

## RESEARCH METHODS IN MOTIVATION SCIENCE

## Valid Replications Require Valid Methods: Recommendations for Best Methodological Practices With Lab Experiments

Eddie Harmon-Jones<sup>1</sup>, Cindy Harmon-Jones<sup>2</sup>, David M. Amodio<sup>3</sup>, Philip A. Gable<sup>4</sup>,  
and Brandon J. Schmeichel<sup>5</sup><sup>1</sup> School of Psychology, The University of New South Wales<sup>2</sup> School of Psychology, Western Sydney University<sup>3</sup> Faculty of Social and Behavioural Sciences, University of Amsterdam<sup>4</sup> Department of Psychological and Brain Sciences, University of Delaware<sup>5</sup> Department of Psychological and Brain Sciences, Texas A&M University

This article considers an important but relatively neglected contributor to the failure to replicate results from experiments in motivation and emotion—messy methods. By methods, we mean the procedures used to collect data and test hypotheses, which concern issues such as experimental design and validity. We offer a set of recommendations for establishing strong and valid experimental methods for both original research and replications. We first consider the lab room setup, starting with the physical environment participants see on the way to the lab, as well as the physical environment of the lab room itself. Then, we explain the importance of a cover story and the psychological state of the participants prior to the beginning of the experiment. In addition, we consider experimenters' and confederates' behavior and appearance, the need for experimenters to be blind to conditions, and the difficulties of having multiple experimenters conduct one experiment. Next, we discuss the construction of strong independent variables, interactions between independent variables, manipulation checks, how the psychological meaning of an independent variable can change over time and place, demand characteristics, and confounds. When considering dependent variables, we explain how to construct sensitive ones and the importance of pretesting. Then, we apply these recommendations to replications and finish by considering some data management and statistical issues. We hope this article will be a useful resource for both students and experienced scientists.

*Keywords:* research methods, replication, internal validity, manipulations, cover story


Several explanations for failed replications in psychology have been offered: p-hacking (Kerr, 1998); small, underpowered samples (Cohen, 1962); questionable research practices (Simmons et al., 2011); natural wide variability in *p* values (Cumming, 2013); and fraud. All these things have likely contributed to failures to replicate. Although these explanations focus primarily on issues related to

statistics, here we address what we believe is a crucial but often neglected additional contributor—messy methods. Although methods are sometimes equated with statistics in psychological science, methods and statistics are not the same. Whereas statistics concern the quantitative analysis of data, methods refer to the procedures used to collect data and test hypotheses. As such, methods concern issues such as experimental design and measurement, which in turn determine the validity of an experiment's measures, manipulations, and causal inferences. In this article, we discuss issues of methodology as they relate to replication failures in psychology experiments, with a focus on experiments on motivation and emotion. We start with some general observations on the importance of strong methods and, based on lessons we have learned over our careers of conducting experiments, offer some recommendations for enhancing replicability in lab experiments.

## Prologue

When two of us were in graduate school, we tried to grow mushrooms in our kitchen. We bought a kit, followed the instructions carefully, waited the appropriate amount of time, and then checked the Petri dishes only to find that we had grown mold and other things,

Rex A. Wright served as action editor.

Eddie Harmon-Jones  <https://orcid.org/0000-0003-4771-3043>

We thank Greg Hajcak for offering helpful comments on an earlier version of this article.

Open Access funding provided by The University of New South Wales: This work is licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0; <https://creativecommons.org/licenses/by/4.0>). This license permits copying and redistributing the work in any medium or format, as well as adapting the material for any purpose, even commercially.

Correspondence concerning this article should be addressed to Eddie Harmon-Jones, School of Psychology, The University of New South Wales, 1105 Mathews Building, UNSW Sydney, NSW 2052, Australia. Email: eddiehj@gmail.com

but no mushrooms. Does this mean mushrooms do not grow in Petri dishes? Of course not. It simply means that the conditions were not appropriate for growing mushrooms. The spores require a sterile environment to grow. Otherwise, other fungi and organisms may crowd out the effect of interest (mushrooms). Although we tried to follow the instructions, we somehow allowed the growing medium to become contaminated with unwanted organisms. Apparently, there was more to the technique of growing mushrooms than was conveyed in the instructions. We believe that something similar may occur with many experiments that fail to support solid hypotheses, as we will elaborate in this article. Humans are much more psychologically complex than mushrooms, particularly when it comes to our motivations and our emotional reactions, so the wide array of possible extraneous (unmeasured) variables that might contaminate psychological processes in a given experiment is almost infinite. Indeed, computer simulations of thousands of ideal replications have revealed that measurement error alone accounts for large differences in study results (Stanley & Spence, 2014).

Clearly, we did not know how to grow mushrooms, even though we had successfully grown vegetables for years (one of us grew up on a farm and grew acres of watermelons and other vegetables). Along these lines, expertise in specific methodologies likely predicts replication success, as evidence has suggested (Bench et al., 2017). Apparently, method sections lack details about setting up lab experiments that may be assumed knowledge by experts. If a method section is an experiment's "recipe," then this additional knowledge is akin to the expertise, gained from years of training and experience, that the chef brings to the kitchen to prepare the dish.

We do not always know what will cause a psychological experiment to work. However, we do have a few thoughts about what might make experiments more likely to work, based on our collective experiences. In our observations, it seems as though some replication attempts employ an input-output mindset when approaching experiments. That is, they seem to believe that simply presenting individuals with a certain input will evoke a certain output or response. However, we believe it is more appropriate to consider what occurs between input and output, and much of this "in-between" concerns the motivations, emotions, and cognitions of the individual—that is, the experience of participants as they take part in a study. As explained below, we suspect that many of the "contaminants" in an experiment involve the individual's motivations and emotions, which may be irrelevant to the experiment's input but likely influence the output.

### Guidelines for Conducting Lab Experiments

Method sections in journal articles, often by necessity, cannot describe every feature of a laboratory setup and design; yet these details—perhaps assumed to be shared knowledge within a field—can be crucial for conducting a valid test. Moreover, while preregistration (e.g., of hypotheses, sample sizes, and statistical analyses) is an important tool for promoting research transparency in both original research and replications, it is not intended to detail the countless minutiae of practices required to produce a strong experimental method. The goal of this article is to explicate many of these crucial yet often implicit details. In what follows, we describe examples of these details, offer guidelines for addressing them, and discuss their implications for replication research.

### Premanipulation Setup

#### *On the Way to the Testing Room*

First, we suggest that experimenters avoid signaling the purpose of their research to participants prior to the experiment, as this may induce the motivational state of demand (motivation to confirm the experimenter's hypothesis, or an oppositional demand, motivation to disconfirm it). One of us once worked at a university where a researcher had "Self-Regulation Lab" posted in big letters on the door to their testing room. We suspect some participants saw the sign as they entered the lab and then began to consider how they should engage in self-regulation (or not). Another researcher had "Culture and Emotions" posted in big letters on their lab door. Others displayed conference posters about their research in the hallway outside the experiment rooms. We suggest avoiding such displays, as they could induce demand, defiance with demand, or suspicion, all of which may motivate individuals away from acting "naturally."

#### *The Testing Room*

When setting up the lab room, we recommend that researchers avoid having mirrors, video cameras, etc. Even computer monitors with black backgrounds can show the participants' reflections. If observation of participants is necessary, then find ways of hiding the methods of observation (e.g., using a hidden camera). This recommendation is based on evidence that mirrors and video cameras can increase the negative affective state of self-consciousness, which then motivates behavior in predictable ways (Duval & Wicklund, 1972).

Related to this issue, a large registered experiment (Wagenmakers et al., 2016) failed to replicate the results of Strack et al. (1988), which found that individuals rate cartoons as more amusing when holding a pen in their teeth (similar to smiling) as compared to their lips (suppressing a smile). However, this replication experiment openly monitored participants with a video camera, whereas the original experiment did not. The inclusion of the camera might have led to the failure to replicate. Indeed, a subsequent experiment by Noah et al. (2018) manipulated the presence versus absence of the video camera and found that the camera's presence made a difference: When there was no camera as in the original Strack et al. (1988) experiment, the pen-in-the-teeth caused greater amusement ratings of the cartoons than the pen-in-the-lips, whereas when there was a camera as in the replication experiment (Wagenmakers et al., 2016), no condition differences emerged. Self-consciousness because of being observed may have suppressed the emotion of amusement.

It is also important to consider the physical size of the lab room in which the experiment is conducted. In one study that attempted to replicate a power pose effect (Garrison et al., 2016), a small lab room was used, and one of us had the intuition that taking up space with a power pose in such a small room may have caused participants to feel embarrassed. Perhaps as a consequence, the previously-observed power pose effect was not replicated in our study (meta-analyses suggest an effect; Gronau et al., 2017). A quick look at the majority of research articles in psychology journals reveals that most never specify the lab room's dimensions and set-up (e.g., lighting), even though these things may matter. For instance, one study manipulated room size (architectural space) and found

that larger rooms led to more self-disclosure in conversations (Okken et al., 2013). Another study revealed that manipulated ceiling height influenced cognitive processing; for example, a high versus low ceiling primed more freedom-related concepts and also caused more relational or commonality processing (Meyers-Levy & Zhu, 2007). When one of us once met with vision neuroscientists to help them find the source of noise in their electroencephalogram (EEG) data, he discovered their EEG setup had the participant sitting right next to the experimenter with no isolation from the noise/distraction/social dynamic whatsoever. The vision neuroscientists, who were not accustomed to thinking about the situational effects of psychological processes, assumed the noise was electrical and needed to be solved with a filter. They did not consider how environmental or emotional factors, such as discomfort from having a stranger in one's personal space, might influence brain activity.

### ***Cover Story***

In social psychology experiments that involve deception, a cover story is crucial to set the stage and disguise the purpose of the study (Aronson et al., 1998; E. Harmon-Jones et al., 2007). Even if an experiment does not involve deception, it is important to provide participants with a plausible reason for taking part in the experiment. Many individuals are inquisitive and may be motivated to discover what the experiment is about. Providing a plausible introduction will reduce suspicion and allow participants to focus on processing the important parts of the experiment.

### ***Psychological State at the Start of the Experiment***

In studies involving a motivational or emotional manipulation, it is important to ensure that participants are in a neutral state when they begin the study. In our preliminary studies examining the effects of low versus high approach-motivated positive affect on attentional processes (Gable & Harmon-Jones, 2008), we discovered that when we manipulated these different affective states between subjects, the manipulation checks revealed that the inductions had not influenced positive affect. Upon recalling how some participants may arrive at the lab with anxiety or in other emotional states, we began presenting participants with a neutral task or video before beginning the main study, so that participants could acclimate to the lab and start the study all within a similar neutral state. Once we made this change, our manipulations had the intended effects on manipulation checks.

### ***Experimenters/Confederates***

When conducting in-person experiments with multiple experimenters who each run a portion of the participants, researchers should train experimenters to be as consistent with each other as possible and to be consistent within themselves across experiment sessions. Differences in presentation styles, moods, dress, etc. could influence participant behavior, which would then create unwanted variance among participants. For example, differences in attractiveness or authoritativeness might affect participants' motivation to follow instructions and/or ingratiate themselves with the experimenter. Experimenters should also be instructed to dress appropriately. Imagine conducting a terror management experiment in which half of the participants are randomly assigned to think about death, but the experimenter wears a death-metal t-shirt and thus places

everyone in a mortality salience condition. Also, encourage experimenters not to drastically alter their appearance during the course of running one experiment (e.g., shaving, make-up, cologne, hair color). When some of us conduct EEG experiments, our experimenters wear lab coats to appear professional and similar to each other. Also, wearing a lab coat may place participants at ease around the "medical" looking EEG equipment knowing that they are working with a skilled technician. A lab coat is not always appropriate, however, as in some contexts it could induce unwanted emotional or motivational responses.

To our knowledge, the effect of experimenters on participants' responses has not been examined systematically. However, a few example studies have addressed this issue. Based on the idea that mortality salience would be more likely to increase worldview defense (i.e., ingroup favoritism and outgroup derogation) when participants were in an experiential rather than rational mode of processing, an experiment was conducted in which the experimenter was randomly assigned to act and dress in a more laid-back, informal manner or a more formal manner toward participants (Simon et al., 1997). In the informal condition, the experimenter wore a t-shirt, shorts, and sandals, and spoke informally using his hands in expressive ways while saying "you guys" and "okay" often. In the formal condition, he wore black glasses and a lab coat over long trousers, and spoke in formal ways while sitting stiffly at a desk. One inspiration for this manipulation of the experimenter's dress and behavior came from the observation that mortality salience effects were more likely to occur when experimenters appeared "informal, laid-back, and comfortable" (Simon et al., 1997, p. 1134). Results from this experiment revealed that as compared to the formal experimenter, the informal experimenter caused participants to write about their mortality in more experiential ways, which then caused them to engage in more worldview defense.

Another study examined the effect of experimenter characteristics on participants' responses by having male participants rate female experimenters on attractiveness (Wacker et al., 2013). This study assessed two biological measures associated with approach motivation as well as self-reported trait approach motivation. Results revealed that these approach motivation measures were positively correlated with each other primarily when male participants rated their female experimenter as attractive, suggesting that this approach motivation context, which is a context typically considered incidental to a manipulation/design, increased the correlations. These results suggest that the inclusion of measures of participants' ratings of experimenters' attractiveness and other relevant variables should be incorporated depending on the variables being manipulated and measured.

Experimenters should be trained to be calm and professional, to avoid inducing unwanted motivations and emotions. One of us once worked with a student researcher whose study included a measure of participants' self-reported baseline emotions. We noticed that participants' baseline negative affect was higher than was typically produced by some of our negative affect inductions. Unobtrusive observation of the student researcher interacting with some participants revealed that he came across as very anxious. He talked fast, did not look at participants, and moved in fast, herky-jerky ways. If experimenters induce a lot of negative (or positive) emotion in participants at the beginning of the study, this effect is likely to weaken and interfere with subsequent emotion or motivation manipulations.

Another one of us once conducted a study on interracial interaction anxiety with a Black versus White confederate as one of the

independent variables. As it turned out, the Black confederate was friendly and socially skilled, whereas the White confederate was shy and awkward, and this difference overwhelmed any effects of race on participants' responses to the confederates. Experiments designed to test these types of hypotheses should ideally use several different experimenters from each race so that the effects can be attributed to race rather than personality differences between two individual experimenters.

As a best practice, we recommend that, when multiple experimenters (confederates) are involved, they should be assigned to run equal numbers of participants from each condition (i.e., to avoid an experimenter confound). If one experimenter runs mostly participants from condition A while another runs participants mostly from condition B, then this presents a confound, as differences between conditions could be because of an experimenter effect rather than the manipulation. In other words, block randomization for each experimenter should be used, so that each experimenter runs an equal number of participants in each condition. In addition, it would be wise to include an experimenter variable in preliminary statistical analyses to check for differences in the dependent variables and manipulation checks by experimenter. If a main effect of experimenter is found, it does not necessarily present a problem, as long as each experimenter runs an equal number of participants in each condition. On the other hand, if interaction effects of the manipulation by experimenter are found, this is more problematic. However, in practice, a manipulation by experimenter interaction may be difficult to detect with the sample sizes typically used in most lab experiments. In this case, as a rough check on the influence of the experimenter on the manipulation effect, especially when null effects occur, researchers could check whether the manipulation effect is present (in the predicted direction and of a relatively similar magnitude) for each experimenter. Along these lines, it would also be important to note how many participants each experimenter ran.

To reduce the self-presentational motivations of participants, some researchers have only certain types of experimenters run certain types of participants. For example, in research on White prejudice toward Blacks, researchers will have only White experimenters (Amodio et al., 2008). Research on helping behavior often has experimenter gender match participant gender (Batson et al., 1997).

Finally, experimenters should remain blind to the condition. Typically, this is more of an issue in between-subjects designs than in within-subjects designs. Although the importance of being blind to condition was emphasized long ago by Rosenthal's excellent work (1966), we are surprised to learn how many researchers do not follow this simple and important guideline. In discussions with researchers who do not follow this guideline, we have learned that they simply do not think it is important and assume that being "unblind" will not taint their results. To this, we always remind them of Rosenthal's classic finding that unblind experimenters can even unwittingly influence the behavior of laboratory rats (see meta-analysis by Rosenthal & Rubin, 1978), likely because experimenters are subconsciously motivated to confirm the predicted effects.

## Independent Variables

Generally speaking, the independent variable should be carefully constructed so that it is as strong as possible. In psychology research, we are typically interested in testing theoretical ideas. However, these theoretical variables can often be operationalized in a variety

of different ways. As such, researchers should begin by establishing the construct validity of the manipulations by performing studies that determine whether the manipulations are doing what they are predicted to do, as has been done with other constructs (Chester & Lasko, 2021; Cronbach & Meehl, 1955).

## Strong Manipulations

In his graduate seminar, Jack Brehm—known for his foundational work on several motivational issues (e.g., Brehm, 1956; Brehm et al., 1983)—informed students that the independent variable should be like a sledgehammer, within ethical constraints. That is, it should be as strong as possible from a methodological standpoint. Researchers who study nonhuman animals are aware of this. For instance, in studies where responses are motivated by food reward, they food-restrict their subjects to 85% of their normal weight to make food rewards more appealing (Goedhoop et al., 2023). Creating an incredibly strong approach motivation by bringing starving folks into the lab and presenting them with opportunities for delicious food would likely be a sledgehammer. However, sledgehammer manipulations are not always practically or ethically feasible.

Moreover, using a sledgehammer manipulation might even present methodological problems if the researcher were interested in measuring some computer responses unrelated to eating because participants might be too distracted by their motivation to acquire food to attend to the computer task. In this case, a sledgehammer manipulation might be so strong that it could overwhelm other interacting manipulations and weaken predicted interaction effects.

The creation of an independent variable also needs to consider other variables in the design. In general, when more than one variable is being manipulated, the strength of the two (or more) manipulations should be as equal as possible. Otherwise, the effect of the stronger manipulation may overwhelm the weaker one, obscuring its effect. Researchers interested in testing interactions must balance the strength of the orthogonal variables.

## Manipulation Checks

One way to ensure a manipulation is working as expected is with manipulation checks. As Sigall and Mills (1998) noted, the term manipulation check can be used to mean an assessment of the conceptual independent variable or a check on whether the differences between the conditions were perceived (e.g., recall measures). In this article, we are referring to the former meaning. Ideally, the manipulation check should measure the psychological process the manipulation was designed to influence. Often this is done with self-report assessments, which, despite some limitations (as we discuss later), can be quite useful in discovering whether the manipulation is working. Ideally, manipulation checks should reveal effect sizes that are quite large (Cohen's  $d > 2.0$ ), so that there is almost no overlap between the conditions in terms of responses to it. For example, with an anger-inducing manipulation, folks in the anger condition should score between 3 and 5 on a 5-point scale measuring the intensity of anger, whereas folks in the neutral condition should score between 1 and 2 on this same scale. Researchers should be aware, however, that self-reported manipulation checks may be influenced by experimenter demand and thus not provide perfect evidence of whether the independent variable was effectively manipulated (Sigall & Mills, 1998).



In some cases, however, manipulation checks are not easily obtained because the construct being manipulated is not verbalizable or because research on the theory or hypothesis has not developed to the point of knowing how best to check the manipulation. For example, the manipulated effect of interest may operate without awareness, and thus should not be observable in a self-reported manipulation check (the opposite may occur as well, where a psychological state may be easily measured with self-reports [e.g., nostalgia] but not captured with current physiological assessments).

Research on terror management theory provides a good example of this. Originally, terror management research did not measure death-construct accessibility, but once it was discovered that death-construct accessibility was immediately suppressed after mortality salience and then later increased (Greenberg et al., 1994), this accessibility measure could be used as a manipulation check of this suppression-rebound process that was incorporated into the theory.

In addition, a manipulation could influence several processes, but only one (or a subset) of these processes may be the mechanism through which it influences the dependent variable of interest. That is, it is possible that a manipulation check measures a different variable than the process directly driving the effect on the dependent variable, as when the manipulation check is self-report, but the mechanism is nonconscious.

However, there may be situations in which the inclusion of a manipulation check is not feasible, as noted by Sigall and Mills (1998). A manipulation check might reveal the purpose of the experiment and thus increase suspicion from participants. In addition, the inclusion of a manipulation check might undermine the effectiveness of the manipulation because it causes participants to doubt the truthfulness of the instructions that were part of the manipulation. Last, if the manipulation check is presented to participants prior to the dependent variable, it could influence responses to the dependent variable, and if placed after the dependent variable, it may not be as sensitive because of the time delay between the manipulation and the check on it.

Following Sigall and Mills (1998), the inclusion of manipulation checks is not necessary for experiments. As they noted (p. 226), “their inclusion does not solve the fundamental problem in experiments of eliminating plausible alternative explanations for the effect of the experimental treatment on the dependent measure.” However, despite the limitations of manipulation checks, a large effect on these measures is a positive sign for the experimenter that the manipulation is effective.

### ***Manipulations May Depend on Context***

The manipulation of the precise psychological construct may not be as simple as having your participants do exactly what another researcher’s participants did in a different time and place. You should always pretest the manipulation and manipulation check with your specific sample (Stroebe & Strack, 2014). Things may differ because of time and place, as nicely illustrated in a study that coded the context sensitivity of each study of the 100 studies used in the 2015 Reproducibility Project (Van Bavel et al., 2016). Van Bavel et al. (2016, p. 6455) defined context sensitivity “as differing in time (e.g., prerecession vs. postrecession), culture (e.g., individualistic vs. collectivistic culture), location (e.g., rural vs. urban setting), or population (e.g., a racially diverse population vs. a

predominantly White population).” Results from this study found that studies that concerned topics that were more context-sensitive were less likely to be replicated.

As Wright et al. (2019) noted when considering ego depletion, task difficulty will interact with the ability of participants and the importance of task success to influence fatigue. Thus, holding the motivation for success constant, if one sample has lower ability on a task, then what may appear to be an easy task would be difficult for them and thus evoke greater effort and possibly later fatigue; a more difficult task would not evoke much effort or later fatigue because it is perceived as impossible or not worth the effort for the lower ability sample. If another sample has higher ability, then an easy task would not evoke much effort and not much fatigue, but a more difficult task would.

### ***Confounds***

A manipulation should not have confounds or features that potentially lead to response biases, such as demand characteristics. For example, in recent studies (E. Harmon-Jones et al., 2024), we wanted to use some texts that had been published as part of a well-cited set of studies (van Prooijen et al., 2022). However, upon consideration of the texts, we saw two sentences that may have induced experimenter demand (Orne, 1962) and also added a confound. This research manipulated texts to be low versus high in conspiratorial information. In the low conspiratorial condition, the text had this sentence, “There is little reason to question the official reading of this event.” In the high conspiratorial condition, the text had this sentence, “There is, however, ample reason to question the official reading of this event.” In our opinion, the sentences basically tell the participants how to respond to the subsequent questions about belief in a conspiracy theory (i.e., experimenter demand). Moreover, these two sentences created a confound between the two conditions; that is, in addition to the manipulation of details about whether Jeffery Epstein, a wealthy convicted sex offender, was murdered or committed suicide in his jail cell, participants in one condition were told to not question “the official reading of this event,” whereas participants in the other condition were told to question it. We removed these sentences, and we also did not replicate the past research’s mediational evidence. Perhaps the removal of these sentences was the cause of this failure to replicate. Either way, those sentences should have never been included in the texts because of the experimenter demand and confound.

### ***Dependent Variables***

The dependent variable should be as sensitive as possible to detect effects, just as the independent variables should be as strong as feasible. In general, and especially when considering motivational variables, self-reports are often less sensitive than measures of psychophysiological processes. The work on Brehm’s (Brehm et al., 1983) influential motivational intensity theory has been primarily tested with cardiovascular measures (Gendolla et al., 2012). The theory could be tested by asking participants how much effort they plan to expend on upcoming tasks that vary in difficulty etc., but these self-reports have been found to be insensitive (Wright et al., 1990). Perhaps the lack of sensitivity in effort reports is because of the fact that many participants are motivated to appear like good participants and over-report their effort, regardless of task

difficulty, or participants may not be aware of this process. However, another type of self-report, which is more indirect, has been found to be sensitive to manipulations of variables derived from Brehm's theory. That is, several experiments have found that self-reported goal value varies directly with actual motivation (Brehm et al., 1983).

Even when using self-report measures, steps can be taken to ensure that the measures are more sensitive and less subject to various biases and noise. For example, manipulations such as mortality salience and social rejection had been found not to influence self-reported affect in past research (e.g., Greenberg et al., 1995; Twenge et al., 2003). However, these past studies measured emotion using the Positive Affect and Negative Affect Schedule (PANAS; Watson et al., 1988), which is intended to measure general positive and negative affect. Our research suggested that measuring general affect may be less sensitive than measuring the specific discrete emotion that the manipulation would be expected to influence (C. Harmon-Jones et al., 2016). For example, when we administered the social rejection manipulation, we replicated the past null effects when participants completed the PANAS. However, we found that social rejection significantly increased reported sadness (C. Harmon-Jones et al., 2016). Similarly, the effect of mortality salience on PANAS negative affect was small, but this manipulation produced a large increase in reported fear and anxiety (C. Harmon-Jones et al., 2016).

In these same studies, we found that the instructions given to participants may have important effects on the sensitivity of self-report measures of emotion (C. Harmon-Jones et al., 2016). The effects of emotion manipulations were larger when participants were asked to report the emotions they had felt during the manipulation (while viewing emotional photos or completing writing tasks), compared to when participants were asked to report the emotions they were feeling right now (after the manipulation was over). This suggests that when researchers attempt to replicate a study, it is important to be aware of the specific instructions used, because these may influence the strength of an effect and thus the likelihood of a successful replication.

While on the topic of self-reported emotions, some scientists recommend including more emotion words to assess a construct because the inclusion of more items increases reliability. However, the inclusion of more emotion words may move the measure away from the construct of primary interest, thus reducing construct validity and sensitivity. An example of this occurred in a study on guilt wherein a single item best captured the construct of interest because seemingly related terms (e.g., self-disappointment and self-dissatisfaction) did not quite capture it (Amodio et al., 2007).

Relatedly, the order in which events occur in some studies may influence the results. In one article, we replicated a previously observed negative correlation between analytic thinking and religious belief—but only when analytic thinking was measured (hence, primed) first. When the two variables were measured the other way around, their correlation disappeared (Finley et al., 2015).

### ***Pilot Testing***

With regard to all of these recommendations, we advocate careful pretesting (pilot testing) prior to starting the study. Pretesting can accomplish several goals: (a) check the validity of the manipulation, (b) discover whether the manipulation check yields a large effect, (c) test the sensitivity of the dependent variable, and (d) estimate

the effect size of the independent-dependent variable pairing (which can be used for a priori power analyses for subsequent studies). All of the issues are important to pretest and they may require separate pretests. Pretesting can also be used to discover whether participants are suspicious of any aspect of the study or just think any of it is weird. To discover these latter issues, a careful one-on-one debriefing with participants is necessary in which the researcher sits down with the participant and slowly asks them questions about these important aspects. See E. Harmon-Jones et al. (2007) for examples.

Although these recommendations are based primarily on our experience as researchers on specific topics pertaining to motivation and emotion and thus may not readily generalize to research on other topics, we believe they provide an important lesson: When starting research on a topic that is new to you, seek the advice of those who already have experience in that area. In our careers, we have gained much assistance in this way. A summary of these guidelines is displayed in Table 1.

### **Applying These Guidelines to Replications**

Replication is a crucial part of a strong psychological science (e.g., Crandall & Sherman, 2016; Popper, 1959). Although direct replication is important, conceptual replications may be even more important for theory testing, as they operationalize the variables of interest in different ways. When conceptual replications are successful, they lend more robust support to the theory relative to the direct replication of a specific paradigm that was previously tested.

Moreover, many theories and hypotheses are often extended by replicating a basic effect and then adding moderator variables. Such studies are an important source of successful replication research that, in the past, may not have received notice as such.

### **Selecting Studies to Replicate**

Following others (Albarracín & Dai, 2021), we recommend that researchers who are interested in reproducing a theoretical finding select studies for replication that have a reasonable chance of being replicated. If the goal is to determine support for a theory, and not for a particular procedure or operationalization, then researchers should choose approaches with the strongest methods. If a study is selected because it is suspected to involve a weak procedure, then the replication is less theoretically informative because it intentionally provides a weak test of the theory. However, if the goal is to demonstrate that a procedure is weak and nonreplicable—which appears to be a frequent goal (E. Harmon-Jones & Harmon-Jones, 2024)—then it should be acknowledged that the replication (or nonreplication) is more informative regarding the specific operationalization than the theory. Beyond limiting its usefulness for theory testing, this approach is also limited by experimenter demand, such that the expectation of nonreplication might influence the behaviors of the researchers and the ultimate outcome. As Bornstein (1990, p. 76) noted,

consider a situation in which the experimenter is highly motivated to retain the null hypothesis (i.e., a situation in which the experimenter wants to not replicate a finding). In this situation, there are many things that the experimenter can do—consciously or unconsciously—to obtain the desired result. Generally speaking, anything that introduces “noise” into the experiment will tend to obscure any experimental effects or group differences that actually exist, thereby increasing the likelihood that the null hypothesis will be retained.

**Table 1***Summary Guidelines for Conducting Lab Experiments in Motivation and Emotion*


---

Premanipulation setup
<ul style="list-style-type: none"> <li>• Consider the physical environment on the way to the lab room</li> <li>• Consider the physical environment of the lab room (e.g., cameras, mirror, room size, social setting)</li> <li>• Create a plausible cover story</li> <li>• Be mindful of the psychological state of participants at the start of the experiment</li> </ul>
Experimenters/confederates
<ul style="list-style-type: none"> <li>• Ensure consistency between multiple experimenters/confederates</li> <li>• Compare the effect of multiple experimenters statistically</li> <li>• Keep appearance/clothing consistent</li> <li>• Be cognizant of informal/formal style of interacting</li> <li>• Measure differences in physical attractiveness</li> <li>• Track emotion and personality variables of experimenters</li> <li>• Keep blind to condition</li> </ul>
IVs
<ul style="list-style-type: none"> <li>• Use a strong IV—aim for a psychological sledgehammer</li> <li>• Be careful that IV is not so strong that participants ignore DV</li> <li>• Ensure multiple IVs are similar in strength</li> <li>• Check the effectiveness of manipulation and appreciate its complexity</li> <li>• Pretest to ensure IVs are the same psychologically in different times and places</li> <li>• Avoid experimenter demand</li> <li>• Avoid confounds</li> </ul>
DVs
<ul style="list-style-type: none"> <li>• Create sensitive DVs</li> <li>• Consider challenges with self-reported effort</li> <li>• Consider challenges with self-reported emotion</li> <li>• Be aware that the order of presentation of variables might influence results</li> <li>• Pretest</li> </ul>
Applying guidelines to replications
<ul style="list-style-type: none"> <li>• Consider the benefits of conceptual replications for theory testing</li> <li>• Select studies for replication that have a reasonable chance of being replicated</li> <li>• Note that expectations of nonreplication influence researchers and outcomes</li> <li>• Use the exact original method when replicating a specific effect</li> <li>• Test correlations that logically follow from experimental hypotheses if possible</li> <li>• Ensure participating labs in projects follow instructions of the study coordinators</li> </ul>
Data management
<ul style="list-style-type: none"> <li>• Take steps to avoid errors in data before posting</li> <li>• Share data files in widely accessible formats (e.g., *.csv)</li> <li>• Use variable names that are easy to understand</li> <li>• Provide a coding sheet that explains the variable labels</li> <li>• Have a collaborator/colleague rerun all analyses prior to posting the data file online</li> <li>• Convey handling of excluded participants in shared data files</li> </ul>
Statistical issues
<ul style="list-style-type: none"> <li>• Be aware of difficulties with falsifying hypotheses</li> <li>• Know problems with null hypothesis significance testing</li> <li>• Consider promises and pitfalls of meta-analyses</li> <li>• Know problems with effect sizes</li> </ul>

---

*Note.* IVs = independent variables; DVs = dependent variables.

Above, we have noted a number of ways to avoid a successful replication, such as by using weak manipulations, using manipulations that differ in strength when looking for an interaction, using insensitive dependent measures, introducing noise variance by having multiple experimenters, and drawing from a population that is very different from the original sample. A solid attempt to show that an effect does not exist would use the strongest feasible manipulation, most sensitive dependent variable, and statistical tests intended to disconfirm the hypothesis (see below).

Relatedly, we recommend that replicators do not focus on one experiment in a program of research, but rather consider the entire body of related experiments and contact the original researchers to ask for recommendations for testing the strongest, most important experiment from the body of research. Often, the original researchers will have helpful recommendations about which versions work consistently and which versions are a bit fragile.

When conducting a study with the primary purpose of testing whether an effect replicates, ensure that the method is as close as possible to the original method in terms of technical details such as stimulus timing. A good example of a replication along these lines is one conducted by Domachowska et al. (2016), who replicated the effect of high approach motivation decreasing the breadth of attention using the same stimuli and computer program as used in the original experiment (Gable & Harmon-Jones, 2008). These researchers went on to examine different stimuli as well, noting the importance of cultural context (German vs. United States participants) in their stimuli choice.

### Validity Problems in Replications

However, examples of replication experiments that did not follow the original experiment also exist. For example, the Open Science

Collaboration (2015) replicated a study (Amodio et al., 2008) that used a speeded sequential priming task to examine White participants' motivations to respond without prejudice toward Black people. However, several critical aspects of the design were not followed in the replication. For example, the wrong equipment was used for the task, resulting in much slower presentations of the main priming task (as well as the inability to produce a well-established flankers effect in a control task). In addition, the sample included only 39% White subjects, with others being from racial/ethnic minority groups (Asian, Latino, and Black), which was not appropriate for testing a question about White people's prejudice toward Black people (Amodio et al., 2025). Given these and other problems with replication, this failed attempt should have been described as inconclusive rather than as a nonreplication (Giner-Sorolla et al., 2018).

In another replication study, of E. Harmon-Jones et al. (2008), participants rated decision alternatives prior to making a difficult decision, and then they rerated these alternatives after the decision (using the design created by Brehm, 1956). This research paradigm is designed to test the dissonance theory prediction that following a difficult decision, the chosen alternative will be rated more positively than the rejected alternative (spreading of alternatives). In this free-choice/difficult decision experiment, the decision alternatives should be rated approximately equal in attractiveness prior to the decision. However, in this replication experiment, the decision alternatives were not rated equally in the predecision phase and this was particularly the case in the condition in which the most predecision to postdecision attitude change was predicted to occur; this introduced a confound. If the different decision alternatives are not rated almost equally prior to the decision, then there will be less dissonance and participants will be less motivated to change their attitudes. Moreover, statistically speaking, this predecision spreading of alternatives should make it more difficult to see a significant increase in postdecision spreading of alternatives. One potential reason the predecision ratings were so different is because of the difference in the quality of the replication study, as compared to the original experiment. The organizing replication group gave the study lead of this replication (one of the current article's co-authors) 4 months to gather study materials, ethics approval, train experimenters, and run nearly 80 participants (one at a time as a result of the EEG measure). Thus, the replication had to use four different experimenters. In contrast, the original study only had one experimenter. The large number of experimenters and rushed replication could have contributed to the problems with this replication.

When replicating an experimental design that comes from a theory with obvious hypotheses regarding the relationship between variables, we recommend testing the associated correlational hypotheses if possible (e.g., the manipulation check on the independent variable is used as a continuous predictor). As an example, consider a replication of the induced compliance effect associated with cognitive dissonance theory (Vaidis et al., 2024). This research tested whether making a counter-attitudinal statement about an issue would cause attitude change about the issue in the direction of the statement, and whether perceived choice would moderate the effect, as has been observed in hundreds of past experiments (E. Harmon-Jones, 2019). However, the replication research did not test or report the correlation of self-rated perceived choice and attitudes, even though dissonance theory would predict that greater perceived choice would relate to more attitude change following counter-attitudinal advocacy. Subsequent

analyses by other researchers revealed that perceived choice did relate to attitude change (Lishner, 2024; Pauer et al., 2024), as would be expected (E. Harmon-Jones & Harmon-Jones, 2024), thus replicating the original theoretical prediction.

### Mistakes Will Be Made

It is probably all too common to have small but consequential mistakes in single-investigator, single-lab studies, and the likelihood of such screw-ups only multiplies when multiple independent labs (each varying in experience, interest, etc.) collaborate to test an effect. In a large multilab replication project on the ego depletion effect (Vohs et al., 2021), one lab started data collection before anybody was supposed to start, with some of the loose ends yet to be tied up. Then, that lab turned in one of the largest effects in the direction opposite to the prediction. We cannot say for sure that their failure to follow directions mattered for the results, but the point is that participating labs do not always follow the well-intentioned instructions of the study coordinators (and these failures may go unnoticed).

One of us is currently involved in another multilab replication attempt regarding Wegner's classic thought suppression findings (e.g., Wegner & Erber, 1992). As of this writing, the project is over 1 year past due, with participating laboratories confused by and in disagreement about how to score the primary dependent measure. This confusion persists even though the scoring method was reviewed and approved prior to data collection as a registered replication report.

### Data Management

Flaws in data processing can lead to failures to test hypotheses and replicate past results. One of us recently witnessed a flaw in data shared from a published article (Vaidis et al., 2024). This article was a many-labs replication of the induced compliance paradigm, described above. We planned to conduct some new analyses of the shared data. When we first downloaded the data from Open Science Framework on March 5, 2024, we observed that the mean preexisting attitude rating was much larger in the shared data set ( $M \sim 2.8$ ) than had been ( $M \sim 1.6$ ) reported in Vaidis et al. (2024). The mean that we discovered in the shared data set would indicate that, for many participants, the "counter-attitudinal" essay they wrote may not have been counter to their attitude. Without a counter-attitudinal essay, dissonance should not occur and thus there should be no motivation for attitude change. These differences between means occurred across the entire sample and within each condition. We contacted the editor about this discrepancy, and he contacted the authors, who replied and only said that they had reuploaded the correct data files. Another similar instance of researchers posting data with errors was uncovered in a multilab project replicating the effect of mortality salience on death-thought accessibility (Rife et al., in press). The original version of the article reported a partial failure to replicate but the reanalysis of the data with the errors corrected produced results consistent with the original experiment (D. Lishner, personal communication, December 18, 2024). Fortunately, the errors were uncovered by researchers not involved in the replication project prior to publication, and the authors of the project acknowledged David Lishner and Christopher Groves for finding the errors.



Following these examples, we encourage researchers to triple-check their data. We have found that data analyses are easier to replicate when data sheets are set up in a way that anyone with statistics experience could replicate the results. It is also good practice to share all data and analysis scripts. Indeed, many funding agencies and journals now require data sharing. Moreover, because many software packages are not free to use, and not all researchers have these packages and/or know how to use them, it is best practice to share data as \*.csv files with an explanation of variables and so on in an associated document, so that interested parties can reanalyze the data with any software package. Once all of this is done, ask a collaborator or colleague to use these files to rerun the analyses to make sure they replicate what is reported in your article.

When several individuals collaborate on one research project, the data file can be shared with several individuals, and then there might be multiple versions of the data file, with more variables, fewer variables, different variable labels, different excluded/included participants, etc. To avoid problems, it is ideal to have one person control the data file (locked and edited by this one person). In terms of participants who are excluded, we make the exclusions prior to data analysis, mark these exclusions in one version of the data file, and save it with a name that indicates such. Then, copy that file and physically remove the excluded participants (e.g., “\_\_nonvalid\_subjects”), and save the file with a name that indicates such (e.g., “\_\_valid\_subjects”). This approach maintains transparency, by making both versions available, but serves to avoid mistakes in the reanalysis of reported results.

## Statistical Issues

One of us recently heard someone with a PhD in psychology say something like, “Failing to replicate an effect at  $p < .05$  is important because it fulfills the goal of falsification of theories.” There are several problems with this statement.

It is incredibly difficult to falsify theories, especially broad theories like cognitive dissonance theory. Such a theory has been tested in multiple ways, and one nonreplication of a single experiment does not mean the theory, or all other examples of the effect, are wrong.

One experiment with a  $p > .05$  is not going to do it. Reviewers will (should) ask questions about the operationalization of the independent and dependent variables as well as extraneous variables. More broadly, good theories generate multiple hypotheses, so it would take a considerable amount of work to falsify a theory. Falsification is too lofty a goal for a single study.

Moreover, null hypothesis statistical tests are not designed to falsify. They involve convoluted logic that many researchers do not quite understand. As Gigerenzer et al. (2004, p. 3) noted, “A  $p$ -value is the probability of the observed data (or of more extreme data points), given that the null hypothesis  $H_0$  is true ...” If the probability is small, we then conclude that the null hypothesis is false. The converse, “if the probability is not small, we conclude that the null hypothesis is true,” is not appropriate. For example, with  $p < .10$ , the probability of the observed data is less than 10%, which is still small even if not quite as small as 5%.

To attempt to prove the null or show no effect exists, the commonly used null hypothesis statistical test cannot be used, as several have discussed (Gallistel, 2009). Rather, one needs to use Bayesian statistics, which can give a probability of the null being true or false. However, as noted throughout this article, several factors can cause “false” null effects.

In addition, one individual experiment provides little information, as Cumming (2014) illustrated in his work on the dance of the  $p$  values. Meta-analyses are needed to fully understand an effect, but meta-analyses often lump good with bad methods, and thus can be misleading. In addition, because meta-analyses require multiple studies testing the same effect, they are usually unable to address nuanced interactive predictions because so few of these types of tests are conducted.

Psychological scientists often recommend reporting effect sizes, and some label effect sizes of certain magnitudes as small, medium, or large. As Funder and Ozer (2019) explained, these labels are misleading and depend on the type of study being conducted. Relatedly, some obvious and relatively simple methods produce larger effect sizes (e.g., Stroop, 1935) compared to less obvious but more interesting methods (e.g., Festinger & Carlsmith, 1959).

A good portion of research on motivation is aimed at testing theory in laboratory experiments to ultimately understand what the mind-brain can do. This type of basic research stands in contrast to research aimed at testing applications. With basic research, attempting to determine effect sizes might be a futile exercise, as Baumeister (2020, p. 803) explained:

The artificial environment of a psychological laboratory experiment offers an excellent method for testing whether a causal relationship exists,—but it is mostly useless for predicting the size and power of such effects in normal life. In comparison with effects out in the world, laboratory effects are often artificially large, because the laboratory situation is set up precisely to capture this effect, with extraneous factors screened out. Equally problematic, laboratory effects are often artificially small, given practical and ethical constraints that make laboratory situations watered-down echoes of what happens in life. Furthermore, in many cases the very notion of a true effect size (as if it were constant across different manipulations and dependent variables) is absurd.

## Conclusion

We would like to stress that as psychological scientists, we are tasked with finding the truth about psychological processes. Our aim should be to produce strong science, which includes replication with strong methods. We hope that this article will aid researchers as they conduct and interpret replication research.

## References

- Albarracín, D., & Dai, W. (2021). Priming effects on behavior and priming behavioral concepts: A commentary on Sherman and Rivers (2020). *Psychological Inquiry*, 32(1), 24–28. <https://doi.org/10.1080/1047840X.2021.1889319>
- Amodio, D. M., Devine, P. G., & Harmon-Jones, E. (2007). A dynamic model of guilt: Implications for motivation and self-regulation in the context of prejudice. *Psychological Science*, 18(6), 524–530. <https://doi.org/10.1111/j.1467-9280.2007.01933.x>
- Amodio, D. M., Devine, P. G., & Harmon-Jones, E. (2008). Individual differences in the regulation of intergroup bias: The role of conflict monitoring and neural signals for control. *Journal of Personality and Social Psychology*, 94(1), 60–74. <https://doi.org/10.1037/0022-3514.94.1.60>
- Amodio, D. M., Devine, P. G., & Harmon-Jones, E. (2025, February 25). *An invalid test of the original experiment: Comment on the Reproducibility Project's attempt to replicate Amodio, Devine, & Harmon-Jones (2008)*. [https://doi.org/10.31234/osf.io/ey79r\\_v1](https://doi.org/10.31234/osf.io/ey79r_v1)

- Aronson, E., Wilson, T. D., & Brewer, M. B. (1998). Experimentation in social psychology. In D. T. Gilbert, S. T. Fiske, & G. Lindzey (Eds.), *The handbook of social psychology* (4th ed., pp. 99–142). McGraw-Hill.
- Batson, C. D., Polycarpou, M. P., Harmon-Jones, E., Imhoff, H. J., Mitchener, E. C., Bednar, L. L., Klein, T. R., & Highberger, L. (1997). Empathy and attitudes: Can feeling for a member of a stigmatized group improve feelings toward the group? *Journal of Personality and Social Psychology*, 72(1), 105–118. <https://doi.org/10.1037/0022-3514.72.1.105>
- Baumeister, R. (2020). Do effect sizes in psychology laboratory experiments mean anything in reality?. *Psychology: Journal of the Higher School of Economics*, 17(4), 803–811. <https://doi.org/10.17323/1813-8918-2020-4-803-811>
- Bench, S. W., Rivera, G. N., Schlegel, R. J., Hicks, J. A., & Lench, H. C. (2017). Does expertise matter in replication? An examination of the reproducibility project: Psychology. *Journal of Experimental Social Psychology*, 68, 181–184. <https://doi.org/10.1016/j.jesp.2016.07.003>
- Bornstein, R. F. (1990). Publication politics, experimenter bias and the replication process in social science research. *Journal of Social Behavior and Personality*, 5(4), 71–81.
- Brehm, J. W. (1956). Postdecision changes in the desirability of alternatives. *The Journal of Abnormal and Social Psychology*, 52(3), 384–389. <https://doi.org/10.1037/h0041006>
- Brehm, J. W., Wright, R. A., Solomon, S., Silka, L., & Greenberg, J. (1983). Perceived difficulty, energization, and the magnitude of goal valence. *Journal of Experimental Social Psychology*, 19(1), 21–48. [https://doi.org/10.1016/0022-1031\(83\)90003-3](https://doi.org/10.1016/0022-1031(83)90003-3)
- Chester, D. S., & Lasko, E. N. (2021). Construct validation of experimental manipulations in social psychology: Current practices and recommendations for the future. *Perspectives on Psychological Science*, 16(2), 377–395. <https://doi.org/10.1177/1745691620950684>
- Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *The Journal of Abnormal and Social Psychology*, 65(3), 145–153. <https://doi.org/10.1037/h0045186>
- Crandall, C. S., & Sherman, J. W. (2016). On the scientific superiority of conceptual replications for scientific progress. *Journal of Experimental Social Psychology*, 66, 93–99. <https://doi.org/10.1016/j.jesp.2015.10.002>
- Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52(4), 281–302. <https://doi.org/10.1037/h0040957>
- Cumming, G. (2013). *Understanding the new statistics: effect sizes, confidence intervals, and meta-analysis*. Routledge. <https://doi.org/10.4324/9780203807002>
- Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*, 25(1), 7–29. <https://doi.org/10.1177/0956797613504966>
- Domachowska, I., Heitmann, C., Deutsch, R., Goschke, T., Scherbaum, S., & Bolte, A. (2016). Approach-motivated positive affect reduces breadth of attention: Registered replication report of Gable and Harmon-Jones (2008). *Journal of Experimental Social Psychology*, 67, 50–56. <https://doi.org/10.1016/j.jesp.2015.09.003>
- Duval, S., & Wicklund, R. A. (1972). *A theory of objective self awareness* (pp. x, 238). Academic Press.
- Festinger, L., & Carlsmith, J. M. (1959). Cognitive consequences of forced compliance. *The Journal of Abnormal and Social Psychology*, 58(2), 203–210. <https://doi.org/10.1037/h0041593>
- Finley, A. J., Tang, D., & Schmeichel, B. J. (2015). Revisiting the relationship between individual differences in analytic thinking and religious belief: Evidence that measurement order moderates their inverse correlation. *PLoS ONE*, 10(9), Article e0138922. <https://doi.org/10.1371/journal.pone.0138922>
- Funder, D. C., & Ozer, D. J. (2019). Evaluating effect size in psychological research: Sense and nonsense. *Advances in Methods and Practices in Psychological Science*, 2(2), 156–168. <https://doi.org/10.1177/251524519847202>
- Gable, P. A., & Harmon-Jones, E. (2008). Approach-motivated positive affect reduces breadth of attention. *Psychological Science*, 19(5), 476–482. <https://doi.org/10.1111/j.1467-9280.2008.02112.x>
- Gallistel, C. R. (2009). The importance of proving the null. *Psychological Review*, 116(2), 439–453. <https://doi.org/10.1037/a0015251>
- Garrison, K. E., Tang, D., & Schmeichel, B. J. (2016). Embodying power: A preregistered replication and extension of the power pose effect. *Social Psychological and Personality Science*, 7(7), 623–630. <https://doi.org/10.1177/1948550616652209>
- Gendolla, G. H., Wright, R. A., & Richter, M. (2012). Effort intensity: Some insights from the cardiovascular system. In R. M. Ryan (Ed.), *The Oxford handbook of human motivation* (pp. 420–438). Oxford University Press.
- Gigerenzer, G., Krauss, S., & Vitouch, O. (2004). The null ritual: What you always wanted to know about significance testing but were afraid to ask. In D. Kaplan (Ed.), *The Sage handbook of quantitative methodology for the social sciences* (pp. 391–408). Sage.
- Giner-Sorolla, R., Amodio, D. M., & van Kleef, G. A. (2018). Three strong moves to improve research and replications alike. *Behavioral and Brain Sciences*, 41, Article e130. <https://doi.org/10.1017/S0140525X18000651>
- Goedhoop, J., Arbab, T., & Willuhn, I. (2023). Anticipation of appetitive operant action induces sustained dopamine release in the nucleus accumbens. *Journal of Neuroscience*, 43(21), 3922–3932. <https://doi.org/10.1523/JNEUROSCI.1527-22.2023>
- Greenberg, J., Pyszczynski, T., Solomon, S., Simon, L., & Breus, M. (1994). Role of consciousness and accessibility of death-related thoughts in mortality salience effects. *Journal of Personality and Social Psychology*, 67(4), 627–637. <https://doi.org/10.1037/0022-3514.67.4.627>
- Greenberg, J., Simon, L., Harmon-Jones, E., Solomon, S., Pyszczynski, T., & Lyon, D. (1995). Testing alternative explanations for mortality salience effects: Terror management, value accessibility, or worrisome thoughts? *European Journal of Social Psychology*, 12(4), 417–433. <https://doi.org/10.1002/ejsp.2420250406>
- Gronau, Q. F., Van Erp, S., Heck, D. W., Cesario, J., Jonas, K. J., & Wagenmakers, E.-J. (2017). A Bayesian model-averaged meta-analysis of the power pose effect with informed and default priors: the case of felt power. *Comprehensive Results in Social Psychology*, 2(1), 123–138. <https://doi.org/10.1080/23743603.2017.1326760>
- Harmon-Jones, C., Bastian, B., & Harmon-Jones, E. (2016). Detecting transient emotional responses with improved self-report measures and instructions. *Emotion*, 16(7), 1086–1096. <https://doi.org/10.1037/emo0000216>
- Harmon-Jones, E. (2019). *Cognitive dissonance: Reexamining a pivotal theory in psychology* (2nd ed.). American Psychological Association. <https://doi.org/10.1037/0000135-000>
- Harmon-Jones, E., Amodio, D. M., & Zinner, L. R. (2007). Social psychological methods of emotion elicitation. In J. A. Coan & J. J. B. Allen (Eds.), *Handbook of emotion elicitation and assessment* (pp. 91–105). Oxford University Press.
- Harmon-Jones, E., & Harmon-Jones, C. (2024). Dissonance in the induced-compliance paradigm: A commentary on Vaidis et al. (2024). *Advances in Methods and Practices in Psychological Science*, 7(4), Article 25152459241268308. <https://doi.org/10.1177/25152459241268308>
- Harmon-Jones, E., Harmon-Jones, C., Fearn, M., Sigelman, J. D., & Johnson, P. (2008). Left frontal cortical activation and spreading of alternatives: Tests of the action-based model of dissonance. *Journal of Personality and Social Psychology*, 94(1), 1–15. <https://doi.org/10.1037/0022-3514.94.1.1>
- Harmon-Jones, E., Szymaniak, K., Edgeworth, D., Sebban, G., & Harmon-Jones, C. (2024). Evil perceptions but not entertainment value appraisals relate to conspiracy beliefs. *Frontiers in Social Psychology*, 2, Article 1350584. <https://doi.org/10.3389/frsps.2024.1350584>
- Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review*, 2(3), 196–217. [https://doi.org/10.1207/s15327957pspr0203\\_4](https://doi.org/10.1207/s15327957pspr0203_4)

- Lishner, D. A. (2024). But did they really perceive no (low) choice? Comment on Vaidis et al. (2024). *Advances in Methods and Practices in Psychological Science*, 7(4), Article 25152459241267915. <https://doi.org/10.1177/25152459241267915>
- Meyers-Levy, J., & Zhu, R. (2007). The influence of ceiling height: The effect of priming on the type of processing that people use. *Journal of Consumer Research*, 34(2), 174–186. <https://doi.org/10.1086/519146>
- 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. <https://doi.org/10.1037/pspa0000121>
- Okken, V., van Rompay, T., & Pruijn, A. (2013). Room to move: On spatial constraints and self-disclosure during intimate conversations. *Environment and Behavior*, 45(6), 737–760. <https://doi.org/10.1177/0013916512444780>
- Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251), Article aac4716. <https://doi.org/10.1126/science.aac4716>
- Orne, M. T. (1962). On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications. *American Psychologist*, 17(11), 776–783. <https://doi.org/10.1037/h0043424>
- Pauer, S., Linne, R., & Erb, H.-P. (2024). From the illusion of choice to actual control: Reconsidering the induced-compliance paradigm of cognitive dissonance. *Advances in Methods and Practices in Psychological Science*, 7(4), Article 25152459241265002. <https://doi.org/10.1177/25152459241265002>
- Popper, K. R. (1959). *The logic of scientific discovery*. Hutchinson.
- Rife, S. C., Lambert, Q., Calin-Jageman, R., Matus, A., Banik, G., Barberia, I., Beaudry, J. L., Bernauer, H., Calvillo, D., Chopik, W. J., David, L., de Beijer, I., Evans, T. R., Hartanto, A., Kačmár, P., Legate, N., Martončík, M., Massar, K., Moreau, D., ... Wiggins, B. J. (in press). Registered replication report: Study 3 from Trafimow and Hughes (2012). *Advances in Methods and Practices in Psychological Science*. [https://osf.io/preprints/psyarxiv/esu9z\\_v2](https://osf.io/preprints/psyarxiv/esu9z_v2)
- Rosenthal, R. (1966). *Experimenter effects in behavioral research*. Appleton-Century-Crofts.
- Rosenthal, R., & Rubin, D. B. (1978). Interpersonal expectancy effects: The first 345 studies. *Behavioral and Brain Sciences*, 1(3), 377–386. <https://doi.org/10.1017/S0140525X00075506>
- Sigall, H., & Mills, J. (1998). Measures of independent variables and mediators are useful in social psychology experiments: But are they necessary? *Personality and Social Psychology Review*, 2(3), 218–226. [https://doi.org/10.1207/s15327957pspr0203\\_5](https://doi.org/10.1207/s15327957pspr0203_5)
- Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, 22(11), 1359–1366. <https://doi.org/10.1177/0956797611417632>
- Simon, L., Greenberg, J., Harmon-Jones, E., Solomon, S., Pyszczynski, T., Arndt, J., & Abend, T. (1997). Terror management and cognitive-experiential self-theory: Evidence that terror management occurs in the experiential system. *Journal of Personality and Social Psychology*, 72(5), 1132–1146. <https://doi.org/10.1037/0022-3514.72.5.1132>
- Stanley, D. J., & Spence, J. R. (2014). Expectations for replications: Are yours realistic? *Perspectives on Psychological Science*, 9(3), 305–318. <https://doi.org/10.1177/1745691614528518>
- 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. <https://doi.org/10.1037/0022-3514.54.5.768>
- Stroebe, W., & Strack, F. (2014). The alleged crisis and the illusion of exact replication. *Perspectives on Psychological Science*, 9(1), 59–71. <https://doi.org/10.1177/1745691613514450>
- Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, 18(6), 643–662. <https://doi.org/10.1037/h0054651>
- Twenge, J. M., Catanese, K. R., & Baumeister, R. F. (2003). Social exclusion and the deconstructed state: Time perception, meaninglessness, lethargy, lack of emotion, and self-awareness. *Journal of Personality and Social Psychology*, 85(3), 409–423. <https://doi.org/10.1037/0022-3514.85.3.409>
- Vaidis, D. C., Slegers, W. W. A., van Leeuwen, F., DeMarree, K. G., Sætrevik, B., Ross, R. M., Schmidt, K., Protzko, J., Morvinski, C., Ghasemi, O., Roberts, A. J., Stone, J., Bran, A., Gourdon-Kanhukamwe, A., Gunsoy, C., Moussaoui, L. S., Smith, A. R., Nugier, A., Fayant, M.-P., ... Priolo, D. (2024). A multilab replication of the induced-compliance paradigm of cognitive dissonance. *Advances in Methods and Practices in Psychological Science*, 7(1), Article 25152459231213375. <https://doi.org/10.1177/25152459231213375>
- Van Bavel, J. J., Mende-Siedlecki, P., Brady, W. J., & Reinero, D. A. (2016). Contextual sensitivity in scientific reproducibility. *Proceedings of the National Academy of Sciences*, 113(23), 6454–6459. <https://doi.org/10.1073/pnas.1521897113>
- van Prooijen, J.-W., Ligthart, J., Rosema, S., & Xu, Y. (2022). The entertainment value of conspiracy theories. *British Journal of Psychology*, 113(1), 25–48. <https://doi.org/10.1111/bjop.12522>
- Vohs, K. D., Schmeichel, B. J., Lohmann, S., Gronau, Q. F., Finley, A. J., Ainsworth, S. E., Alquist, J. L., Baker, M. D., Brizi, A., Bunyi, A., Butschek, G. J., Campbell, C., Capaldi, J., Cau, C., Chambers, H., Chatzisarantis, N. L. D., Christensen, W. J., Clay, S. L., Curtis, J., ... Albarracín, D. (2021). A multisite preregistered paradigmatic test of the ego-depletion effect. *Psychological Science*, 32(10), 1566–1581. <https://doi.org/10.1177/0956797621989733>
- Wacker, J., Mueller, E. M., Pizzagalli, D. A., Hennig, J., & Stemmler, G. (2013). Dopamine-D2-receptor blockade reverses the association between trait approach motivation and frontal asymmetry in an approach-motivation context. *Psychological Science*, 24(4), 489–497. <https://doi.org/10.1177/0956797612458935>
- Wagenmakers, E.-J., Beek, T., Dijkhoff, L., Gronau, Q. F., Acosta, A., Adams, R. B., 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. <https://doi.org/10.1177/1745691616674458>
- Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology*, 54(6), 1063–1070. <https://doi.org/10.1037/0022-3514.54.6.1063>
- Wegner, D. M., & Erber, R. (1992). The hyperaccessibility of suppressed thoughts. *Journal of Personality and Social Psychology*, 63(6), 903–912. <https://doi.org/10.1037/0022-3514.63.6.903>
- Wright, R. A., Mlynski, C., & Carbajal, I. (2019). Outsiders' thoughts on generating self-regulatory-depletion (fatigue) effects in limited-resource experiments. *Perspectives on Psychological Science*, 14(3), 469–480. <https://doi.org/10.1177/1745691618815654>
- Wright, R. A., Shaw, L. L., & Jones, C. R. (1990). Task demand and cardiovascular response magnitude: Further evidence of the mediating role of success importance. *Journal of Personality and Social Psychology*, 59(6), 1250–1260. <https://doi.org/10.1037/0022-3514.59.6.1250>

Received February 28, 2025

Revision received March 27, 2025

Accepted March 29, 2025 ■