ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/326142587

Experimental Design and the Reliability of Priming Effects: Reconsidering the "Train Wreck"

Preprint · January 2018

DOI: 10.31234/osf.io/r7pd3

CITATIONS
2

READS
543

2 authors:

Andrew M. Rivers
University of British Columbia - Vancouver
29 PUBLICATIONS **3,071** CITATIONS
SEE PROFILE

Jeffrey W Sherman
University of California, Davis
116 PUBLICATIONS **4,876** CITATIONS
SEE PROFILE

All content following this page was uploaded by Andrew M. Rivers on 01 January 2020.

The user has requested enhancement of the downloaded file.

Experimental Design and the Reliability of Priming Effects: Examining the “Train Wreck” through the Lens of Statistical Power

Andrew M. Rivers
University of British Columbia

Jeffrey W. Sherman
University of California-Davis

Failures to replicate high-profile priming effects have raised questions about the reliability of priming phenomena. Studies at the discussion’s center, labeled “social priming,” have been interpreted as a specific indictment of priming that is social in nature. However, “social priming” differs from other priming effects in multiple ways. The present research examines one important difference: whether effects have been demonstrated with within- or between-subjects experimental designs. To examine the significance of this feature, we assess the reliability of four well-known priming effects from the cognitive and social psychological literatures using both between- and within-subjects designs and analyses. All four priming effects are reliable when tested using a within-subjects approach. In contrast, only one priming effect reaches that statistical threshold when using a between-subjects approach. This demonstration serves as a salient illustration of the underappreciated importance of experimental design for statistical power, generally, and for the reliability of priming effects, specifically.

Keywords: priming; replication; statistical power; experimental design

2010 kicked off a volatile decade in psychological science. Researchers have begun questioning the current research paradigm in which the field operates, leading to an introspective period commonly referred to as a “crisis of confidence” or a “supposed crisis of confidence,” depending on one’s perspective (Spellman, 2015). The phenomenon of priming, specifically, has proven to be one of the most polarizing topics of discussion.

In the present article we argue that discussions about the replicability of priming effects have overlooked the importance of research design in contributing to successful replication of priming effects. The topics discussed within this manuscript should not be surprising, or even novel, to most readers. In fact, our central argument—that within-subjects designs are more powerful than between-subjects designs—is taught in introductory research methods classes at most universities. Despite having learned the importance of research design in statistical power and replication, we believe that its importance may continue to be overlooked because scientists lack concrete examples that tangibly illustrate the importance of design factors. To remedy this, we seek to provide concrete examples using priming effects that are familiar to researchers within many domains of psychology. In doing so, we hope to advance and sharpen the discussion surrounding the reliability of priming phenomena.

What is priming?

Most broadly, priming refers to the incidental influence of environmental context on cognition and behavior (Logan, 1980). Frequently, priming effects are thought to result from the activation of mental representations that facilitate or interfere with related subsequent behavior (Molden, 2014). Research traditions have utilized priming tasks to investigate many different theoretical questions of interest. Cognitive psychologists initially utilized priming tasks to probe the organization of mental representations (e.g., Neely, 1991). Social psychologists initially adapted priming methodologies to understand how activated knowledge and evaluations influence perception and behavior (e.g., Fazio, Sanbonmatsu, Powell, & Kardes, 1986). At a more granular level, different areas of investigation have developed unique paradigms to elicit priming effects and have generated a plethora of theoretical models to explain those effects.

Across research traditions, priming involves exposing subjects to different environmental contexts–or ‘primes’–that are incidental to, and yet still influence, subsequent behavior. For example, work from the Lexical Decision Task (Schvaneveldt & Meyer, 1973) demonstrates that people are faster to correctly identify words (e.g., ‘doctor’) following the presentation of a related prime word (e.g., ‘nurse’) compared to an unrelated prime word (e.g., ‘paper’). In other words, the target behavior – correctly indicating that ‘doctor’ is a word – occurs more quickly in contexts in which related, incidental stimuli are observed than in contexts in which unrelated stimuli are observed.

Controversies in Priming

Several recent replication studies have failed to find evidence for priming phenomena, spawning sometimes heated debates regarding the replicability of priming effects. A few, high-profile effects have been at the center of this debate, and have stayed in the spotlight for a number of years.

Arguably the most prominent single effect in the discussion is reported in Experiments 2a and 2b from Bargh, Chen, and Burrows (1996). In these experiments, participants unscramble sentences that either contained words stereotypically associated with elderly Americans (e.g., often too early retired they) or not (e.g., they her see outside normally). The dependent measure was the amount of time taken to walk from the experimental room to a nearby elevator. The key observation was that participants exposed to the elderly primes took longer to walk to the elevator than did control participants. This result was taken to demonstrate that incidentally activating the elderly stereotype produced behavior related to the stereotyped group (i.e., the elderly walk more slowly).

Two independent research teams sought to reproduce this finding using similar priming manipulations and speed of walking as the dependent measure (Doyen, Klein, Pichon, & Cleeremans, 2012; Pashler, Coburn, & Harris, 2008). Neither research team detected a priming effect similar to that reported in Bargh et al.’s (1996) Experiment 2. Doyen and colleagues’ (2012) failure to replicate received considerable press in Discover magazine, which has been shared over 900 times on social media and has been cited in academic literature more than 20 times (Yong, 2012). A few months later, Daniel Kahneman (2012) wrote an open letter to researchers working in the field of ‘social priming’ that described a “train wreck looming,” in which he describes the field as the “poster child for doubts about the integrity of psychological research.” This letter attracted considerable professional and media attention, much of which consisted of critiques of ‘social priming.’ A Google search of “Kahneman Train Wreck” yielded almost 7000 results, and the letter has been cited almost 60 times in academic journal articles.

The implication of Kahneman’s letter was that there was something particularly amiss among priming studies with a social component. However, exactly what constituted “social priming” was then and remains ambiguous and controversial. There is no consensually agreed upon definition of the term. Nevertheless, it continues to be a lightning rod for replication controversy, and the narrative of social priming as a train wreck continues to hold sway (Engber, 2017; McCook, 2017; Meyer & Chabris, 2014; Poole, 2016; Schimmack, Heene, & Kesavan, 2017).

An Alternative View of the Train Wreck: Research Design and Statistical Power

The proposed dichotomy between social and non-social priming has obscured important methodological differences among the identified priming effects and fails to account for differences in the relative reliability of various priming effects (Molden, 2014; Payne, Brown-Iannuzzi, & Loersch, 2016). Those effects identified as “social priming” differed from other priming effects in a variety of ways beyond their status as social versus non-social. For example, the types of behaviors measured and the time lags between primes and targets differed between studies identified as social and non-social priming. In the present research, we examine another important methodological difference between studies identified as social versus non-social priming; the status of the study as a within- versus between-subjects design. Specifically, the studies identified as “social priming” most frequently use between-subjects designs, whereas studies considered to exemplify non-social priming most frequently use within-subjects designs. Thus, the dichotomy between social and non-social priming largely overlaps with the dichotomy between different experimental designs. Between-subjects designs most often require larger samples of subjects to achieve the same statistical power as within-subjects designs investigating similar effects. If researchers using between-subjects designs do not obtain larger samples (as was the case in Bargh et al. 1996), their observations will be underpowered and results reported in the literature will be more difficult to replicate. This in turn suggests a new, and we argue more useful, dichotomy through which we can view the train wreck—a dichotomy between priming phenomena that are adequately- or inadequately-powered to detect priming effects.

Within-subjects designs more effectively reduce residual error by sampling from the same participants under different experimental treatments. As a result, within-subject designs offer greater power to detect effects than between-subjects designs. In the case of priming studies, the difference is reflected in dozens or even hundreds of response trials per condition in within-subject designs versus a single response trial in strictly between-subjects designs. Thus, it is not surprising that the most controversial and difficult to replicate priming effects have used between-subjects designs that were inadequately-powered to observe all but the largest effects, priming or otherwise (e.g., Bargh et al., 1996; Dijksterhuis & van Knippenberg, 1998). For example, Bargh et al.'s (1996) studies used a fully between-subjects design with two conditions and one critical measurement per participant. Those studies sampled from 15 participants per condition, a sample size that is not uncommon when studying priming effects in within-subjects designs. However, the between-subjects design used in Bargh et al. (1996) requires an assumed effect size of $d_s = 1.06$ to achieve Cohen's (1988) recommended level of power, $1\text{-}\beta = 0.8$. According to Cohen's descriptions, this effect would be 'grossly perceptible.' Among other things, the fact that the effect was found to be so large ($d_s = 1.06$ and $d_s = .82$) was viewed skeptically. Both of the teams that failed to replicate Bargh et al (1996) also used between subjects designs. Pashler et al. (2008) sampled from 66 subjects, which requires an effect size of $d_s = .70$ to achieve 80% power. Doyen et al. (2012) recruited a sample of 120 participants, which requires an effect size of $d_s = .52$ to achieve 80% power. Underpowered research, in any domain, not only reduces the ability to detect effects but also increases the rate of false discoveries in the literature through the selective publication of results (e.g., Simonsohn, 2015; Smaldino &McElreath, 2016). If the statistical observations of Bargh et al. (1996) were the product of selective publication, it should not be surprising that the effects did not emerge in subsequent replications of the work. Such a conclusion does not imply that priming effects with social stimuli are, on the whole, less reliable; instead, it implies that underpowered research (i.e., between-subjects designs in combination with small samples) published in the literature will be less reliable.

In their blog post "Reconstruction of a train wreck: How priming research went off the rails," Schimmack et al., (2017) analyze experimental evidence cited in the fourth chapter of Thinking Fast and Slow (Kahneman, 2011), the chapter in which Kahneman described the priming effects he later labeled "social priming" and lamented as unreliable. This analysis showed that 29 [1] separate experiments rejected the null hypothesis (i.e., achieved the NHST threshold of $p < .05$) even though the studies had an estimated 57% average power (see Schimmack, 2012; 2017). Schimmack et al. concluded that, "selective reporting of studies that worked is at the core of the replicability crisis in social psychology." Of most relevance to our discussion, all of the studies included in the Schimmack et al. analyses used a fully between-subjects design. Their analysis excluded "cognitive priming effects," a body of literature that almost exclusively uses within-subjects designs, indicating that no citations for such studies were provided by Kahneman (2011).

[1] Note that Schimmack et al. (2017) indicate they analyzed 31 experimental results, though only 29 statistics are presented. This discrepancy may arise from the inclusion of test statistics from Schimmack et al.'s point 4.5. In any case, the discrepancy has no implications for the present discussion.

In other words, the studies described as "social priming" by Schimmack et al. (2017) used between-subjects designs, whereas the studies considered to reflect non-social priming (i.e., semantic priming) used within-subjects designs. We propose that the key determinant of replicability is not whether the phenomenon is social or non-social, but is instead whether reported experiments had sufficient power to detect priming effects—a factor that is inexorably linked to experimental design.

Payne et al. (2016) similarly proposed the importance of design type in accounting for variation in replicability, but did not directly test it. In his critique of Payne et al.'s work, Shanks (2017) argued that consideration of the within- versus between-subject nature of different priming studies "sheds minimal light on the priming controversy" (p. 1221). In the present research, we provide a concrete demonstration of the influence of design type on the replicability of priming effects. Among these effects, we demonstrate that differences in reliability are not dependent on the social versus non-social nature of the priming task, but rather are dependent on design type. To do this, we investigate two relatively robust priming effects from the literature in cognitive psychology and two robust effects from the priming literature in social psychology, and manipulate whether those effects are tested in a within- versus between-subjects design.

Experiments

Overview of Analysis Approach

For each priming effect, we performed three separate analyses that conceptually mirror design choices made by researchers studying priming effects. In the first "full information" approach, we analyze the data with traditionally used repeated-measures analyses. Each participant receives an estimate of their performance in the two trial types of interest that is informed by their behavioral performance on multiple relevant trials.

The second "one-shot" approach most closely approximates a between-subjects design used in studies like Bargh et al. (1996). In this approach, each participant is randomized to one level of the experimental design based on the first experimental trial to which they are exposed. The effect of priming is then analyzed using fully between-subjects tests.

We supplement the "one-shot" approach with a third "representative-shot" analysis. In this approach, we randomly assign each participant to one experimental level, and then randomly sample one trial from each participant's full set of responses (rather than only the first relevant trial).

Methods

Stroop Task. Data for the Stroop task come from 3,337 participants who took part in the ManyLabs 3 project (Ebersole, Atherton, Belanger, Skulborstad, Allen, et al., 2016). The Stroop task is a simultaneous priming procedure in which participants view a stimulus word that names a color that is printed in ink that is either congruent or incongruent with the word (see Table 1 for technical specifications).[2] The Stroop effect refers to longer latencies for identifying color on incongruent versus congruent trials. Ebersole et al. (2016) report *D* scoring indices that filter out trials with latencies greater than 10,000-ms. Additionally, incorrect responses are replaced with the mean latency for the corresponding trial type plus a 600-ms penalty. Our analyses compare latencies on congruent versus incongruent trials for a random sample of 115 participants.[3]

[2] Though the Stroop task is not traditionally thought of as a priming task, it meets all the criteria for scientific definitions of priming. Specifically, as in other priming tasks, participants view a stimulus array that contains both task-irrelevant (prime) and task-relevant (target) features. Priming captures the degree to which cognition and behavior are influenced by irrelevant or incidental stimuli (the word vs. its color). Traditionally, priming tasks have presented the prime, an interval without a stimulus that determines stimulus onset asynchrony, and then a target. Thus, conventionally, priming tasks have a positive SOA, but this is not necessary. Stroop represents a case in which SOA is zero–both task-relevant and task-irrelevant stimuli are presented simultaneously. A number of researchers have investigated the mechanisms of priming be varying the SOA from negative (the target precedes the prime) to zero, to positive (e.g., Gawronski & Ye, 2013; Musch & Klauer, 1997; see Glaser and Glaser, 1982 and Logan 1980, for examples within the Stroop literature).

[3] We sought to sample from at least 100 participants for the WIT and SMT Experiments, and randomly sampled 115 cases from the LDT and Stroop data repositories. We report

Lexical Decision Task (LDT). Data for the LDT come from 512 participants who completed the task as part of the Semantic Priming Project (Hutchison, Balota, Neely, Cortese, Cohen-Shikora, et al., 2013). The LDT is a sequential priming procedure in which participants view a prime word that is followed by one of three types of target strings (related words, unrelated words, or non-words). Participants are to press one of two keys, as quickly as possible without making too many errors, to indicate whether the target is a word or nonword. In the present data, related target words were first associates of prime words that directly preceded[4] these targets. Unrelated target words were first associates for primes that were presented in other trials; unrelated targets instead were preceded by primes that they were not highly related to. Non-word targets consisted of target words that were modified by changing one or two letters to produce a pronounceable, but non-existent word (see Hutchison et al., 2013 for more details).

The semantic priming effect refers to reduced latency for identifying target words on trials in which primes are related versus unrelated. Hutchison et al. (2013) trim data on trials in which latencies differ from each participant's mean by greater than +/- 3SD. Additionally, only correct identifications are included in the analyses. We specifically analyze data for the two critical trial types that are used to illustrate the semantic priming effect: first-associate target words that are either related to or unrelated to the prime word. Our analyses compare response latencies on related versus unrelated trials for a sample of 115 participants.

Weapons Identification Task (WIT). Data for the WIT (Payne, 2001) come from a sample of 120 undergraduates at the University of California, Davis. The WIT is a sequential priming procedure in which participants view prime head and shoulders photographs of Black or White men, followed by targets consisting of black and white sketches of guns or tools. Participants are to press one of two keys to identify whether the target shown was a gun or a tool. Further, participants are instructed to respond as quickly as possible and were shown a message to "Please try to respond faster!" written in bolded red text if they registered a response after the 450-ms response deadline.

The WIT effect refers to a greater proportion of identification errors on stereotype incongruent trials (Black-Tool; White-Gun) versus congruent trials (White-Tool; Black-Gun). Computer errors rendered incomplete data for 2 participants. We excluded casewise data from one participant for making an inordinate number of errors (85% versus a sample average error rate of 25%). Thus, the final sample consisted of 117 participants. Our analyses compare the proportion of errors on stereotype congruent versus incongruent trials.

Stereotype Misperception Task (SMT). Data for the SMT (Krieglmeyer & Sherman, 2012) come from a sample of 114 undergraduates at the University of California, Davis. The SMT is a sequential priming procedure in which prime images of Black or White faces precede pixelated target faces generated with face-morphing software. Participants are to press one of two keys to identify whether each target is "more threatening" or "less threatening" than the average target encountered during the task.[5] Participants complete a series of twelve practice trials to acquaint them with the task and to calibrate them to relatively more and less threatening targets. Participants are asked to make identifications quickly and further are explicitly instructed to avoid the influence of prime stimuli while identifying target stimuli.

The SMT effect refers to a greater proportion of more threatening target identifications on Black prime trials versus White prime trials. Three participants were excluded from the analyses for utilizing the same response key on all trials. Thus, the final sample consisted of 111 participants. Our analyses compare the

all data exclusions. Additionally, data and a list of additional measures collected during the WIT and SMT studies is available at https://osf.io/wqn9r/

[4] The reported mean strength of first-associates was 0.31 for related words and 0.05 for unrelated words, meaning that 31 percent of participants in norming studies would write a related target word as the first word that came to mind upon seeing the prime word (see Hutchison et al., 2013). We sampled directly from the *Semantic Priming Project* data base and did not apply any further criterion for related vs. unrelated word pairings.

[5] Target images are freely available for download at https://osf.io/pqbhf/

proportion of “more threatening” responses aggregated for Black and White prime trials.

Results

For comparability, we randomly sampled 115 cases from the Stroop and LDT data repositories and used the full samples for the WIT (N=117) and SMT (N=111). We report corrected effect sizes of Hedges gav and gs, which control for repeated-measures correlation and permit comparisons of effect size across between- and within-subjects analyses. To analyze between-subjects judgment data in WIT and SMT, we fit logistic regression models and report odds ratios and their conversions to gs. In addition to Hedges gav, which controls for the correlation of repeated measurements, we also report Cohen’s d_z, preferred for power analyses in within-subjects designs (see Lakens, 2013).

Full Information. Repeated-measures analyses detected priming effects in all four priming tasks. Incongruent Stroop trials had longer response latencies than congruent Stroop trials, $t(114) = 6.052, p < .001$, Hedges $g_{\mathrm{av}} = .220$; 95% CI$_{g\mathrm{av}}$ [.142, .296], $d_z = .565$; 95% CI$_{dz}$ [.369, .761]. Related LDT prime trials had shorter response latencies than unrelated prime trials, $t(114) = 6.955$, $p < .001$, $g_{\mathrm{av}} = .193$; 95% CI$_{g\mathrm{av}}$ [.133, .254], $d_z = .649$; 95% CI$_{dz}$ [.449, .850].[6] Stereotype-incongruent WIT trials produced more errors than stereotype-congruent trials, $t(116) = 3.334$, $p = .001$, $g_{\mathrm{av}} = .226$; 95% CI$_{g\mathrm{av}}$ [.159, .293], $d_z = .309$; 95% CI$_{dz}$ [.124, .494]. Black SMT trials produced a greater proportion of “more threatening” responses than White prime trials, $t(110) = 3.078$, $p = .003$, $g_{\mathrm{av}} = .394$; 95% CI$_{g\mathrm{av}}$ [.163, .604], $d_z = .292$; 95% CI$_{dz}$ [.102, .482].

One-Shot. This analysis included only the first response from each participant, making the design entirely between-subjects. To be clear, each participant is assigned to only one condition of each priming experiment based on the first trial they completed.[7] Students t-tests failed to detect priming effects in either the Stroop or the LDT. Incongruent Stroop trials were directionally, but not reliably, slower than congruent trials, $t(113) = 1.244$, $p = .216$, $g_s = .247$; 95% CI$_{gs}$ [-.146, .638]. Related LDT prime trials had directionally, but not reliably, shorter latencies than unrelated trials, $t(113) = 1.097$, $p = .275$, $g_s = .203$; 95% CI$_{gs}$ [-.162, .571].

Logistic regression analyses indicated that WIT target identification errors were more likely on stereotype-incongruent versus stereotype-congruent trials, $\chi2(1) = 3.918$, $p = .048$, $OR = 2.472$, $g_s = .496$; 95% CI$_{gs}$ [.005, .986].[8] SMT threat responses were directionally, but not reliably, more likely on Black versus White prime trials, $\chi2(1) = .680$, $p = .410$, $OR = 1.378$, $g_s = .177$; 95% CI$_{gs}$ [-.243, .597].

Representative Shot. This analysis included only a single response from each participant that was selected randomly from all responses.[9] As such, this is also an entirely between-subjects design where each participant was assigned to only one condition of each priming experiment. Students t-tests failed to detect priming effects in either the Stroop or the LDT. Incongruent Stroop trials were directionally, but not reliably, slower than congruent trials, $t(113) = 1.117$, $p = .267$, $g_s = .207$; 95% CI$_{gs}$ [-.158, .573]. Related LDT prime trials had directionally, but not reliably, shorter latencies than related trials, $t(113) = 1.261$, $p = .210$, $g_s = .234$; 95% CI$_{gs}$ [-.133, .602].

Logistic regression analyses indicated that identification errors were marginally more likely on stereotype-incongruent versus stereotype-congruent WIT trials, $\chi2(1) = 3.409$, $p = .065$, $OR = 2.429$, $g_s = .489$; 95% CI$_{gs}$ [-.030, 1.009]. “More threatening” responses on the SMT were directionally, but not reliably, more likely on

[6] Analysis of all available 512 participants yields an almost identical effect size of $d_z = .648$

[7] Because the Stroop task employed an unequal ratio of congruent vs. incongruent trials, this resulted in an unbalanced design. The present non-significant statistical result is unchanged if a sampling constraint is added to ensure a balanced design (consisting of both first and second Stroop trials). Further, we constrain sampling to reflect a balanced design in the ‘representative shot’ analyses reported below.

[8] Odds ratios were converted to Hedges g_s using the formulas provided by Borenstein, Hedges, Higgins, & Rothstein (2009) and Lakens (2013).

[9] For Stroop data, we set an additional sampling constraint to select from congruent vs. incongruent trials with equal frequency. To do this, participants were first randomly assigned to a condition and trial number. If the selected trial did not match the assigned condition, we sampled from the nearest trial that did match the condition.

Black versus White prime SMT trials, $\chi 2(1) = 0.435$, $p = .510$, $OR = 1.294$, $g_s = .142$; 95% CI_{gs} [-.280, .564].

Statistical Power. To gain a clearer estimate of how many participants would be required to produce reliable effects on each task, we conducted a set of power analyses using the observed effect sizes from each of the present results. All power analyses were conducted in the freely available G*Power package (Faul, Erdfelder, Buchner, & Lang, 2009). Specifically, we sought to determine the number of participants required to achieve Cohen's (1988) recommended $1 - \beta = .8$ level of power. For within-subjects designs, this is most commonly computed using Cohen's d_z effect size metric, and for between-subjects analyses, we use g_s. Because we are examining the replicability of previous work, all power analyses use 1-tailed distributions.[10]

The Stroop task yielded a Cohen's $d_z = .565$; 95% CI_{dz} [.369, .761]. Sampling from 21 participants would achieve the $1 - \beta = .8$ level of power in a within-subjects design (see Table 2). For a comparable level of power in a between-subjects design, results from the one-shot analysis indicated that researchers would need to sample from 404 participants. Results from the representative shot analyses indicated that a sample of 574 participants would achieve the .8 level of power.

--- Table 2 here ---

The LDT analyses yielded a Cohen's $d_z = .649$; 95% CI_{dz} [.449, .850]. Sampling from 17 participants would achieve the $1 - \beta = .8$ level of power in a within-subjects design. Results from the one-shot and representative-shot analyses indicated that researchers would need to sample from 602 and 454 participants respectively to achieve a comparable level of power.

The WIT yielded a Cohen's $d_z = .309$; 95% CI_{dz} [.124, .494]. Sampling from 67 participants would achieve the $1 - \beta = .8$ level of power in a within-subjects design. For a comparable level of power in a between-subjects design, results from the one-shot and representative-shot analyses indicated that researchers would need to sample from 102 and 106 participants respectively.

The SMT yielded a Cohen's $d_z = .292$; 95% CI_{dz} [.102, .482]. Sampling from 74 participants would achieve the $1 - \beta = .8$ level of power in a within-subjects design. For a comparable level of power in a between-subjects design, results from the one-shot and representative-shot analyses indicated that researchers would need to sample from 792 and 1228 participants respectively.

Discussion

Much debate has surrounded the question of when priming effects are and are not robust and reliable. In these studies, we provided a clear demonstration of the importance of experimental design and statistical power if we wish to have a full accounting of priming effects' replicability. Priming effects reliably emerged when the research design and analyses used a within-subjects approach in which the experiment was highly-powered to detect priming effects. Depending on the task, target N for reaching 80% power ranged between 17 and 74 participants. (see Table 2). In contrast, priming effects did not emerge, with one exception, when the research design adopted a between-subjects approach in which the experiment lacked statistical power to reliably detect priming effects. Depending on the task, target N for reaching 80% power ranged between 102 and 1228 participants.

Given these results, it is not surprising that the most controversial and difficult to replicate priming studies used between-subjects designs with small samples of participants. It so happens that the highest profile cases were conducted by social psychologists and published in social psychological journals. And, indeed, there are reasons why social psychological priming research may be more likely to rely on between-subjects designs than non-social priming research. In particular, concerns about carryover effects in within-subjects designs have sometimes precluded the use of such designs in

[10] Note that the degree to which an effect is detectable ($1 - \beta$) is a function of the correlation between repeated-measures for within-subjects analyses and is implicit in their corresponding effect sizes (i.e., Cohen's d_z). In between-subjects analyses, there is necessarily no contribution from correlated measures. Thus, effects that are easily observed in the within-subjects paradigm may be more difficult to replicate in a between-subjects context (for example, see results for the Lexical Decision Task).

social psychological experiments (see Greenwald, 1976). Nevertheless, the current data make clear that characterizing the replicability of priming effects requires accounting for the statistical power afforded by the research design. In contrast, we see no reason to expect that the replicability of priming effects would depend on their social versus non-social nature. The present analyses revealed robust priming effects with tasks that use socially-relevant primes and targets, were developed by social psychologists, and were published in social psychology journals. We also showed that both social and non-social priming effects are similarly affected by the choice of within- versus between-subjects design. However, a full investigation of the robustness of social versus non-social priming effects, controlling for design and power, is beyond the scope and goals of this paper.

It is not our intention to suggest that failures to replicate priming effects can be solely attributed to research design. Recent meta-analyses indicate that publication bias is present in literatures investigating several priming effects (e.g., priming religious concepts; van Elk, Matzke, Gronau, Guan, Vandekerckhove, & Wagenmakers, 2015; priming mating motives; Shanks et al., 2015). When publication bias is evident, it is unclear how informative the existing literature is in determining whether a specific effect does or does not exist. Notably, meta-analytic bias-correction techniques may fail when studies in a particular literature are underpowered (see Stanley, 2017). Under these conditions, highly powered direct replication can be an irreplacable tool to investigate the veracity of an effect. There are several examples of research teams that have carried out high-powered replications, even with between-subjects paradigms that required large samples. Gomes and McCullough (2015) were unable to replicate the effect of religious priming on decisions in an economic game despite 455 participants across two critical between-subjects conditions. Similarly, Shanks and colleagues (2015), across 9 separate experiments (N=1,325), failed to find evidence consistent with previous work showing that priming mating motives affects people's spending behavior. Shanks et al. (2015) had at least 80% power to detect a between-subjects effect as small as $d_s = .14$. This highly powered failure to replicate in tandem with demonstrable evidence of publication bias indicates that the reported effect – priming mating motives influences consumer behavior – is most likely a false discovery (i.e., Type I error).

We anticipate that some critics will not be satisfied that we have examined "social priming." After Payne et al. (2016) published a robust priming task that altered people's betting behavior in a gambling situation, the work was criticized as "not comparable to the kind of 'social priming' as practiced by Staple [sic] and others" (Neuroskeptic, 2017). Indeed, there are many differences between the gambling paradigm developed by Payne and colleagues (2016) and more frequently discussed paradigms such as in Bargh et al. (1996). However, we are unable to determine why the label "social priming" applies to one type or another. Other robust priming tasks, such as WIT and SMT, that use socially relevant primes and targets, were developed by social psychologists, and were published in social psychological journals are similarly reliable, yet are frequently overlooked or discounted in "social priming" critiques (see Cesario & Jonas, 2014 for a similar discussion). The most common objections are that the term "social priming" should be reserved to describe effects on particular kinds of behaviors and/or on measures that include a relatively long gap between prime and target. We would note that there is nothing inherently "social" about either of these features of priming tasks. For example, it is not clear what is particularly "social" about walking down a hallway (Bargh et al., 1996). It also is not clear what is "social" about inserting a relatively long delay between prime and target. Indeed, cognitive psychologists too have employed relatively longer delays after priming using the Deese/Roediger-McDermott paradigm and a lexical decision task to investigate semantic priming effects up to 30 seconds after priming (see Tse & Neely, 2005; 2007). Rather, these are design features, such as using a within- versus between-subject design that may vary across content and across research domains. Research directly examining these features would be useful for investigating the moderators and mechanisms of priming effects.

To this point, no one has provided clear guidance as to what kinds of effects should "count" as "social priming," but a few influential recently published works are frequently cited in this area. Weingarten et al.'s (2016) meta-analysis is frequently interpreted as an all-encompassing test of the "social priming" literature. However, the authors synthesized research more narrowly, investigating incidental presentation of word primes on what they referred to as "behavior measures." Thus, the researchers focus on studies with particular design features (e.g., word priming; behavior measures) to test a specific theory—the perception-behavior model—rather than attempt to quantify the entire social priming literature. Payne and colleagues (2016) also examined a subset of research, in which primes impact downstream actions and decisions. They refer to this type of priming as "behavior priming," and it may or may not have social aspects. Payne et al. (2016) explicitly state in their manuscript that, "there is no such thing as 'social priming'" (pp. 1270). In contrast to these two influential recent articles, we speak to "social priming" more broadly. In our view, priming tasks that use socially relevant primes and/or targets are measures of social priming, regardless of design features, such as the specific format of the primes (e.g., words vs. faces), the particular behavior measured (button presses vs. walking speed), the within- versus between-subject nature of the design, or the SOA (negative; zero; 300 milliseconds; 3 minutes). As it stands, the term "social priming" offers nothing in the way of explaining the relative robustness of different priming effects. Moreover, use of that term obscures important design features that do help to account for the relative robustness of priming effects.

So, then, what to make of the findings of Bargh et al. (1996) and other priming effects that have relied on strictly between-subjects designs? At this point, we believe that the data are simply uninformative. If we assume that the effect size of Bargh et al.'s (1996) elderly study is comparable to a simple average of the four present priming effects (a very optimistic estimate in our view), then we would require at least 388 participants over 2 between-subjects conditions to achieve the recommended .8 level of power. If we combine all reported data available on Curate Science (Bargh et al., 1996; Cesario et al., 2007; Doyen et al., 2012; Hull et al., 2002; Pashler et al., 2008), the resulting sample (N = 447) would provide the recommended level of power for a single experimental test of the hypothesis. Using a within-subject design, we might reasonably power 6 or 7 experimental tests of the hypothesis. Unfortunately, it is not possible to ask subjects to walk down the same hallway 300 times after exposure to different primes. Given the available evidence from the initially reported effect and subsequent failures to replicate, it seems that, at best, the effect must be substantially smaller than initially reported.

We also want to be clear that we are not suggesting that statistical power should be prioritized above developing psychological theory or above the study of behavior in natural contexts. There are a multitude of important theoretical questions that cannot be answered using within-subjects approaches and, similarly, there are many difficulties that preclude the use of within-subjects approaches in naturalistic experiments. Many social issues of critical importance are likely only possible to study using between-subjects designs and analyses. We do not wish to suggest that psychologists abandon the study of important psychological principles or of naturalistic behavior simply because within-subjects approaches are not possible. Instead, researchers employing between-subjects designs should seek to maximize statistical power through all means available (e.g., Chartier et al., 2017; Judd, Westfall, & Kenny, 2017; Wang, Sparks, Gonzalez, Hess, & Ledgerwood, 2017).

Conclusion

We hope that the present demonstration will serve as a salient example to illustrate the still-underappreciated importance of research design for statistical power and the reliability of research findings. We also hope that this research shifts the debate from content area (social vs. nonsocial) to statistical power as an explanation for the robustness of priming effects. Whereas the former debate leads to a dead-end (i.e., don't investigate priming of social concepts), the latter debate informs the way we evaluate and design future priming research.

References

Aspendorpf, J. B., Conner, M., De Fruyt, F., De Houwer, J., Denissen, J. J. A., Fiedler, K., …Wicherts, J. M. (2013). Recommendations for increasing replicability in psychology. European Journal of Personality, 27(2), 108-118.

Bargh, J. A., Chen, M., & Burrows, L. (1996). Automaticity of social behavior: Direct effects of trait construct and stereotype activation on action. Journal of Personality and Social Psychology, 71(2), 230-244.

Cesario, J., Plaks, J. E., & Higgins, E. T. (2006). Automatic social behavior as motivated preparation to interact. Journal of Personality and Social Psychology, 90(6), 893-910.

Cesario, J., & Jonas, K. (2014). Expectations of replicability and variability in priming effects, part 1: Setting the scope and some basic definitions. Open Science Collaboration Blog. Retrieved from http://osc.centerforopenscience.org/2014/04/09/expectations-1/

Chartier, C. R., McCarthy, R. J., Williams, S. R., Ebersole, C. R., Hamlin, K., Lucas, R. E., … Lenne, R. L. (2017). StudySwap: A platform for interlab replication, collaboration, and research resource exchanged. Retrieved from osf.io/9aj5g

Cohen, J. (1988). Statistical power analysis for the behavioral sciences. Hillsdale, NJ: Lawrence Erlbaum.

Dijksterhuis, A., & van Knippenberg, A. (1998). The relation between perception and behavior, or how to win at a game of Trivial Pursuit. Journal of Personality and Social Psychology, 74(4), 865-877.

Dimberg, U., Thunberg, M., & Grunedal, S. (2002). Facial reactions to emotional stimuli: Automatically controlled emotional responses. Cognition & Emotion, 16(4), 449-471.

Doyen, S., Klein, O., Pichon, C., & Cleeremans, A. (2012). Behavioral priming: It's all in the mind, but whose mind? PLoS One, 7(1), e29081.

Ebersole, C. R., Atherton, O. E., Belanger, A. L., Skullborstad, H. M., Allen, J. M., …Nosek, B. A. (2016). Many Labs 3: Evaluating participant pool quality across the academic semester via replication. Journal of Experimental Social Psychology, 67, 68-82.

van Elk, M., Matzke, D., Gronau, Q. F., Guan, M., Vandekerckhove, J., & Wagenmakers E. J. (2015). Meta-analyses are no substitute for registered replications: A skeptical perspective on religious priming. Frontiers in Psychology, 6(1365).

Engber, D. (2017, May). Daryl Bem proved ESP is real: Which means science is broken. Slate. Accessed at: https://slate.com/health-and-science/2017/06/daryl-bem-proved-esp-is-real-showed-science-is-broken.html

Faul, F., Erdfelder, E., Buchner, A., & Lang, A. (2009). Statistical power analysis using G*Power 3.1: Tests for correlation and regression analyses. Behavior Research Methods, 41(4), 1149-1160.

Fazio, R. H., Sanbonmatsu, D. M., Powell, M. C., & Kardes, F. R. (1986). On the automatic activation of attitudes. Journal of Personality and Social Psychology, 50(2), 229-238.

Ferguson, M. J., & Mann, T. C. (2014). Effects of evaluation: An example of robust "social" priming. Social Cognition, 32, 33-46.

Gawronski, B., & Ye, Y. (2014). What drives priming effects in the affect misattribution procedure? Personality and Social Psychology Bulletin, 40(1), 3-15.

Glaser M. O., & Glaser, W. R. (1982). Time course analysis of the Stroop phenomenon. Journal of Experimental Psychology: Human Perception and Performance, 8(6), 875-894.

Gomes, C. M., & McCullough, M. E. (2015). The effects of implicit religious primes on dictator game allocations: A preregistered replication experiment. Journal of Experimental Psychology: General, 144(6), 94-104.

Greenwald, A. G. (1976). Within-subjects designs: To use or not to use. Psychological Bulletin, 83(2), 314-320.

Hull, J. G., Slone, L. B., Meteyer, K. B., & Matthews, A. R. (2002). The nonconsciousness of self-consciousness. Journal of Personality and Social Psychology, 83(2), 406-424.

Hutchison, K. A., Balota, D. A., Neely, J. H., Cortese, M. J., Cohen-Shikora, E. R., …Buchanan, E. (2013). The semantic priming project. Behavioral Research Methods, 45(4), 1099-1114.

Judd, C. M., Westfall, J., & Kenny, D. A. (2017). Experiments with more than one random factor: Designs, analytic methods, and statistical power. Annual Review of Psychology, 68, 601-625.

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. Frontiers in Psychology, 4(863).

LeBel, E. P., Vanpaemel, W., McCarthy, R., Earp, B., & Elson, M. (2017). A unified framework to quantify the trustworthiness of empirical research. Preprint accessed at: https://psyarxiv.com/uwmr8

Letzter, R. (2016). One of the most controversial ideas in psychology just got a boost. Business Insider. Accessed at: http://www.businessinsider.com/a-new-paper-shows-real-evidence-for-social-priming-2016-10

Logan, G. D. (1980). Attention and automaticity in the Stroop and priming tasks: Theory and data. Cognitive Psychology, 12(4), 523-553.

Kahneman, D. (2011). Thinking, fast and slow. New York: Farrar, Straus, and Giroux.

Kahneman, D. (2012). A proposal to deal with questions about priming effects.

Krieglmeyer, R., & Sherman, J. W. (2012). Disentangling stereotype activation and stereotype application in the Stereotype Misperception Task. Journal Personality and Social Psychology, 103(2), 205-224.

McCook, A. (2017, February). ‘I placed too much faith in underpowered studies:’ Nobel Prize winner admits mistakes. Retraction Watch.

Meyer, M. N., & Chabris, C. (2014). Why psychologists' food fight matters. Slate. Accessed at: http://www.slate.com/articles/health_and_science/science/2014/07/replication_controversy_in_psychology_bullying_file_drawer_effect_blog_posts.html

Molden, D. (2014). Understanding priming effects in social psychology: What is “social priming” and how does it occur? Social Cognition, 32, 1-11.

Musch, J., & Klauer, K. C. (1997). The proportion affect in affective priming: Replication and evaluation of a theoretical explanation. Zeitschrift Experimental Psychology, 44(2), 266-292.

Neely, J. H. (1991). Semantic priming effects in visual word recognition: A selective review of current findings and theories. Basic Processes in Reading: Visual Word Recognition, 11, 264-336.

Neuroskeptic (2016). Social priming – does it work after all? Discover Blogs. Accessed at: http://blogs.discovermagazine.com/neuroskeptic/2016/10/13/social-priming-works-after-all/

Neuroskeptic (2017). More on ‘behavior priming’ and unconscious influences. Discover Blogs. Accessed at: http://blogs.discovermagazine.com/neuroskeptic/2017/08/16/behavior-priming-controversy/

Pashler, H., Harris, C., & Coburn, N. (2008). Elderly-Related Words Prime Slow Walking. Accessed at: http://www.PsychFileDrawer.org/replication.php?attempt=MTU%3D

Payne, B. K. (2001). Prejudice and perception: The role of automatic and controlled processes in misperceiving a weapon. Journal of Personality and Social Psychology, 81(2), 181-192.

Payne, B. K., Brown-Iannuzzi, J. L., & Loersch, C. (2016). Replicable effects of primes on human behavior. Journal of Experimental Psychology: General, 145(10), 1269-1279.

Poole, S. (2016, June). Why bad ideas refuse to die. Guardian. Accessed at: https://www.theguardian.com/science/2016/jun/28/why-bad-ideas-refuse-die

Ramscar, M. (2016). Learning and the replicability of priming effects. Current Opinion in Psychology, 12, 80-84.

Schimmack, U. (2012). The ironic effect of significant results on the credibility of multiple-study articles. Psychological Methods, 17(4), 551-566.

Schimmack, U., Heene, M., Kesavan, K. (2017, February). Reconstruction of a train wreck: How priming research went off the rails. Replicability Index. Accessed at: https://replicationindex.wordpress.com/2017/02/02/reconstruction-of-a-train-wreck-how-priming-research-went-of-the-rails

Schvaneveldt, R. W. & Meyer, D. E. (1973). Retrieval and comparison processes in semantic memory. In S. Kornblum (ed.), Attention and Performance IV. New York, Academic Press.

Shanks, D. R., Vadillo, M. A., Riedel, B., Clymo, A., Govind, S., Hickin, N., Tamman, A. J., & Puhlmann, L. M. (2015). Romance, risk, and replication: Can consumer choices and risk-taking be primed by mating motives? Journal of Experimental Psychology: General, 144(6), 142-158.

Simonsohn, U. (2015). Small telescopes: Detectability and the evaluation of replication results. Psychological Science, 26(5), 559-569.

Smaldino, P. E., & McElreath, R. (2016). The natural selection of bad science. Royal Society of Open Science, 3.

Spellman, B. A. (2015). A short (personal) future history of revolution 2.0. Perspectives on Psychological Science, 10(6), 886-899.

Stanley, T. D. (2017). Limitations of PET-PEESE and other meta-analysis methods. Social Psychological and Personality Science. doi: 10.1177/1948550617693062

Tse, C-S., & Neely, J. H. (2005). Assessing activation without source monitoring in the DRM false memory paradigm. Journal of Memory and Language, 53(4), 532-550.

Tse, C-S., & Neely, J. H. (2007). Semantic and repetition priming effects for the Deese/Roediger-McDermott (DRM) critical items and associates produced by DRM and unrelated study lists. Memory & Cognition, 35(5), 1047-1066.

Wang, Y. A., Sparks, J., Gonzalez, J. E., Hess, Y. D., & Ledgerwood, A. (2017). Using independent covariates in experimental designs: Quantifying the trade-off between power boost and Type I error inflation. Journal of Experimental Social Psychology, 72, 118-124.

Weingarten, E., Chen, Q., McAdams, M., Yi, J., Helper, J., & Albarracín (2016). From primed concepts to action: A meta-analysis of the behavioral effects of incidentally presented words. Psychological Bulletin, 142(5), 472-497.

Wilson, A. D. (2013). Social priming: Of course it only kind of works. Psychology Today. Accessed at: https://www.psychologytoday.com/blog/cognition-without-borders/201310/social-priming-course-it-only-kind-works

Yong, E. (2012). A failed replication draws a scathing personal attack from a psychology professor. Discover blogs: Not Exactly Rocket Science. Accessed at: http://blogs.discovermagazine.com/notrocketscience/2012/03/10/failed-replication-bargh-psychology-study-doye

Table 1. Task specifications for Stroop, LDT, WIT, and SMT.

	Total Trials	Critical Trials	Fixation Cross	Prime Duration	ISI / SOA
Stroop	63	63	N/A	w/target	NA/0-ms
LDT	416	104	500-ms	150-ms	50-/200-ms
WIT	144	96	500-ms	150-ms	100-/250-ms
SMT	144	96	500-ms	150-ms	50-/200-ms

	Target Duration	Backward Mask	Response Deadline	Inter-trial Interval	Comp:incomp:irrelevant
Stroop	until ID	No	N/A	500-ms	1:2:0
LDT	until ID	No	3000-ms	1500-ms	1:1:2
WIT	100-ms	Yes	450-ms	500-ms	1:1:1
SMT	100-ms	Yes	N/A	500-ms	1:1:1

Table 2. Test statistics, effect sizes, and N (in participants) required for 80% power by experimental design type.

	Test statistic	*p*	Raw mean difference [11]	Effect size (ES)	ES 95% CI	N for 1-β = .80
Stroop color-naming task (N=115)						
Full-information	$t(114) = 6.052$	<.001	67.97	$g_{av} = .220$	[.142, .296]	21
One-shot	$t(113) = 1.244$	.216	160.90	$g_s = .247$	[-.146, .638]	404
Representative-shot	$t(113) = 1.117$	.267	61.04	$g_s = .207$	[-.158, .573]	574
Lexical Decision Task (LDT; N=115)						
Full-information	$t(114) = 6.955$	<.001	29.26	$g_{av} = 193$	[.133, .254]	17
One-shot	$t(113) = 1.097$	.275	45.63	$g_s = .203$	[-.162, .571]	602
Representative-shot	$t(113) = 1.389$	.168	61.80	$g_s = .257$	[-.109, .626]	454
Weapons Identification Task (WIT; N=117)						
Full-information	$t(116) = 3.334$	.001	0.04	$g_{av} = .226$	[.159, .293]	67
One-shot	$\chi^2(1) = 3.918$	.048	0.16	$g_s = .496$	[.005, .986]	102
Representative-shot	$\chi^2(1) = 3.409$	.065	0.14	$g_s = .489$	[-.030, 1.009]	106
Stereotype Misperception Task (SMT; N=111)						
Full-information	$t(110) = 3.078$	.003	0.09	$g_{av} = .394$	[.163, .624]	74
One-shot	$\chi^2(1) = 2.251$	.134	0.06	$g_s = .177$	[-.243, .597]	792
Representative-shot	$\chi^2(1) = 0.925$	.336	0.01	$g_s = .142$	[-.280, .564]	1228

[11] Note: Raw means differences in milliseconds (Stroop/LDT), proportion identification errors (WIT), and proportion “more threatening” responses

View publication stats