The role of experimenter belief in social priming

Gilder, Thandiwe; Heerey, Erin

Psychological Science

DOI:
10.1177/0956797617737128

Published: 01/03/2018

The role of experimenter belief in social priming

Thandiwe S E Gilder[1] and Erin A Heerey[1,2]

[1] Bangor University School of Psychology

[2] Western University, Department of Psychology

Running Head: EXPERIMENTER EFFECTS IN SOCIAL PRIMING

Word Count (excluding abstract, methods, results and references): 1999
Abstract: 150
Tables: 1
Figures: 3
References: 40

Address for Correspondence:
Erin A Heerey
Department of Psychology
Western University
Social Sciences Centre
Room 7418
London, Ontario, Canada, N6A 5C2
Email: eheerey@uwo.ca

Abstract

Research suggests that stimuli that prime social concepts can fundamentally alter people's behavior. However, most priming studies fail to explicitly report double-blind procedures. Because experimenter expectations may influence participant behavior, we ask whether a short pre-experiment interaction between participants and experimenters contributes to priming effects when experimenters are not blind to participant condition. An initial double-blind experiment failed to demonstrate expected effects of a social prime on executive cognition. To determine whether double-blinding procedures caused this result, we independently manipulated participants' exposure to a prime and experimenters' belief about which prime participants received. Across four experiments, we found that experimenter belief, rather than prime condition, altered participant behavior. Experimenter belief also altered participants' perceptions of their experimenter, suggesting that differences in experimenter behavior across conditions caused the effect. Findings reinforce double-blind designs as experimental best practice and suggest that people's prior beliefs have important consequences for shaping interaction partner behavior.

Key Words: Social power, priming, experimenter effects

Introduction

Priming, the act of influencing another's behavior via indirect cues, is a common experimental manipulation in social psychology (e.g., Anderson & Galinsky, 2006; Dreisbach & Boettcher, 2011; Fan & Gruenfeld, 1998; Galinsky, Magee, Gruenfeld, Whitson, & Liljenquist, 2008; Overbeck & Park, 2006). In many social psychological priming paradigms, an experimenter asks participants to do a task that 'primes' or activates a particular concept, such as age or social power. The prime's effect is then examined in a subsequent task. Although participants appear to be unaware of the relationship between the prime and target-task, the prime nonetheless affects target-task performance.

Although priming appears to be one way of influencing behavior, the subtle social cues people exchange in face-to-face interactions also have powerful effects (Rosenthal, 1994). Indeed, the beliefs and stereotypes people bring to interactions shape both the behaviors they produce (Wheeler & Petty, 2001) and interaction partners' responses (Herr, Sherman, & Fazio, 1983). For example, when one interaction partner holds a stereotype about another, that partner is likely to behave more stereotypically, even when the belief holder does not intend to transmit the stereotype (Snyder & Stukas, 1999).

In research settings, experimenters' expectations may have the insidious effect of confounding task results. Indeed, research shows that when experimenters are motivated to find significant effects, they are more likely to do so (Sheldrake, 1998). Because expectations can be quite powerful, they may bias experimenter behavior when experimenters are aware of both study hypotheses and participants' conditions (Rosenthal, 1994); but see Barber (1978).

Changes in experimenter behavior may subsequently cause changes in participant behavior, independent of experimental manipulations.

Unfortunately, many papers in the priming literature do not explicitly describe double-blind experimental designs (e.g., Galinsky, Magee, Inesi, & Gruenfeld, 2006). This may be problematic for interpreting results. Indeed, several recent empirical papers have independently reported failures to replicate findings from the priming literature under double-blind conditions (e.g., Doyen, Klein, Pichon, & Cleeremans, 2012; Harris, Coburn, Rohrer, & Pashler, 2013; Pashler, Coburn, & Harris, 2012; Shanks et al., 2013) and research suggests that a failure to double-blind may be endemic (Klein et al., 2012).

For our experiments, we chose tasks thought to prime social power, defined as the ability to access, control and distribute resources within a group (Keltner, Gruenfeld, & Anderson, 2003). Power primes have been the focus of much research, with results generally indicating reliable effects (Galinsky, Gruenfeld, & Magee, 2003; Galinsky et al., 2008; Magee, Galinsky, & Gruenfeld, 2007). For example, research suggests that experimentally priming high versus low social power may improve executive cognition, including more flexible attention, reasoning, cognitive control and better ability to inhibit the influence of distractors (Galinsky et al., 2003; Guinote, 2007; Smith, Jostmann, Galinsky, & van Dijk, 2008; Willis, Rodriguez-Bailon, & Lupianez, 2011). High-power primes may also enhance abstract thinking, risk taking, approach behavior and optimism (Anderson & Galinsky, 2006; Maner, Gailliot, Butz, & Peruche, 2007; Smith & Trope, 2006). However, none of these experimental reports explicitly describes a double-blind design. If experimenters were aware of both participants' conditions and research hypotheses, they may have inadvertently altered their behavior based on this knowledge, thereby communicating expectations to participants.

In Experiment 1 we used a computer-administered role-play task to assign participants to low (“employee”) versus high power (“boss”) conditions in a double-blind design. Our aim was a conceptual replication of work demonstrating that high- versus low-power roles enhance the ability to inhibit distractors during target detection (Guinote, 2007). Despite robust effects of the power manipulation, we failed to find evidence of the expected power-priming effect on a flanker task.

However, we worried that our double-blinding procedure, which diverges from typical experimental reports in this area, might explain our failure to find predicted results. We therefore sought to manipulate both priming condition and experimenter belief about priming condition simultaneously. In each of four independent experiments, involving 11 experimenters and a total of 824 participants, we used a computerized version of a common priming task to activate feelings of high or low social power while independently manipulating experimenter knowledge about participant condition and therefore about expected results. Experiment 2 measured word categorization speed (Smith & Trope, 2006; Experiment 1), Experiment 3 examined risk-taking using the Columbia Card Task (Figner, Mackinlay, Wilkening, & Weber, 2009), Experiment 4 assessed abstract versus concrete categorizations of everyday behaviors (Smith & Trope, 2006; Experiment 2), and Experiment 5 approach behavior (Smith & Bargh, 2008; Experiment 2). In response to reviewer comments, the final experiment had increased experimental power and was preregistered at the Open Science Framework (Heerey & Gilder, 2016).

Experiment 1 Introduction

The goal of Experiment 1 was a conceptual replication of previous work demonstrating that priming high- versus low-power would enhance participants' ability to ignore distractors (Guinote, 2007) in a flanker task. However, given that we planned a between-subjects design, we wanted to ensure that experimenter expectations would not bias data collection. We therefore used a computerized priming task to guarantee that the experimenter was entirely unaware of prime condition prior to debriefing participants.

Experiment 1 Methods

Participants. One hundred and eighteen undergraduate psychology students participated in a study about "personality and cognition" in exchange for partial course credit and a small monetary bonus. We excluded one participant's data due to a computer failure that caused data loss on ~70% of trials. We also excluded four participants' data because they indicated suspicions about the link between the prime- and target-tasks. The final sample size was 113 participants (86 female, age: *M*=20.48, *SD*=3.850). Sample sizes were selected *a priori*, based on a power analysis (two-tailed α=.05, effect size *d*=.70, and experimental power=.80) using typical reported effect sizes (e.g., Smith & Bargh, 2008; Smith & Trope, 2006). Participants gave written consent before participating and were fully debriefed upon study completion. The University's Ethics Committee approved all procedures (likewise for Experiments 2-5 below).

Experimenter. One female experimenter (TSEG) completed all the data collection for this study as part of a PhD thesis. The experimenter had read and discussed the power and executive cognition literature with several collaborators. The experimenter believed she was extending findings in the power-priming literature to a flanker task, and fully expected to find priming effects.

Priming task. This experiment used a strong explicit power manipulation in which participants were assigned to high-power (“boss”), low-power (“employee”), or equal status (“control”) groups for a computerized role-play task. Participants were consented and instructed in pairs to give the impression that they would be working together in the task (in reality, all participants completed the task individually). They were then shown to adjacent rooms for the experimental procedure. After this “instruction” stage of the experiment, the experimenter had no further contact with participants until debriefing.

The computer randomly assigned participants to one of two power-related roles (boss [n=37; “high power”] or employee [n=38; “low power”]) for a target-detection game. They believed they were working with the partner to earn bonus money in the game. For comparison, a third group of participants was assigned to a cooperative “control” condition (n=38). Because the computer assigned participants to priming conditions and administered task instructions accordingly, the experimenter was blind to condition until the debriefing phase of the study.

Although all participants completed the same game, the instructions differed depending on computer-assigned roles. Participants were told that their primary task was to press a key whenever they detected a target (colored square) on the left side of the screen (see Supplementary Materials for full detail). “Bosses” were told that, as an added responsibility of their role, they should also detect and respond to targets on the right side of the screen. Employees were told that the boss had assigned them this same duty. Participants in the cooperative condition believed they were working as a team and both partners would respond to both left and right targets. Regardless of actual performance, participants learned that together they had earned £4.98. Bosses then assigned any amount of this bonus to their employees, retaining the remainder for themselves. On average, bosses in the study behaved relatively fairly, assigning 43.98% (SD=17.75%) of the total bonus to their employees. To emphasize the power differential however, employees were told that they had been allocated 35% of the bonus. In the cooperative condition, participants were told at the task outset that they would each receive 50% of the bonus.

Following the power induction, participants completed a 4-item questionnaire to measure their sense of fairness about the task ("To what extent do you feel like the workload division was fair?" "To what extent do you feel like the bonus money was divided fairly?"), effort expended ("To what extent did you feel like you performed the task to the best of your ability?"), and power ("To what extent did you feel powerful or in control in the task?"). These questions served as the manipulation check.

Target Task. To assess power-related differences in cognitive and attentional control, participants then completed a flanker task (Eriksen & Eriksen, 1974). Participants made speeded left or right button presses to indicate the direction of a central target arrow. A pair of left- or rightward pointing arrows served as distractors. Trials began with a fixation cross for 500ms, followed by a target arrow (50% pointed left) surrounded by distractor arrows pointing in either the same (congruent; 50% of trials) or the opposite direction (incongruent). The target/flanker display remained visible for 500ms before being replaced by a blank screen until the response. Participants then saw feedback about whether they were correct (1000ms). They completed three blocks of 60 randomly ordered trials. At the end of session, the experimenter fully debriefed participants and probed them for suspicion. All participants received the same monetary bonus (£5). The experimental protocol was fully automatized using E-prime (version 1.2; Psychology Software Tools, Inc.).

Data Analysis. We calculated the proportion of correct trials and the mean reaction times (excluding error trials) for congruent and incongruent trials as a measure of the flanker effect. Because we consider the absence of an effect to be equally important as its presence, we examined these data using Bayesian ANOVAs with power condition (high, low, control) as the between-subjects variable. In Bayesian analysis, the presence and absence of an effect are evaluated with different models. Prior probability distributions for the coefficients under each model are specified. This allows us to calculate each model's marginal likelihood given the observed data. The ratio of the two models' marginal likelihoods is the Bayes factor. For model comparison, we report the Bayes factor (BF_{10}), the ratio of the probability of the observed data under the alternate model, relative to that under the null model. Note that the Bayes factor automatically penalizes for model complexity, such that in the absence of any effect, the evidence will favor the simpler over the more complex model. A $BF_{10}>1$ indicates that the evidence favors the alternate model and a $BF_{10}<1$ suggests that the evidence favors the null model. BF_{10}s ranging from 3 to 20 are considered positive evidence in favor of the alternate model, whereas BF_{10}s of .33 to .10 constitutes moderate evidence in favor of the null model (see Jarosz & Wiley, 2014). Note that we report the BF_{01} (the ratio of the probability of the observed data under the null model, relative to that under the alternate model) where evidence appears to favor the null model. We gave each of the models (e.g., null, prime condition) an equal (uninformative) prior probability. Traditional ANOVA results appear in Supplementary Materials. All Bayesian analyses were conducted using JASP (version 0.8.2, JASP Team, 2017).

Experiment 1 Results

Manipulation Check. To test the efficacy of the power prime, we conducted a set of Bayesian ANOVAs, with effort, fairness and power ratings as dependent variables and power condition as the independent variable. With respect to self-reported effort, the results were non-diagnostic (BF_{10}=1.481). That is, even though low-power participants appeared to report slightly more effort than others (Figure 1A), the data did not conclusively support either the null model or an effect of prime condition. In contrast, analyses suggested that prime condition was highly effective at influencing perceptions of both task fairness (BF_{10}=1.868x10^8) and experienced power (BF_{10}=3.745x10^5). Specifically, low power participants thought the task was less fair than other participants and felt less powerful, especially relative to high-power

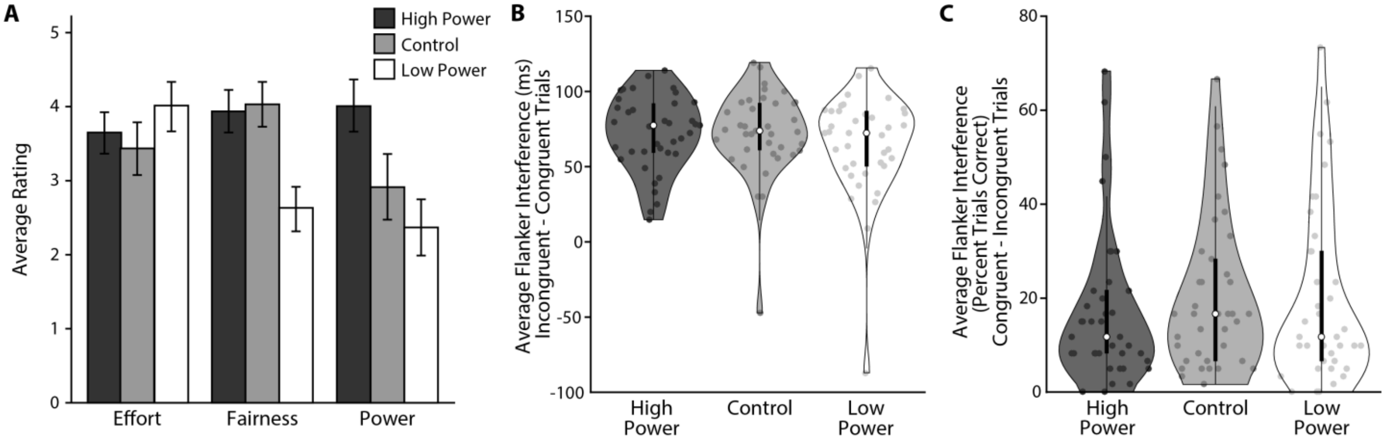


Figure 1. Experiment 1 task and results (N=113). A) Perceptions of effort, fairness and power during the task. Error bars show the 95% credible interval. Violin plots (including individual data points) of the flanker effects for B) reaction time (incongruent trials–congruent trials) and C) accuracy (congruent trials–incongruent trials) across participant conditions. The white dots indicate the median and the central boxes show the interquartile range. The whiskers show the 95%CI of the median.

participants. These results suggest that the power manipulation effectively induced feelings of high and low power. Data files for Experiment 1 are available at https://osf.io/pnvjf/ (likewise for Experiments 2-5 below).

Target Task. The reaction time (Figure 1B[1]) and accuracy (Figure 1C) results from the flanker task show strong evidence for the presence of the typical flanker effect. Participants responded both more quickly and accurately on trials with congruent versus incongruent distractors. Interestingly, when the experimenter was unaware of the power condition to which participants had been assigned, there was no indication that this effect was modulated by the prime (Speed: BF_{01}=5.952; Accuracy: BF_{01}=8.333). Thus, when the experimenter was blind to prime condition, the data were almost six times more likely (for response speed; eight times for accuracy) under the null than the prime-effect hypothesis.

Experiment 1 Discussion

Under double blind conditions, we found no evidence that power primes affected behavior in a subsequent flanker task, despite robust effects on our manipulation check. We can think of three possibilities for why this occurred. First, the flanker task has not, to our knowledge, been used with a power-prime. Nonetheless, tasks tapping similar facets of executive cognition have shown power-priming effects (Guinote, 2007; Smith et al., 2008) and the flanker task itself may be sensitive to a social status prime (Dreisbach & Boettcher, 2011). Second, although the power manipulation we used is based on previous role-play priming tasks, we did not actually ask participants to interact with an experimenter or each other as is typical

[1] Because we chose to plot individual data points, readers may note the presence of outliers in Figure 1 and other figures. Excluding these participants does not substantially change the findings (data available at https://osf.io/pnvjf/).

(e.g., Galinsky et al., 2003). However, computerization of this task was necessary to ensure that the experimenter remained blind to participants' condition. Finally, our double-blind design may have played a role in the present results. We tested this idea across four experiments.

Experiment 2-5 Introduction

In these experiments, we ask whether experimenters' knowledge of participants' priming condition might influence results, independent of participants' actual task condition. To test this question, we orthogonally manipulated experimenters' belief about which prime condition each participant experienced and the actual prime condition that a participant received. In all experiments, we used a computerized version of a power prime that has been frequently used to prime social power (e.g., Smith & Bargh, 2008; Smith & Trope, 2006). Each experiment involved an independent set of experimenters and a different target-task.

Experiment 2-5 General Methods

Experiments 2-5 all followed the same general protocol. We begin by describing this common methodology. We then describe the unique aspects and main results of each experiment, reserving manipulation check data and additional experimenter-related results for a general results section at the end. Our University Ethics Committees approved all study procedures and all participants provided fully informed consent after debriefing.

Experimenter Selection and Training. Experimenters were either Master's-level (Experiments 2-4) or Honors undergraduates (Experiment 5) who conducted the research in the context of thesis projects[2]. To ensure that they understood the literature and expected findings, they participated in journal clubs, in which they read and discussed a series of papers from the

[2] We explain our approach to ethical issues pertaining to having misled student researchers in Supplementary Materials.

relevant power priming literature. On the basis of these discussions, they developed hypotheses and selected target tasks. In all cases, they believed that they were replicating (conceptually or directly) and extending the relevant literature to account for the effects of both mood and power on their target tasks. In the context of training, they each learned a script for instructing participants (see Supplementary Methods), completed the experimental session as if they were participants, and practiced running one another on the task.

Experimenter Belief Manipulation. Each experimenter independently collected data from a sample of participants. Working from a list that ostensibly assigned participant ID codes to power-prime conditions, experimenters started the computer program before each participant arrived. After entering a participant's ID, they typed "H" for high- or "L" for low-power to start the task. They believed that this procedure caused the computer to administer the high- and low-power primes. Unbeknownst to experimenters, only half the participants completed the priming condition to which the experimenter "assigned" them. In these cases, the experimenter's belief about the prime condition and the actual prime condition were congruent, as in past research. The remaining participants completed the opposite condition to which the experimenter believed they had been assigned, meaning that the actual prime condition differed from the experimenter's belief about it.

Experimenters tested participants individually and consented and instructed them using a script (see Supplementary Materials). They also answered any questions a participant chose to ask. This procedure took about five minutes. Once participants began the computerized portion of the experimental session, they had no further contact with experimenters until debriefing.

Throughout the data collection phase of the experiment, experimenters remained blind to this manipulation. Therefore, they only had knowledge about the condition they believed participants to have completed and the results expected based on that belief. We fully debriefed experimenters at the completion of data collection and all experimenters provided informed consent for their data to be reported in this paper. None reported any suspicion about the manipulation.

Power Priming Task. The cover story maintained that that the experiments involved unrelated tasks and that we wanted to control for individual differences in participants' moods in our analyses. Participants were therefore told that they would complete a baseline mood measure before each of the tasks. Consistent with this story, the computer administered the Positive and Negative Affect Scales (PANAS, Watson, Clark, & Tellegen, 1988) before both prime and target tasks, with a randomized word order. We also embedded five power-related words into the PANAS at random points (powerless, unimportant, dominant, self-assured, influential; the first two of these were reverse-scored and the words appeared in random order). These data allowed us to measure change in feelings of power from pre- to post-manipulation and served as a manipulation check for the power-prime. Cronbach's alpha analysis showed that the set of items had moderate to good reliability (α=.729) and principle components analysis confirmed that the items loaded onto a single factor with loadings>.638. Embedding these words within the PANAS helped to conceal the nature of the experimental manipulation. Participants rated the degree to which they felt each item "right now" on a 100-point visual analogue scale using a mouse click.

After the PANAS, participants completed the power prime, a computerized version of the same 17-item scrambled-sentence priming task reported in Smith and Trope (2006; Experiment 2). On each trial, they made grammatically correct sentences by using a mouse to select and organize four of five randomly ordered words (e.g., in one item participants re-ordered four of the following words to make a sentence: “class,” “he,” “dominates,” “the,” “chooses”). In the high-power condition, half of the sentences included high-power associated words (“dominates,” “commands,” etc.) and in the low-power condition, half of the sentences contained words associated with low power (“subordinate,” “obeys,” etc.). Participants spent as long as they liked working on each sentence and could click an “undo” button if they made a mistake. Task items and word orders were identical to those in previous research (Smith & Bargh, 2008; Smith & Trope, 2006). After the second PANAS, participants completed the target task associated with their experiment.

Finally, we wanted to assess whether experimenters’ expectations altered the impressions they made on participants. To achieve this, the computer asked participants to rate the experimenter on a 7-point Likert scale (*1=not at all; 7= extremely*) after the target task. Participants responded to the prompt, “To what extent do you think the experimenter is:” and rated the experimenter on the following adjectives: attractive, competent, friendly, and trustworthy. Experimenters were unaware that participants made these ratings. The experimental protocol was fully automatized, and presented using E-prime (version 1.2 [Experiments 2-4] or 2 [Experiment 5]; Psychology Software Tools, Inc.). All participants were tested individually. At the end of the session, the experimenter returned to the room to debrief and probe each participant for suspicion about the purpose of the experiment and the relationship between the tasks using a funnel-debriefing procedure (Bargh & Chartrand, 2014).

Experiment 2: Specific Methods

Participants. One hundred and sixteen psychology undergraduates participated in a study about "cognition and mood" in exchange for partial course credit. We excluded five participants' data, based on poor English fluency (they all needed the aid of a dictionary during the target task). The final sample therefore included data from 111 participants (77 female, age: M=21.64, SD=4.44). Sample sizes sought to balance experimental power (assuming two-tailed α=.05, effect size d=.70 [e.g., Smith & Trope, 2006], and power=.80), as well as feasibility of project completion within the allotted time.

One male and one female experimenter collected data for this experiment. They believed the project was a conceptual extension of Smith and Trope's (2006; Experiment 1) findings on the effects of power and abstract thinking. They thought they were extending previous findings by examining participants' reaction times on a word categorization task (unreported in the original) and changing the priming task from a prompted writing task to our computerized scrambled sentences task.

Target Task. To measure the influence of prime and experimenter expectation on abstract thinking ability, participants then completed an English-language version of the word categorization task, reported in Experiment 1 of Smith & Trope (2006). We used the same categories/exemplars as Smith and Trope (vehicles, furniture, and clothing), presented in random order. On each trial, participants saw the category name at the top of the screen with a category exemplar below it. They rated how well they thought the exemplar fit into the category (using a 10 point scale: 0=item does not belong in this category; 9=item definitely belongs in this category; see Supplementary Figure 1A for example). Participants responded “as quickly as possible” and saw a total of eighteen exemplars in each category, six of which were weak exemplars (e.g., “feet” is a weak exemplar of a vehicle), six were moderate (e.g., “helicopter”) and six strong (e.g., “car”). The first item from a category was always a strong exemplar and remaining items appeared in random order. The experimenters believed that participants receiving the high-power prime would classify category exemplars more quickly. The dependent variable for this task was mean reaction time across all trials. We analyzed these data using Bayesian ANOVAs with experimenter belief (high, low) and prime condition (high, low) as between-subjects factors.

Experiment 2: Specific Results

As Figure 2A shows, the mean RTs for the two priming conditions appeared to be similar. Accordingly, Bayesian analysis suggested that the data were almost 5 times more likely under the null model than under the priming-effect model (BF_{01} = 4.926). In contrast, analyses showed positive evidence in favor of the experimenter-effect model, relative to the null model (BF_{10} = 3.179). Full Bayesian results for all tested models (e.g., the interaction) appear in supplementary materials (likewise for Experiments 3-5). These results provide moderate evidence for a model that included an experimenter effect, and suggest that the null model was superior to the model allowing for a priming effect. We also note that we failed to find evidence of priming effects on actual categorization ratings, contrary to the original report (see Supplementary Materials).

Experiment 3: Specific Methods

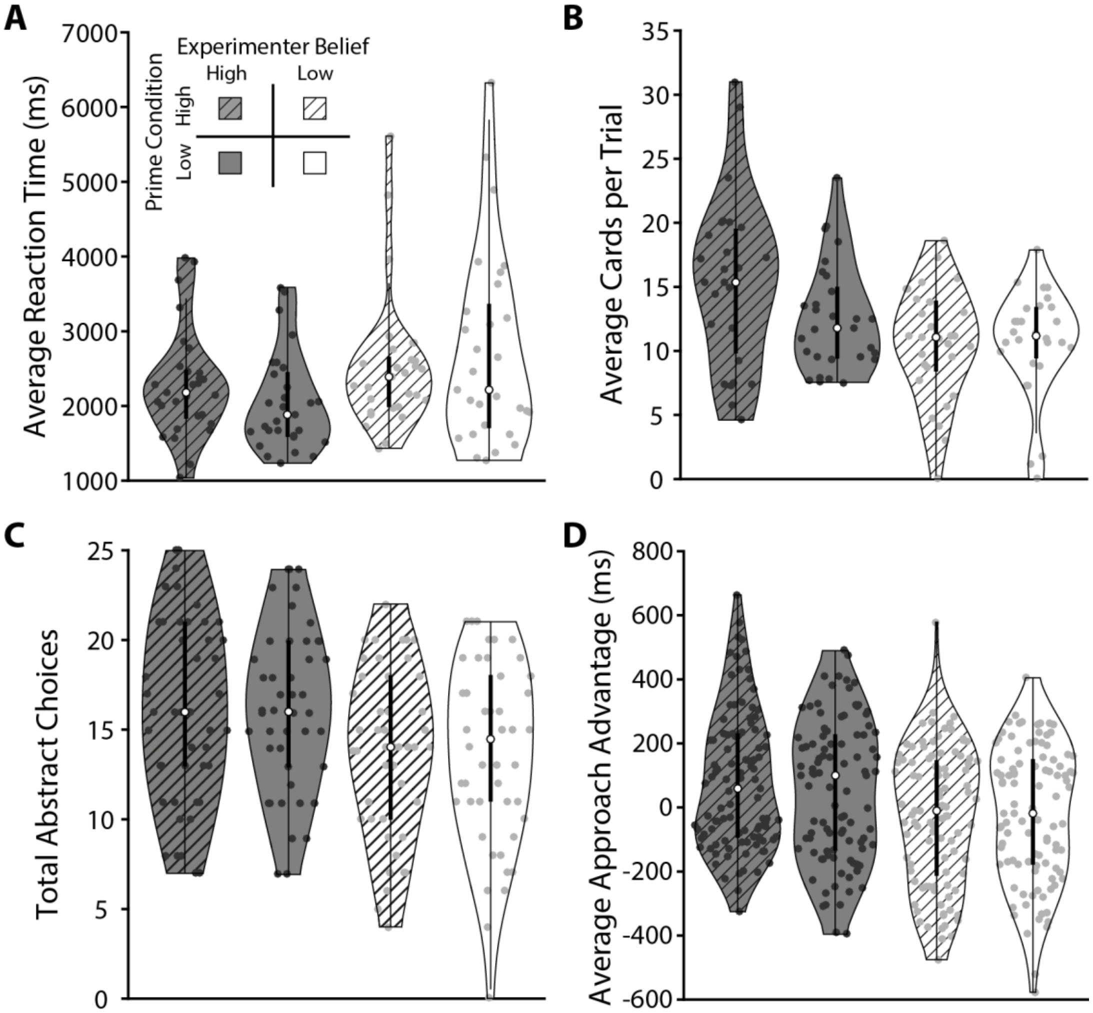


Figure 2. Results of Experiments 2 (N=111), 3 (N=110), 4 (N=179), and 5 (N=400). Violin plots (including individual data points) for A) Reaction time for exemplar classification task (Experiment 2). B) Average number of cards selected per trial on the Columbia Card Task (Experiment 3). C) Total number of abstract choices on the Behavior Identification Form (Experiment 4). D) Average "approach advantage" (avoid trials–approach trials) in the lexical decision task (Experiment 5). The white dots indicate the median and the central boxes show the interquartile range. The whiskers show the 95%CI of the median.

Participants. One hundred and ten undergraduate psychology students (66 female; age: M=21.18, SD=3.71) participated in a study about "motivation and mood" in exchange for partial course credit and a small performance-based monetary bonus. One male and one female experimenter collected the data. They believed the project was a conceptual replication of Maner, et al. (2007), in which high-power-primed participants took more risks.

Target Task. To assess risk-taking behavior, participants completed the “hot” version of the Columbia Card Task (CCT, see Figner et al., 2009). The CCT is a sequential risk-taking task, in which participants make a series of selections from a field of cards (Supplementary Figure 1B). Each field contains mostly “gain” cards (yellow happy face), for which they earn points, and up to three “loss” cards (green unhappy face) that lead to punishment if uncovered. Participants click on cards, one-at-a-time, to reveal outcomes. If the click reveals a gain card, participants earn points and may choose another card. If it reveals a punishment card, the trial immediately ends and the loss is deducted from the trial earnings. As long as no loss card has been revealed, participants may stop a trial at any time (even if they have not selected any cards). Because each selected gain card increases the ratio of loss:gain cards, each click is more risky than the previous (see Supplementary Methods for additional detail). Participants completed 27 trials of the task and received a small cash bonus equal to the number of points they earned on three randomly selected trials at the end of the experiment.

As a measure of risk-taking, we used the average number of cards selected per trial. Because the loss cards were randomly distributed in each deck, occasionally the trial ended during an early click. To ensure that these random occurrences did not influence our dependent measure, we only used trials in which participants stopped voluntarily (Figner et al., 2009).

Experiment 3: Specific Results

Although experimenters expected high-power-primed participants to engage in more risk taking, the evidence did not strongly suggest either the null model or the priming-effects model, BF_{01}=2.674. However, there appeared to be a strong influence of experimenter belief on participants’ risk-taking behavior (Figure 2B). Bayesian ANOVA indicated that the data were 25 times more likely under the experimenter-effects model than under the null model (BF_{10}=25.088; see Figure 3B).

Experiment 4: Specific Methods

Participants. One hundred eighty-one undergraduate psychology students participated in a study about “cognition and mood” in exchange for partial course credit. We excluded two female participants, one for extremely fast responding throughout the task (all RTs<200ms, suggesting that she had not read the items) and one who indicated suspicion about the prime’s relationship to the target-task. The final sample included 179 participants (151 female; age: M=20.26, SD=3.47) and 3 female experimenters.

Target Task. Experiment 4 was a direct replication of Smith and Trope’s (2006) Experiment 2 finding on abstract categorizations of everyday behavior. Participants completed the Behavior Identification Form (BIF, Vallacher, Wegner, & Somoza, 1989), which lists 25 common behaviors, each followed by two alternative descriptors for the behavior (e.g., “reading” might be classified as “following lines of print” or “gaining knowledge;” see Supplementary Figure 1C). Participants chose the descriptor that best characterized each action for them. One of the descriptors was always classified as more abstract and the other was a more concrete description of the behavior. The dependent variable was the number of abstract classifications participants made. Experimenters predicted that high-power primed participants would make more abstract categorizations than low-power primed participants.

Experiment 4: Specific Results

The data (see Figure 2C) showed positive evidence favoring the null model over the model including the priming effect, BF_{01}=6.098. As above, however, the evidence strongly supported the experimenter-belief model, relative to the null model, BF_{10}=20.760. Thus, the effects of experimenter belief appeared to be more likely than priming effects.

Experiment 5: Specific Methods

Participants. Here, we enhanced our sample size to ensure adequate statistical power in response to reviewer feedback. An *a priori* G*Power 3.1 analysis (α=.05, η^2=.04) suggested that a sample size of 400 participants was sufficient to achieve 95% power to detect a main effect of prime condition based on previously reported effects (e.g., Smith & Bargh, 2008). In exchange for partial course credit, 417 undergraduate psychology students participated in a study about “cognition and mood”. Per reviewer suggestion, this experiment (including methods, sample size, exclusion criteria, hypotheses, and analyses) was preregistered at the Open Science Framework (Heerey & Gilder, 2016) prior to data collection. To ensure that the experimenter belief manipulation remained secret, we embargoed relevant aspects of the protocol until after experimenter debriefing. Following preregistered procedures, we excluded data from 17 participants due to poor task performance (>20% of trials affected by errors, reaction times<250ms, or reaction times >a participants' grand mean +3 standard deviations) or speaking to the experimenter during the experimental session, achieving a final sample of 400 participants (291 female, age: *M*=18.480, *SD*=1.288). There were three female experimenters and one male experimenter.

Target Task. The target task was a direct replication of the lexical decision task described in Smith and Bargh (2008; Experiment 2), in which participants primed with high-power were faster to engage in approach behavior. The only difference from the published task was a switch of the words from Dutch to English. In the task, participants responded to a series of centrally presented letter strings by using a key press to move a stick figure either toward or away from the letter string, depending on whether it was an English word. Participants completed the task under one of two movement instructions. They either moved the stick figure toward words (approach direction) and away from non-words (avoid direction) or away from words and toward non-words (counterbalanced across experimenter belief and prime condition).

On each trial of the task, participants viewed a central fixation cross for 2000ms. A stick figure then appeared, centered in either the top (50% of trials) or bottom half of the screen. After an onset delay of 750ms, a central letter string appeared and remained visible until participants pressed either the up or down arrow key on the keyboard (Supplementary Figure 1D). After the key press, the stick figure moved toward the center or edge of the screen. After 750ms the next trial began. Participants were told to keep their fingers on the response keys and respond as quickly and accurately as possible.

The computer measured reaction time from the onset of the letter-string to the first key press. There were 24 trials containing English words (in a random set of 12 of these trials, the stick figure appeared above the stimulus, likewise for non-word trials) and 24 trials containing non-words (stimuli available at https://osf.io/pnvjf/). The words were rated as medium in frequency and neutral in valence based on a set of published word norms (Warriner, Kuperman, & Brysbaert, 2013) along with ratings from an independent sample of 98 local participants. Trials appeared in random order. Prior to beginning the task, participants completed 12 practice trials[3] with speed and accuracy feedback after each. There was no feedback during the task (the E-prime program used in data collection is available at https://osf.io/pnvjf/).

[3] See supplementary materials for additional notes.

Because a preliminary analysis indicated that instruction set (i.e., approach words or approach nonwords) did not moderate task results (all p-values>.406), we collapsed across this variable, as in Smith and Bargh (2008). To examine the effects of experimenter belief and prime condition, we calculated the “approach advantage” participants experienced in the task by subtracting mean approach speed from mean avoid speed (excluding error trials and trials in which the reaction time was <250ms or more than 3SDs above a participant’s mean). This preregistered performance index served as the dependent variable. Data analysis was fully automatized, such that it could not be influenced by experimenter expectations.

Experiment 5: Specific Results

Experimenters expected an “approach advantage” for high-power-primed participants. As above, evidence suggested the data were almost 8 times more likely under the null model than under the prime only model, BF_{01}=7.813 (see Figure 2D), and very strongly supported the experimenter belief only model, relative to the null model, BF_{10}=537.388. Thus, across all four of these experiments, results favored an experimenter-effects model relative to the null model, and provided moderate evidence for the null model relative to the priming-effects model. A mini-meta-analysis of our results appears in Supplementary materials.

Experiment 2-5 General Results

Manipulation Check. To ensure that the priming task activated power-related concepts, we used the power items hidden in the PANAS as manipulation check. In this case, we use frequentist analyses to describe our results to allow readers to compare the effects of our implicit power manipulation to those in previous research reports. Bayesian results appear in Supplementary Materials.

Although it is often not directly measured, reports from the power priming literature suggest that the high-power version of the scrambled sentences priming task induces greater feelings of power than the low-power version. We tested whether the prime influenced feelings of power in Experiments 2-5 using the power-related items embedded in the PANAS. We therefore examined whether prime condition influenced post-prime feelings of power using ANCOVA models with pre-prime feelings of power as the covariate.

In Experiment 2, in contrast with predictions (e.g., Galinsky et al., 2003; Smith et al., 2008; Smith & Trope, 2006), the priming task did not appear to have influenced participants' feelings of power, F(1,108)=.306, p=.581, d=.06[CI=-.22, .35] (Adjusted mean High-power=59.61[CI=56.59, 62.62]; Adjusted mean Low-power=58.42[CI=55.43, 61.41]). In Experiment 3, however, the prime condition did have a statistically significant effect on feelings of power such that participants exposed to the high-power prime felt more powerful (Adjusted mean M=61.81[CI=58.02, 65.59]) than did those exposed to the low-power-prime (Adjusted mean M=56.05[CI=52.20, 59.91]), F(1,107)=4.464, p=.037, d=.36[CI=0, .72]. We found similar results in Experiment 4, F(1,176)=5.763, p=.017, d=.18[CI=-.03, .39] (Adjusted mean High-power=56.65[CI=54.91, 58.39]; Adjusted mean Low-power=53.65[CI=51.92, 55.38]), and Experiment 5, F(1,397)=15.580, p<.001, d=.20[.06, .34] (Adjusted mean High-power=59.51[CI=57.97, 61.05]; Adjusted mean Low-power=55.12[CI=53.58, 56.67]). We note, however, that the effect sizes are small and Bayesian analyses suggest anecdotal support at best with respect to power priming effects on the manipulation check (see Supplementary Materials). Nonetheless, with the exception of Experiment 2, these effects (and effect sizes) are similar to those that have been previously reported (e.g., Galinsky et al., 2003), suggesting that the power prime here was effective in changing feelings of power.

Experimenter Effects. Although the experimenters in Experiments 2-5 achieved the empirical results they predicted based on their beliefs about participants' prime condition, they each asserted that this knowledge had not affected their behavior when instructing participants. How did experimenters transmit these effects? To examine this, we asked whether participants' ratings of experimenters depended on experimenter belief. Because people's interpersonal behavior varies dramatically depending on a variety of factors (e.g., personality, Sherman, Rauthmann, Brown, Serfass, & Jones, 2015), we had no *a priori* hypotheses about which experimenter ratings would differ or whether they would do so consistently across the set of experimenters – only that some characteristics would differ for experimenters who produced moderate experimenter effects (as noted in preregistration, Heerey & Gilder, 2016). We conducted frequentist and Bayesian ANOVAs for each experimenter using the trait ratings as dependent variables and experimenter belief as the independent variable (results appear in Table 1). For nine of the eleven experimenters, we found statistically significant effects, although not all of these reached reportable thresholds using Bayesian models.

Detailed analysis suggests that experimenters transmitted their expectations in different ways. Generally, however, when experimenters believed their participants were in the high-power versus low-power condition, they were rated as more trustworthy, often friendlier (although some experimenters were rated as less friendly), and sometimes more attractive (see Table 1). There were no differences in participants' ratings of experimenter competence across the experimenter belief conditions, meaning that it is likely that experimenters presented task instructions clearly regardless of condition.

Interestingly, the two experimenters who were not rated differently based on their beliefs about participants' priming conditions, did not produce experimenter effects on their target tasks (see Table 1). Together, these results suggest that experimenters' prior beliefs shaped participants' target-task behavior, likely via subtle changes in experimenter behavior. The two exceptions suggest that some individuals may be less susceptible to producing experimenter effects than others.

General Discussion

In Experiment 1, under double-blind conditions, we failed to find predicted effects of a social power prime on a flanker task, despite robust differences in participants' experiences of power. Results of Experiments 2-5 provide consistent evidence that experimenters, rather than prime conditions, influenced target-task outcomes, albeit inadvertently. These results show that subtly revealed expectations can shape others' behavior, and suggest that experimenters are a more powerful stimulus than many researchers, ourselves included, might care to imagine.

Of course, there are a number of possible explanations for why we failed to find priming effects, one of these being task choice. Although Experiments 4 and 5 attempted to directly replicate findings in the literature, using the same prime and target tasks (Smith & Trope, 2006, Experiment 2 and Smith & Bargh, 2008, Experiment 2), other experiments used variations on reported studies. Whereas our Experiment 2 used the same target task as Smith and Trope (2006; Experiment 1), these authors primed power with a writing exercise rather than the scrambled sentences task. Our Experiment 3 risk-taking measure has not, to our knowledge, been used in power priming, although research has found power priming effects on similar sequential risk-taking tasks (Jordan, Sivanathan, & Galinsky, 2011; Maner et al., 2007). However, if previously reported power effects are as generalizable as commonly claimed (e.g., Guinote, 2007; Maner et al., 2007; Overbeck & Park, 2006; Smith et al., 2008; Smith & Trope, 2006), the power prime should have influenced behavior on these tasks. Given that Experiments 1 and 3-5 showed expected power effects in the manipulation check, and that experimenter effects were sensitively detected in Experiments 2-5, we do not believe that task choice is responsible for our failure to replicate (conceptually or directly) previous findings.

Another difference between our methods and typical designs is that participants did the manipulation check immediately pre- and post-prime. Pilot testing suggested that this was the most reliable way to detect manipulation-related effects. However, it is possible that this procedure contributed to our failure to find a priming effect (e.g., Loersch & Payne, 2012). While additional experimentation is necessary to establish whether priming effects are observed under double-blind conditions without manipulation check, previous research has found intact priming effects immediately following a manipulation check (e.g., Storbeck & Clore, 2008). Furthermore, power-related test items were hidden within a mood measure, which itself has been shown not to influence priming results when used in this way (Smith & Bargh 2008). Finally, if this manipulation check eliminated the power-priming effect why did it not also eliminate the experimenter effect?

In contrast, our data suggest that experimenters’ expectations about task outcomes influenced participants’ performance. This influence was likely exerted via alteration of experimenter behavior, as revealed by experimenter-ratings. Although exploratory, these results suggest that effects commonly attributed to priming tasks (e.g., better executive cognition, increased risk-taking) might be caused by inadvertent differences in non-double-blind-experimenters' behavior. We therefore believe that this set of findings clearly demonstrates the need for double blind designs, insofar as this is possible, and explicit measurement of experimenter behavior where it is not.

Note that we do not claim that these results invalidate priming research generally, as they do not show that priming tasks must fail under double-blind conditions. Indeed, reports suggest that priming may work when no experimenter is present (e.g., online, Scholl & Sassenberg, 2015). However, our results do reveal a consistent and unexpectedly powerful influence of experimenter belief communicated during a scripted 5-minute interaction. These results suggest that research reports should be regarded skeptically unless authors explicitly report strong double blinding, such that it is impossible for experimenters to become aware of participants' conditions during data collection.

More broadly, our findings suggest that one person's behavior in a social interaction may depend strongly on interaction partner beliefs. For example, people's expectations may shape both their own behavior and their responses to others (Snyder & Stukas, 1999). Interaction partners may use behavioral cues to infer another's expectations, thereby allowing themselves to be "nudged" toward a particular behavior or outcome (Miller & Turnbull, 1986). At a societal level, this result has important implications for understanding how self-fulfilling prophesies arise. For example, teachers may inadvertently favor male students in mathematics and female students in English, leading to gender differences in literacy and numeracy (Nguyen & Ryan, 2008). Thus, these results suggest that understanding the interdependence between social partners' beliefs and behaviors may be important in understanding some intergroup and interpersonal conflicts that arise.

Despite its broad implications, this work has limitations. Because experimenters were exploring priming effects using predictions from the literature, we did not attempt to directly manipulate experimenters' prior beliefs (e.g., inducing experimenters to believe that a high-power prime would impair abstract-thinking ability), although previous research shows that directly altering experimenter beliefs has a similar effect (Doyen et al., 2012). Additionally, we were unable to explicitly examine the specific behaviors that changed experimenter ratings, as we could not directly observe experimenters without alerting them to the manipulation.

Conclusions. These experiments have two important implications. First, they suggest that in order to ensure the integrity of research outputs, authors should carefully consider the potential for experimenter effects during the study design process and take action to prevent these effects (e.g., video-based participant instruction). Second, these findings suggest that people's beliefs about their interaction partners or about the outcomes of their interactions exert a powerful influence on both interaction-level processes and interaction partners' subsequent behavior. Thus, people's beliefs, stereotypes, and expectations may determine the nature, quality and outcomes of their interactions.

Author Notes

We wish to thank the students who served as experimenters, along with the Welsh Institute of Cognitive Neuroscience for financial support. TG and EH were jointly involved in study conceptualization, data collection and analysis. TG and EH jointly wrote the first draft of the manuscript. EH programmed the computer tasks and supervised all students. Neither author has conflicts of interest to report.

References

Anderson, C., & Galinsky, A. D. (2006). Power, optimism and risk-taking. *European Journal of Social Psychology, 36*, 511-536.

Barber, T. X., & Rubin, D. (1978). Interpersonal expectancy effects: the first 345 studies. *Behav Brain Sci, 3*, 377-415.

Bargh, J. A., & Chartrand, T. L. (2014). The mind in the middle: A practical guide to priming and automaticity research. In H. Reis & C. M. Judd (Eds.), *Handbook of Research Methods in Social and Personality Psychology* (2nd ed., pp. 311-344). New York, NY: Cambridge University Press.

Doyen, S., Klein, O., Pichon, C. L., & Cleeremans, A. (2012). Behavioral priming: it's all in the mind, but whose mind? *PLoS One, 7*(1), e29081. doi:10.1371/journal.pone.0029081

Dreisbach, G., & Boettcher, S. (2011). How the social-evaluative context modulates processes of cognitive control. *Psychol Res, 75*, 143-151.

Eriksen, B., & Eriksen, C. W. (1974). Effects of noise letters upon the identificaiton of a target letter in a nonsearch task. *Perception and Psychophysics, 16*(1), 143-149.

Fan, E. T., & Gruenfeld, D. H. (1998). When needs outweigh desires: The effects of resource interdependence and reward interdependence on group problem solving. *Basic and Applied Social Psychology, 20*, 45-56.

Figner, B., Mackinlay, R. J., Wilkening, F., & Weber, E. U. (2009). Affective and deliberative processes in risky choice: age differences in risk taking in the Columbia Card Task. *J Exp Psychol Learn Mem Cogn, 35*(3), 709-730. doi:10.1037/a0014983

Galinsky, A. D., Gruenfeld, D. H., & Magee, J. C. (2003). From power to action. *J Pers Soc Psychol, 85*(3), 453-466. doi:10.1037/0022-3514.85.3.453

Galinsky, A. D., Magee, J. C., Gruenfeld, D. H., Whitson, J. A., & Liljenquist, K. A. (2008). Power reduces the press of the situation: implications for creativity, conformity, and dissonance. *J Pers Soc Psychol, 95*(6), 1450-1466. doi:10.1037/a0012633

Galinsky, A. D., Magee, J. C., Inesi, M. E., & Gruenfeld, D. H. (2006). Power and perspectives not taken. *Psychol Sci, 17*(12), 1068-1074. doi:10.1111/j.1467-9280.2006.01824.x

Guinote, A. (2007). Power affects basic cognition: Increased attentional inhibition and flexibility. *J Exp Soc Psychol, 43*, 685-697.

Harris, C. R., Coburn, N., Rohrer, D., & Pashler, H. (2013). Two failures to replicate high-performance-goal priming effects. *PLoS One, 8*(8), e72467. doi:10.1371/journal.pone.0072467

Heerey, E. A., & Gilder, T. S. E. (2016, August 26, 2016). Mood changes in power priming. Retrieved from osf.io/c4qnz

Herr, P. M., Sherman, S. J., & Fazio, R. H. (1983). On the consequences of priming: Assimilation and contrast effects. *J Exp Soc Psychol, 19*, 323-340.

Jarosz, A. F., & Wiley, J. (2014). What are the odde? A practical guide to computing and reporting Bayes factors. *Journal of Problem Solving, 7*(1), 2-9. doi:10.7771/1932-6246.1167

Jordan, J., Sivanathan, N., & Galinsky, A. D. (2011). Something to lose and nothing to gain: The role of stress in the interactive effect of power and stability on risk taking.

Administrative Science Quarterly, 56(4), 530-558.

JASP Team (2017). JASP (Version 0.8.2) [Computer Software]

Keltner, D., Gruenfeld, D. H., & Anderson, C. (2003). Power, approach, and inhibition. *Psychol Rev, 110*(2), 265-284.

Klein, O., Doyen, S., Leys, C., Magalhaes de Saldanha da Gama, P. A., Miller, S., Questienne, L., & Cleeremans, A. (2012). Low Hopes, High Expectations: Expectancy Effects and the Replicability of Behavioral Experiments. *Perspect Psychol Sci, 7*(6), 572-584. doi:10.1177/1745691612463704

Loersch, C., & Payne, B. K. (2012). On mental contamination: The role of (mis) attribution in behavior priming. *Soc Cogn, 30*(2), 241-252.

Magee, J. C., Galinsky, A. D., & Gruenfeld, D. H. (2007). Power, propensity to negotiate, and moving first in competitive interactions. *Pers Soc Psychol Bull, 33*(2), 200-212. doi:10.1177/0146167206294413

Maner, J. K., Gailliot, M. T., Butz, D. A., & Peruche, B. M. (2007). Power, risk, and the status quo: Does having authority promote riskier or more conservative decision-making? *Pers Soc Psychol Bull, 33*(4), 451-462.

Miller, D. T., & Turnbull, W. (1986). Expectancies and interpersonal processes. *Annu Rev Psychol, 37*(1), 233-256.

Nguyen, H. H., & Ryan, A. M. (2008). Does stereotype threat affect test performance of minorities and women? A meta-analysis of experimental evidence. *J Appl Psychol, 93*(6), 1314-1334. doi:10.1037/a0012702

Overbeck, J. R., & Park, B. (2006). Powerful perceivers, powerless objects: Flexibility of powerholders' social attention. *Organizational Behavior and Human Decision Processes, 99*, 227-243.

Pashler, H., Coburn, N., & Harris, C. R. (2012). Priming of social distance? Failure to replicate effects on social and food judgments. *PLoS One, 7*(8), e42510. doi:10.1371/journal.pone.0042510

Rosenthal, R. (1994). Science and ethics in conducting analyzing and reporting psychological research. *Psychol Sci, 5*(3), 127-134.

Scholl, A., & Sassenberg, K. (2015). Better know when (not) to think twice: how social power impacts prefactual thought. *Pers Soc Psychol Bull, 41*(2), 159-170. doi:10.1177/0146167214559720

Shanks, D. R., Newell, B. R., Lee, E. H., Balakrishnan, D., Ekelund, L., Cenac, Z., . . . Moore, C. (2013). Priming intelligent behavior: an elusive phenomenon. *PLoS One, 8*(4), e56515. doi:10.1371/journal.pone.0056515

Sheldrake, R. (1998). Experimenter effects in scientific research: How widely are they neglected? *Journal of Scientific Exploration, 12*(1), 73-78.

Sherman, R. A., Rauthmann, J. F., Brown, N. A., Serfass, D. G., & Jones, A. B. (2015). The independent effects of personality and situations on real-time expressions of behavior and emotion. *J Pers Soc Psychol, 109*(5), 872-888. doi:10.1037/pspp0000036

Smith, P. K., & Bargh, J. A. (2008). Nonconscious Effects of Power on Basic Approach and Avoidance Tendencies. *Soc Cogn, 26*(1), 1-24.

Smith, P. K., Jostmann, N. B., Galinsky, A. D., & van Dijk, W. W. (2008). Lacking power impairs executive functions. *Psychol Sci, 19*(5), 441-447. doi:10.1111/j.1467-9280.2008.02107.x

Smith, P. K., & Trope, Y. (2006). You focus on the forest when you're in charge of the trees: power priming and abstract information processing. *J Pers Soc Psychol, 90*(4), 578-596. doi:10.1037/0022-3514.90.4.578

Snyder, M., & Stukas, A. A. (1999). Interpersonal processes: The interplay of cognitive, motivational and behavioral activities in social interaction. *Annu Rev Psychol, 50*, 273-303.

Storbeck, J., & Clore, G. L. (2008). The affective regulation of cognitive priming. *Emotion*, *8*(2), 208–215.

Vallacher, R. R., Wegner, D. M., & Somoza, M. P. (1989). That's easy for you to say: action identification and speech fluency. *J Pers Soc Psychol, 56*(2), 199-208.

Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal and dominance for 13,915 English lemmas. *Behav Res Methods, 45*, 1191-1207.

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: the PANAS scales. *J Pers Soc Psychol, 54*(6), 1063-1070.

Wheeler, S. C., & Petty, R. E. (2001). The effects of stereotype activation on behavior: A review of possible mechanisms. *Psychol Bull, 127*(6), 797-826.

Table 1: Individual Experimenter Effects

Experimenter	**Trait**	**Belief Condition**		**F** **(BF$_{10}$)**	**p**	**Cohen's d** **(CI)**
1		**Low** Mean (CI)	**High** Mean (CI)			
N=60 Experimental Effect size: Cohen's d=.516 (-.01, 1.04)	**Attractive**	4.03 (3.45; 4.62)	3.63 (3.05; 4.22)	.928 (.387)	.339	-.253 (-.79; .36)
	Competent	5.77 (5.32; 6.21)	5.73 (5.29; 6.18)	.011 (.264)	.916	-.034 (-.51; .35)
	Friendly	6.17 (5.73; 6.61)	5.27 (4.83; 5.71)	8.386 (7.799)	.005 *	-.760 (-1.30; -.48)
	Trustworthy	5.50 (5.02; 5.98)	5.30 (4.82; 5.78)	.342 (.303)	.561	-.154 (-.63; .32)
2						
N=51 Experimental Effect size: Cohen's d=.486 (-.09, 1.06)	**Attractive**	4.07 (3.41; 4.74)	4.54 (3.83; 5.25)	.934 (.412)	.339	.277 (-.37; .96)
	Competent	5.67 (5.19; 6.14)	5.08 (4.58; 5.59)	2.869 (.900)	.097	-.490 (-1.00; -.05)
	Friendly	5.78 (5.32; 6.23)	4.71 (4.23; 5.19)	10.530 (16.732)	.002 *	-.929 (-1.50; -.59)
	Trustworthy	4.81 (4.29; 5.34)	4.83 (4.28; 5.39)	.002 (.281)	.961	.015 (-.51; .54)
3						
N=60 Experimental Effect size: Cohen's d=.492 (-.04, 1.00)	**Attractive**	5.23 (4.84; 5.63)	5.60 (5.20; 6.00)	1.706 (.535)	.197	.347 (0; .77)
	Competent	5.70 (5.30; 6.10)	5.70 (5.30; 6.10)	<.001 (.262)	1.000	0 (-.37; .41)
	Friendly	5.47 (5.07; 5.86)	6.17 (5.77; 6.56)	6.303 (3.451)	.015 *	.659 (.39; 1.14)
	Trustworthy	5.17 (4.79; 5.54)	5.37 (4.99; 5.74)	.569 (.333)	.454	.198 (-.11; .62)
4						
N=50 Experimental Effect size: Cohen's d=.867 (.25, 1.47)	**Attractive**	4.19 (3.67; 4.72)	5.17 (4.62; 5.71)	6.662 (3.953)	.013 *	.749 (.31; 1.33)
	Competent	5.81 (5.35; 6.27)	5.88 (5.40; 6.35)	.042 (.288)	.839	.061 (-.43; .48)
	Friendly	5.58 (5.06; 6.09)	6.04 (5.51; 6.58)	1.591 (.541)	.213	.361 (-.17; .85)
	Trustworthy	5.00 (4.53; 5.47)	5.75 (5.26; 6.24)	4.919 (2.024)	.031 *	.640 (.21; 1.14)

Table 1: Individual Experimenter Effects (CONTINUED)

Experimenter	Trait	Belief Condition		F (BF_{10})	p	Cohen's d (CI)
		Low Mean (CI)	**High** Mean (CI)			
5						
N=60 Experimental Effect size: Cohen's d=.759 (.22, 1.29)	**Attractive**	4.10 (3.54; 4.66)	4.20 (3.64; 4.76)	.063 (.269)	.802	.066 (-.48; .62)
	Competent	5.87 (5.42; 6.31)	5.90 (5.46; 6.35)	.011 (.264)	.916	.025 (-.34; .52)
	Friendly	3.27 (2.75; 3.78)	4.10 (3.59; 4.61)	5.273 (2.291)	.025 *	.600 (.01; .10)
	Trustworthy	4.53 (4.08; 4.99)	5.30 (4.85; 5.76)	5.697 (2.712)	.020 *	.628 (.14; 1.03)
6						
N=60 Experimental Effect size: Cohen's d=.520 (-.01, 1.05)	**Attractive**	4.07 (3.42; 4.72)	4.26 (3.63; 4.88)	.177 (.283)	.676	.111 (-.47; .78)
	Competent	6.28 (5.92; 6.64)	6.03 (5.68; 6.38)	.942 (.280)	.336	-.262 (-.64; .04)
	Friendly	3.69 (3.17; 4.21)	4.48 (3.98; 4.98)	4.897 (1.970)	.031 *	.577 (.04; 1.02)
	Trustworthy	4.72 (4.28; 5.17)	5.39 (4.96; 5.82)	4.594 (1.744)	.036 *	.570 (.12; .97)
7						
N=59 Experimental Effect size: Cohen's d=.212 (-.31, .74)	**Attractive**	4.47 (3.94; 4.99)	4.55 (4.02; 5.08)	.052 (.270)	.820	.057 (-.43; .60)
	Competent	5.30 (4.83; 5.77)	5.62 (5.14; 6.10)	.909 (.387)	.344	.252 (-.19; .74)
	Friendly	3.53 (3.00; 4.07)	3.48 (2.94; 4.02)	.018 (.210)	.894	-.035 (-.57; .49)
	Trustworthy	5.07 (4.66; 5.48)	5.24 (4.83; 5.66)	.360 (.235)	.551	.155 (-.16; .62)
8						
N=101 Experimental Effect size: Cohen's d=.517 (.12, .92)	**Attractive**	3.42 (2.95; 3.89)	3.73 (3.25; 4.20)	.850 (.306)	.359	.188 (-.27; .65)
	Competent	5.48 (5.03; 5.93)	5.47 (5.05; 5.89)	<.001 (.210)	.975	-.007 (-.42; .43)
	Friendly	4.96 (4.47; 5.45)	5.92 (5.63; 6.21)	11.463 (29.038)	.001 *	-.675 (-1.15; -.39)
	Trustworthy	5.40 (4.98; 5.82)	5.90 (5.63; 6.17)	4.076 (1.263)	.046 *	.404 (.14; .82)

Table 1: Individual Experimenter Effects (CONTINUED)

Experimenter	**Trait**	**Belief Condition**		**F (BF$_{10}$)**	**p**	**Cohen's d (CI)**
		Low	**High**			
9		Mean (CI)	Mean (CI)			
N=99[4] Experimental Effect size: Cohen's d=.560 (.15, .97)	**Attractive**	4.94 (4.47; 5.41)	5.36 (4.96; 5.76)	1.899 (.490)	.171	.279 (-.11; .74)
	Competent	6.22 (6.00; 6.45)	6.44 (6.22; 6.66)	1.832 (.476)	.179	.280 (.06; .50)
	Friendly	5.47 (5.08; 5.86)	6.12 (5.85; 6.39)	7.500 (5.514)	.007 *	.464 (.06; .85)
	Trustworthy	5.29 (4.91; 5.67)	6.24 (5.98; 6.50)	17.472 (327.991)	<.001 *	.845 (.59; 1.22)
10						
N=100 Experimental Effect size: Cohen's d=.649 (.24, 1.06)	**Attractive**	4.14 (3.83; 4.45)	4.98 (4.67; 5.29)	14.527 (101.477)	<.001 *	.768 (.46; 1.08)
	Competent	6.08 (5.82; 6.34)	6.18 (5.91; 6.45)	.288 (.240)	.593	.109 (-.16; .36)
	Friendly	5.54 (5.19; 5.89)	6.36 (6.13; 6.59)	15.550 (153.281)	<.001 *	.795 (.57; 1.14)
	Trustworthy	5.52 (5.16; 5.88)	6.14 (5.89; 6.39)	7.948 (6.661)	.006 *	.570 (.33; .93)
11						
N=100 Experimental Effect size: Cohen's d=.013 (-.38, .41)	**Attractive**	3.82 (3.30; 4.34)	3.92 (3.43; 4.41)	.079 (.218)	.779	.057 (-.42; .57)
	Competent	6.22 (5.98; 6.46)	6.10 (5.83; 6.37)	.446 (.257)	.506	-.135 (-.40; .10)
	Friendly	5.92 (5.58; 6.26)	5.66 (5.20; 6.12)	.833 (.305)	.364	-.184 (-.63; .15)
	Trustworthy	6.00 (5.71; 6.29)	5.72 (5.30; 6.14)	1.200 (.359)	.276	-.221 (-.63; .07)

Table 1: Participants' ratings of experimenters by trait, depending on experimenter belief. Experimenters 1 and 2 participated in Experiment 2; Experimenters 3 and 4 participated in Experiment 3; Experimenters 5 – 7 participated in Experiment 4; Experimenters 8 – 11 participated in Experiment 5. The effect size (Cohen's d[CI]) achieved by each experimenter depending on his/her belief about the prime condition is also reported. N=Number of participants included in analyses; CI= 95%CI. *Indicates a statistically significant difference (p<.05).

[4] See supplementary materials for additional notes.

Supplementary Online Material

Experiment 1 Power Priming Game

The game was a fast-paced task in which participants responded to colored squares, appearing (100ms duration) to either the left or right of a fixation cross. Participants made a key press whenever they saw a target (a blue square in a stream of colored squares) on the left. They responded with a different key press to a right-sided target (grey square) whenever it appeared. The computer randomly selected inter-stimulus intervals (independently for stimuli appearing to the left and right of fixation) from normal distributions with means of 1000ms (SD=300ms; left stimuli) or 2500ms (SD=500ms; right stimuli). Participants earned points for each target they detected within 500ms. The game included two, 3-minute blocks of trials, separated by a break.

Experiment 1 NHST Results

Manipulation Check. One-Factor MANOVA – Independent variable: Prime Condition [high/low/control]; Dependent variables: Effort, Fairness and Power ratings.

	df	F	p-value	Effect Size (η^2)
Effort	2,110	3.537	.032	.060
Fairness	2,110	29.751	<.001	.351
Power	2,110	20.201	<.001	.269

Target Task. One-Factor ANOVA – Independent variable: Prime Condition [high/low/control]; Dependent variable: Average Response Speed Difference (incongruent – congruent trials) or Proportion Correct Difference (incongruent – congruent trials).

	df	F	p-value	Effect Size (η^2)
Response Speed	2,110	.859	.427	.015
Proportion Correct	2,110	.454	.636	.008

Experiment 2 Script

BEFORE PARTICIPANT ARRIVES:

- Prepare study information and consent forms
- Start the computer task by entering the participant ID and the participant's condition when prompted. Enter 'H' for high-power and 'L' for low-power (the ID list states the condition to which each participant has been assigned).

AFTER PARTICIPANT ARRIVES:

Hi, welcome to the experiment. My name is _____ and I'm the experimenter for today's study. First, can you please read the study information sheet, which describes what you'll be doing in this experiment? Once you've done that, if you're happy to participate please sign the consent form on the next page.

I'll leave the room while you do that, but I'll be back in a couple of minutes, and then I'll explain what you need to do in the experiment itself. **[Leave the room while participants complete the form.]**

Ok, do you have any questions? **[Check that participants have signed the consent form and collect the completed form.]**

The first thing I'll ask you to do is to complete our demographics questionnaire. For the question that asks about your years of education, please put what year you are in University. **[Leave the room while participants complete the form.]**

[Check that participants have completed the demographics form and collect the completed form.] Now that we are about to start, can you please turn off your mobile phone? **[Wait while participant turns phone off.]** This experiment is about how subtle changes in a person's mood alter performance on a variety of cognitive tasks. To measure this, we will ask you to complete a mood inventory on the computer. Then the computer will ask you to complete a "scrambled sentences" task. In this task, you will see a set of 5 words, presented in a random order. You will need to use the mouse to select and order the words to make a grammatically correct sentence with 4 of them. The computer will give you more instructions about this just before the task.

After the scrambled sentences task, you will complete another mood inventory followed by a second task. This task is a word-rating task where you will see a word and decide how well it fits into a category. For example, if the category was 'pet' you might be asked to rate how well the example 'dog' fits into the category. You should try to rate each word as quickly and accurately as you can. The computer will give you more instructions about the task as well. The word-rating task will be followed by a questionnaire on the computer.

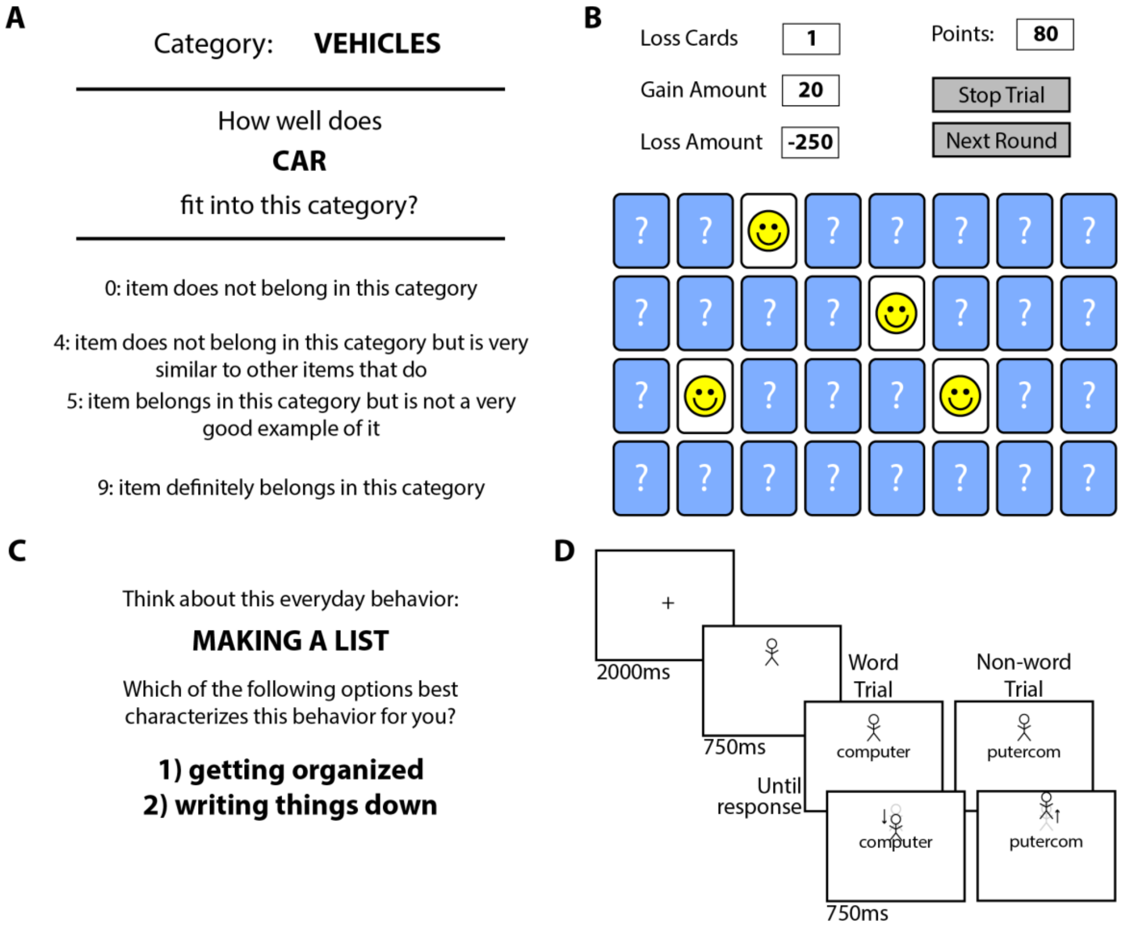
Do you have any questions? **[Answer any questions they have. Then, press the space bar to start the first mood inventory.]** Ok. Here is the first mood inventory. Click the line at the point that indicates how much of this feeling **[point to emotion word]** you are experiencing right now. **[Leave the room while participant completes the entire computer protocol.]**

Experiments 2 – 5 Ethical Considerations

Because these experiments necessarily involved deceiving experimenters, who were all honors undergraduate or master's thesis students of EAH, we undertook a thorough and careful approach to ensuring their rights and the ethical conduct of this research. Each of these experiments underwent a full ethical review. To safeguard confidentiality, an independent experimental administrator initially handled and re-labeled data files to ensure that the final data sets could not be linked to a known experimenter or participant. This kept the research team blind to experimenter identity and necessarily meant that no data were analyzed prior to the completion of data collection on a given project. Experimenters were fully debriefed at the end of data collection phase of the protocol. The main experimental participants were also debriefed at this time via email and offered the opportunity to "opt-out" of the experiment using a web-based survey if they chose (none did so). Experimenters were all offered the opportunity to provide fully informed consent after debriefing. To ensure that they felt free to make whichever consent decision they wanted without repercussion, experimenters did this via an independently administered survey, that was opened only after they had completed their final coursework and prior to the final posting date for course grades. Thus, although they did not know what course marks they had received, final marks had already been submitted to the University registrar and could no longer be altered. No experimenter declined consent, but had one done so, the independent administrator would have kept his/her identity secret from the research team.

Experiments 2 – 5 Task Designs

Supplementary Figure 1 shows example task screens for the target tasks participants completed in Experiments 2-5.


A
Category: VEHICLES
How well does
CAR
fit into this category?
0: item does not belong in this category
4: item does not belong in this category but is very similar to other items that do
5: item belongs in this category but is not a very good example of it
9: item definitely belongs in this category
B
Loss Cards 1
Gain Amount 20
Loss Amount -250
Points: 80
Stop Trial
Next Round
C
Think about this everyday behavior:
MAKING A LIST
Which of the following options best characterizes this behavior for you?
1) getting organized
2) writing things down
D
+
2000ms
750ms
Word Trial
Non-word Trial
computer
putercom
Until response
computer
putercom
750ms

Supplementary Figure 1: Experiment 2-5 target tasks. A) Word-categorization task in Experiment 2; B) Columbia Card Task in Experiment 3; C) Behavior Identification Form in Experiment 4; and D) Lexical decision task in Experiment 5.

Experiment 2 Additional Analyses

Based on their discussions, it is important to note that the Experiment 1 experimenters did not believe they would replicate Smith and Trope's (2006) categorization rating finding that low-power primed participants would rate exemplars as less likely to be category members. They reasoned that empirically all the category exemplars, including the weak ones, were actually category members (see Rosch, 1975) and should be rated as such (e.g., "feet" belongs in the category "vehicle," even though it is a non-typical exemplar). Moreover, the way the categorization task works is that participants see a category and an exemplar and rate the exemplar's membership within the category. Exemplars are always associated with their true categories (e.g., the exemplar "car" is always paired with the category "vehicles" and never with "furniture"). Because it was easy to learn that items were always paired with the appropriate category and never with non-relevant categories, the experimenters thought that this might lead to inflated categorization ratings (thereby reducing group differences). As they predicted, results suggested that the null model was a more likely explanation for the data than either prime condition, F(1,107)=.084, p=.772, d=.065[CI=-.231, .331]; BF_{01}=4.717 (High Power: 7.60[CI=7.23, 7.82]; Low Power: 7.56[CI=7.35, 7.76]), or experimenter belief, F(1,107)=.827, p=.365, d=.169[CI=-.150, .410]; BF_{10}=.295 (High Power: 7.64[CI=7.44, 7.85]; Low Power: 7.51[CI=7.35, 7.71]).

Experiment 3 Columbia Card Task (additional detail)

At the start of each trial, all 32 cards were face down in a 4x8 grid arrangement. At the top of the screen, participants viewed the number of loss cards in the deck (1, 2 or 3), the cost of a loss (−250, −500 or −750) and the point-gain per win (10, 20 or 30). At the end of each trial, regardless of whether the participant had elected to stop the trial or clicked a loss card, the computer revealed all the remaining cards and the final score for that trial. In our version of the task, the three parameters (number of loss cards, loss amount and gain amount) were crossed in a factorial design, such that participants completed one trial under each of the 27 possible conditions.

Experiment 5 Additional Notes

1) Our preregistration and Smith & Bargh 2008 both suggested that we would use 10 practice trials. However, we changed this number to 12 in order to fully counterbalance the stick figure's starting position (above/below the stimulus) and word/non-word trials in order to avoid introducing bias. Although we failed to amend our preregistration document to reflect this change, we note that practice trials are not included in statistical analyses and

therefore should not affect the results.

2) Experimenter 3 initially ran 400 useable participants prior to being debriefed but advanced analyses showed that one additional participant had failed to meet performance criteria and so was removed from the dataset without replacement. This is a small deviation from our preregistration procedure in which we registered a sample size of 100 participants per experimenter. We note that due to the large sample size, the inclusion of this participant does not change our results in any meaningful way.

Null Hypothesis Significance Testing (NHST)

In order to allow readers the opportunity to compare our Bayesian analysis (as described in the text) to the more common frequentist analysis, we include a series of tables showing traditional ANOVA results for each experiment. The means and variance estimates used in these analyses are as described/shown in the text/figures. We leave the interpretation of these results to the reader's discretion.

Experiment 2

Two-Factor ANOVA: Experimenter Belief [high/low] and Prime Condition [high/low]; Dependent variable: Average classification speed.

	df	F	p-value	Cohen's d (95%CI)
Experimenter Belief	1, 107	6.170	.015	.47 (.09, .89)
Prime Condition	1, 107	.032	.859	.02 (-.35, .40)
Belief x Prime Interaction	1, 107	1.397	.240	.22 (-.15, .60)

Experiment 3

Two-Factor ANOVA: Experimenter Belief [high/low] and Prime Condition [high/low];

Dependent variable: Average number of cards selected per trial in the CCT.

	df	F	p-value	Cohen's d (95%CI)
Experimenter Belief	1,103	11.664	<.001	.64 (.25, 1.04)
Prime Condition	1,103	2.029	.157	.22 (-.16, .61)
Belief x Prime Interaction	1,103	2.253	.136	.29 (-.10, .68)

Experiment 4

Two-Factor ANOVA: Experimenter Belief [high/low] and Prime Condition [high/low];

Dependent variable: Average number of abstract choices in the BIF.

	df	F	p-value	Cohen's d (95%CI)
Experimenter Belief	1,175	10.498	.001	.49 (.19, .32)
Prime Condition	1,175	.004	.951	.02 (-.27, .16)
Belief x Prime Interaction	1,175	.036	.850	.03 (-.27, .32)

Experiment 5

Two-Factor ANOVA: Experimenter Belief [high/low] and Prime Condition [high/low];

Dependent variable: Average approach advantage in the lexical decision task.

	df	F	p-value	Cohen's d (95%CI)
Experimenter Belief	1,396	17.818	<.001	.42 (.22, .62)
Prime Condition	1,396	.312	.577	.06 (-.14, .25)
Belief x Prime Interaction	1,396	.770	.381	.09 (-.11, .28)

Additional Bayesian Analyses (Experiments 2-5)

A 2x2 between-subjects, Bayesian ANOVA produces tests of five different effects. In our case these include the null model; the model examining experimenter belief only; the model examining only the prime condition; the experimenter belief + prime condition model; and a model containing both main effects + their interaction. We only report theoretically important models in the text to conserve space. However to enable reviewers to examine our complete results, we report all models here (excluding the null model, for which BF_{10} always = 1.000).

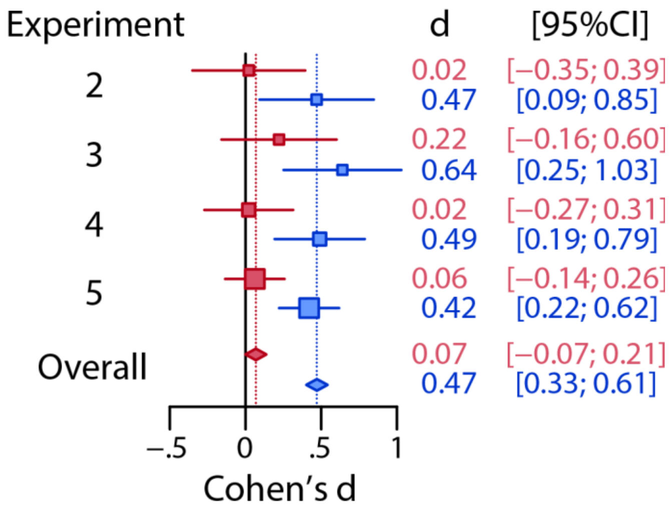
Experiment	Model	BF_{10}
2	Experimenter Belief (EB)	3.179
	Prime Condition (PC)	.203
	EB + PC	.617
	EB + PC + EBxPC	.311
3	EB	25.088
	PC	.374
	EB + PC	11.895
	EB + PC + EBxPC	8.049
4	EB	20.760
	PC	.164
	EB + PC	3.385
	EB + PC + EBxPC	.712
5	EB	537.388
	PC	.128
	EB + PC	67.662
	EB + PC + EBxPC	15.312

Because we reported frequentist analyses of our manipulation check data in the main paper, we have opted to include the Bayesian results here. In this case, we report the model results for Bayesian ANCOVAs, examining the average of the power-related items embedded within the post-prime PANAS as the dependent variable, prime condition as the independent variable and average pre-prime power ratings as the covariate.

Experiment	Model	BF_{10}
2	Covariate (Pre-prime power)	4.117×10^{11}
	Prime Condition (PC)	.202
	Covariate + PC	9.558×10^{10}
3	Covariate	27.523
	PC	1.311
	Covariate + PC	39.597
4	Covariate	6.833×10^{24}
	PC	.189
	Covariate + PC	1.515×10^{25}
5	Covariate	5.012×10^{54}
	PC	2.217
	Covariate + PC	8.726×10^{56}

Mini Meta-Analysis

For comparison purposes, we conducted a small meta-analysis on the effect sizes of Experiments 2-5[1]. We used Cohen's d as our effect size measure and conducted the analysis in r using the package "metafor" (Viechtbauer, 2010) and plotted both the experimenter and priming effects across Experiments 2-5, as well as their averages. Figure S1 shows these results. Overall, these results suggest a small but reliable effect of experimenter across the set of experiments (d=.472[CI=.331,



Supplementary Figure 2. Forest plot of reported effect sizes for effect of prime condition (plotted in red) and experimenter effects (plotted in blue).

[1] We wish to thank the editor for this suggestion.

.612], z=6.564, df=3, p<.0001) but fail to show evidence of a priming effect, at least using this scrambled sentences priming task (d=.067[CI=-.072, .206], z=.943, df=3, p=.346).

Supplementary Reference

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. Journal of Statistical Software, 36(3), 1-48. URL: http://www.jstatsoft.org/v36/i03/