67*, 363–385. doi:10.1146/annurev-psych-010213-115120

Neto, F. (1995). Conformity and independence revisited. *Social Behavior and Personality: An International Journal*, *23*(3), 217–222. doi:10.2224/sbp.1995.23.3.217

Noah, T., Schul, Y., & Mayo, R. (2018). When both the original study and its failed replication are correct: Feeling observed eliminates the facial-feedback effect. *Journal of Personality and Social Psychology*, *114*(5), 657–664. doi:10.1037/pspa0000121.supp (Supplemental)

Nosek, B. A., Hardwicke, T. E., Moshontz, H., Allard, A., Corker, K. S., Dreber, A., Fidler, F., Hilgard, J., Struhl, M. K., Nuijten, M. B., Rohrer, J. M., Romero, F., Scheel, A. M., Scherer, L. D., Schönbrodt, F. D., & Vazire, S. (2022). Replicability, robustness, and reproducibility in psychological science. *Annual Review of Psychology*, *73*, 719–748.

*O'Donnell, M., Nelson, L. D., Ackermann, E., Aczel, B., Akhtar, A., Aldrovandi, S., Alshaif, N., Andringa, R., Aveyard, M., Babincak, P., Balatekin, N., Baldwin, S. A., Banik, G., Baskin, E., Bell, R., Białobrzeska, O., Birt, A. R., Boot, W. R., Braithwaite, S. R., … Sherman, M. F. (2018). Registered replication report: Dijksterhuis and van Knippenberg (1998). *Perspectives on Psychological Science*, *13*(2), 268–294. doi:10.1177/1745691618755704

Olson, M. A., & Fazio, R. H. (2001). Implicit attitude formation through classical conditioning. *Psychological Science*, *12*(5), 413–417. doi:10.1111/1467-9280.00376

*Ong, L. S., Ijzerman, H., & Leung, A. K.-Y. (2015). Is comfort food really good for the soul? A replication of Troisi and Gabriel's (2011) Study 2. *Frontiers in Psychology*, *6.* doi:10.3389/fpsyg.2015.00314

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science*, *349(6251),* 1-8. doi:10.1126/science.aac4716

Oppenheimer, D.M., Meyvis, T., & Davidenko, N. (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology, 45,* 867-872. doi:10.1016/j.jesp.2009.03.009

*Panero, M. E., Weisberg, D. S., Black, J., Goldstein, T. R., Barnes, J. L., Brownell, H., & Winner, E. (2016). Does reading a single passage of literary fiction really improve theory of mind? An attempt at replication. *Journal of Personality and Social Psychology*, *111*(5), e46–e54. doi:10.1037/pspa0000064.supp (Supplemental)

Payne, B. K., Burkley, M. A., & Stokes, M. B. (2008). Why do implicit and explicit attitude tests diverge? The role of structural fit. *Journal of Personality and Social Psychology*, *94*(1), 16–31. doi:10.1037/0022-3514.94.1.16

Petty, R. E., Cacioppo, J. T., & Schumann, D. (1983). Central and peripheral routes to advertising effectiveness: The moderating role of involvement. *Journal of Consumer Research*, *10*(2), 135–146. doi:10.1086/208954

Philpot, L.M., Khokhar, B.A., Roellinger, D.L., Ramar, P., & Ebbert, J.O. (2018). Time of day is associated with opioid prescribing for low back pain in primary care. *Journal of General Internal Medicine*, online publication, doi:10.1007/s11606-018-4521-8

*Pollet, T. V., Costello, J., Groeneboom, L., Peperkoorn, L. S., & Wu, J. (2019). Do red objects enhance sexual attractiveness? No evidence from two large replications. *Displays*, *56*, 23–29. doi:10.1016/j.displa.2018.10.008

*Rabagliati, H., Corley, M., Dering, B., Hancock, P. J. B., King, J. P. J., Levitan, C. A., Loy, J. E., & Millen, A. E. (2020). Many Labs 5: Registered replication of Crosby, Monin, and Richardson (2008). *Advances in Methods and Practices in Psychological Science*, *3*(3), 353–365. doi:10.1177/2515245919870737

Rand, D. G., Greene, J. D., & Nowak, M. A. (2012). Spontaneous giving and calculated greed. *Nature*, *489*(7416), 427–430. doi:10.1038/nature11467

Riegg Cellini, S., & Grueso, H. (2021). Student learning in online college programs. *American Educational Research Association Open*. doi: 10.1177/23328584211008105

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*, 331-363. doi:10.1037/1089-2680.7.4.331

Risen, J. L., & Gilovich, T. (2008). Why people are reluctant to tempt fate. *Journal of Personality and Social Psychology*, *95*(2), 293–307. doi:10.1037/0022-3514.95.2.293

*Ritchie, S. J., Wiseman, R., & French, C. C. (2012). Failing the future: Three unsuccessful attempts to replicate Bem's "retroactive facilitation of recall" effect. *PloS ONE*, *7*(3). doi:10.1371/journal.pone.0033423

*Sanchez, C., Sundermeier, B., Gray, K., & Calin-Jageman, R. J. (2017). Direct replication of Gervais & Norenzayan (2012): No evidence that analytic thinking decreases religious belief. *PloS ONE*, *12*(2). doi:10.1371/journal.pone.0172636

Schuldt, J. P., Konrath, S. H., & Schwarz, N. (2011). "'Global warming"' or "'climate change"'? Whether the planet is warming depends on question wording. *Public Opinion Quarterly*, *75*(1), 115–124. doi:10.1093/poq/nfq073

Schweinsberg, M., Madan, N., Vianello, M., Sommer, S. A., Jordan, J., Tierney, W., Awtrey, E., Zhu, L. L., Diermeier, D., Heinze, J. E., Srinivasan, M., Tannenbaum, D., Bivolaru, E., Dana, J., Davis-Stober, C. P., du Plessis, C., Gronau, Q. F., Hafenbrack, A. C., Liao, E. Y., … Uhlmann, E. L. (2016). The pipeline project: Pre-publication independent replications of a single laboratory's research pipeline. *Journal of Experimental Social Psychology*, *66*, 55–67. doi:10.1016/j.jesp.2015.10.001

*Skorb, L., Aczel, B., Bakos, B. E., Feinberg, L., Hałasa, E., Kauff, M., Kovacs, M., Krasuska, K., Kuchno, K., Manfredi, D., Montealegre, A., Pękala, E., Pieńkosz, D., Ravid, J., Rentzsch, K., Szaszi, B., Schulz-Hardt, S., Sioma, B., Szecsi, P., … Hartshorne, J. K. (2020). Many Labs 5: Replication of van Dijk, van Kleef, Steinel, and van Beest (2008). *Advances in Methods and Practices in Psychological Science*, *3*(3), 418–428. doi:10.1177/2515245920927643

*Soutter, A. R. B., & Mõttus, R. (2020). "global warming" versus "climate change": A replication on the association between political self-identification, question wording, and environmental beliefs. *Journal of Environmental Psychology*. doi:10.1016/j.jenvp.2020.101413

Sripada, C., Kessler, D., & Jonides, J. (2014). Methylphenidate blocks effort-induced depletion of regulatory control in healthy volunteers. *Psychological Science*, *25*, 1227–1234. doi:10.1177/0956797614526415

Srull, T. K., & Wyer, R. S. (1979). The role of category accessibility in the interpretation of information about persons: Some determinants and implications. *Journal of Personality and Social Psychology*, *37*(10), 1660–1672. doi:10.1037/0022-3514.37.10.1660

Strack, F., Martin, L. L., & Stepper, S. (1988). Inhibiting and facilitating conditions of the human smile: A nonobtrusive test of the facial feedback hypothesis. *Journal of Personality and Social Psychology*, *54*(5), 768–777. doi:10.1037/0022-3514.54.5.768

Takano, Y., & Sogon, S. (2008). Are Japanese more collectivistic than Americans? Examining conformity in in-groups and the reference-group effect. *Journal of Cross-Cultural Psychology*, *39*(3), 237–250. doi:10.1177/0022022107313902

Trinh, P., Hoover, D.R., & Sonnenberg, F.A. (2021). Time-of-day changes in physician clinical decision making: A retrospective study. *PLOS One.* doi:10.1371/journal.pone.0257500

Troisi, J. D., & Gabriel, S. (2011). Chicken soup really is good for the soul: "comfort food" fulfills the need to belong. *Psychological Science*, *22*(6), 747–753. doi:10.1177/0956797611407931

Ušto, M., Drače, S., & Hadžiahmetović, N. (2019). Replication of the "Asch effect" in Bosnia and Herzegovina: Evidence for the moderating role of group similarity in conformity. *Psihologijske Teme*, *28*(3), 589–599. doi:10.31820/pt.28.3.7

van Dijk, E., van Kleef, G. A., Steinel, W., & van Beest, I. (2008). A social functional approach to emotions in bargaining: When communicating anger pays and when it backfires. *Journal of Personality and Social Psychology*, *94*(4), 600–614. doi:10.1037/0022-3514.94.4.600

*Verschuere, B., Meijer, E. H., Jim, A., Hoogesteyn, K., Orthey, R., McCarthy, R. J., Skowronski, J. J., Acar, O. A., Aczel, B., Bakos, B. E., Barbosa, F., Baskin, E., Bègue, L., Ben-Shakhar, G., Birt, A. R., Blatz, L., Charman, S. D., Claesen, A., Clay, S. L., … Yıldız, E. (2018). Registered replication report on Mazar, Amir, and Ariely (2008). *Advances in Methods and Practices in Psychological Science*, *1*(3), 299–317. doi:10.1177/2515245918781032

Vianello, M. (2015). Replication of Payne, Burkley, & Stokes (2008, *Journal of Personality and Social Psychology*, Study 4). Retrieved from https://osf.io/79y8g/

Vlaander, G. P., & Van Rooijen, L. (1985). Independence and conformity in Holland: Asch's experiment three decades later. *Gedrag: Tijdschrift Voor Psychologie*, *13*(1), 49–55.

*Vohs, K., Schmeichel, B., Lohmann, S., Gronau, Q. F., Finley, A. J., … Wagenmakers, E. J., & Albarracín, D. (2021). A multi-site preregistered paradigmatic test of the ego depletion effect. *Psychological Science*, *32*(10), 1566-158. doi:10.1177/0956797621989733

Vohs, K. D., & Schooler, J. W. (2008). The value of believing in free will: Encouraging a belief in determinism increases cheating. *Psychological Science*, *19*(1), 49–54. doi:10.1111/j.1467-9280.2008.02045.x

*Wagenmakers, E.-J., Beek, T., Dijkhoff, L., Gronau, Q. F., Acosta, A., Adams, R. B., Jr., Albohn, D. N., Allard, E. S., Benning, S. D., Blouin-Hudon, E.-M., Bulnes, L. C., Caldwell, T. L., Calin-Jageman, R. J., Capaldi, C. A., Carfagno, N. S., Chasten, K. T., Cleeremans, A., Connell, L., DeCicco, J. M., … Zwaan, R. A. (2016). Registered Replication Report: Strack, Martin, & Stepper (1988). *Perspectives on Psychological Science*, *11*(6), 917–928. doi:10.1177/1745691616674458

Webb, M., & Tangney, J. (in press). Too good to be true: Bots and other bad data from Mechanical Turk. *Perspectives on Psychological Science*.

Weingarten, E., Chen, Q., McAdams, M., Yi, J., Hepler, J., & Albarracín, D. (2016). From primed concepts to action: A meta-analysis of the behavioral effects of incidentally presented words. *Psychological Bulletin*, *142*(5), 472–497. doi:10.1037/bul0000030.supp (Supplemental)

Williams, L. E., & Bargh, J. A. (2008). Experiencing physical warmth promotes interpersonal warmth. *Science*, *322*(5901), 606–607. doi:10.1126/science.1162548

Wilson, B.M., & Wixted, J.T. (2018). The prior odds of testing a true effect in cognitive and social psychology. *Advances in Methods and Practices in Psychological Science, 1*, 186-197. Doi: 10.1177/251245918767122.

Zhou, H., & Fishbach, A. (2016). The pitfall of experimenting on the web: How unattended selective attrition leads to surprising (yet false) research conclusions. *Journal of Personality and Social Psychology, 111,* 493-504. doi:10.1037/pspa0000056