Does encouraging a belief in determinism increase cheating?

Reconsidering the value of believing in free will

Thomas Nadelhoffer*, Jason Shepard*, Damien L. Crone, Jim A.C. Everett,
Brian D. Earp, & Neil Levy

* Nadelhoffer and Shepard share first authorship.

***This is a prepublication draft only. The final draft is forthcoming in *Cognition*.**

Abstract

A key source of support for the view that challenging people's beliefs about free will may undermine moral behavior is two classic studies by Vohs and Schooler (2008). These authors reported that exposure to certain prompts suggesting that free will is an illusion increased cheating behavior. In the present paper, we report several attempts to replicate this influential and widely cited work. Over a series of five studies (sample sizes of N = 162, N = 283, N = 268, N = 804, N = 982) (four preregistered) we tested the relationship between (1) anti-free-will prompts and *free will beliefs* and (2) free will beliefs and *immoral behavior*. Our primary task was to closely replicate the findings from Vohs and Schooler (2008) using the same or highly similar manipulations and measurements as the ones used in their original studies. Our efforts were largely unsuccessful. We suggest that manipulating free will beliefs in a robust way is more difficult than has been implied by prior work, and that the proposed link with immoral behavior may not be as consistent as previous work suggests.

Key words:

Free will, skepticism, moral behavior, cheating, replication

---

1. Introduction

Many scientists over the past two decades have become increasingly vocal about their skepticism concerning free will and responsibility (e.g., Bargh, 2008; Cashmore, 2010; Greene & Cohen, 2004; Harris, 2012; Haynes, 2011; Montague, 2008; Ogletree & Oberle, 2008; Wegner, 2003; 2008; cf. Baumeister, 2008; Walter, 2001). With articles covering such skepticism now regularly appearing in the popular press, the debate about free will has

captured the public imagination. One recurring theme in this public discourse is that recent scientific advances (especially in neuroscience) may threaten our traditional picture of agency and responsibility. Whether this really is the case is a complicated metaphysical question that we will not attempt to answer in this paper. Instead, we are interested in the possible ramifications of people *believing* that this is the case as free will skepticism becomes more mainstream.

One potential ramification of the expanding profile of free will skepticism is that it may negatively affect moral behavior. The best-known support for this view comes from two classic studies by Vohs and Schooler (2008) reporting that exposure to certain prompts which suggest that free will is an illusion subsequently increased the proportion of cheating behavior among study participants. As the authors wrote, "The fact that brief exposure to a message asserting that there is no such thing as free will can increase … cheating raises the concern that advocating a deterministic worldview could undermine moral behavior" (Vohs & Schooler, 2008, p. 53). Indeed, if public pronouncements concerning the death of free will or the ascendency of deterministic, mechanistic, and reductionistic explanations of human behavior have the downstream consequence of increasing the prevalence of immoral behavior, this might give free will skeptics grounds for not attempting to publicize their views.

As Vohs and Schooler acknowledge, nothing about their studies speaks to the *truth* of free will skepticism. It is possible that the skeptics are right that free will is an illusion. However, as Vohs and Schooler note, "it is possible that free will is an illusion that nevertheless offers some functionality" in encouraging moral behavior (Vohs & Schooler, 2008, p. 53). If true, then communicating specific views about free will in the public sphere may have negative consequences. At a minimum, "if exposure to deterministic messages

increases the likelihood of unethical actions, then identifying approaches for insulating the public against this danger becomes imperative" (Vohs & Schooler, 2008, p. 54).

But before we start worrying too much about the practical upshot of disbelieving in free will, we need to make sure that the findings from Vohs and Schooler are reliable. In the following section, we describe the methods and findings from Vohs and Schooler (2008) in detail. Then we provide an overview of the wider body of evidence bearing on their results. Following that, we report our own recent attempts to replicate this influential and widely cited work. As we will see, the purported relationship between disbelieving in free will and immoral behavior is more elusive than researchers have assumed.

1.1. The Vohs and Schooler (2008) Paradigm

In Study 1 from Vohs and Schooler (2008), *N* = 30 participants (distributed across two conditions) were presented with one of two excerpts from Francis Crick's *The Astonishing Hypothesis* (1994). In the Anti-FW condition, participations read an excerpt that was deterministic, reductionistic, and skeptical about free will. In the Control condition, participants read an excerpt about consciousness that did not touch on free will. After reading their respective primes, participants responded to the unpublished version of the Free Will and Determinism Scale (FAD, Paulhus & Margesson, 1994) and to the Positive and Negative Affectivity Schedule (PANAS, Watson, Clark, & Tellegen, 1988), which Vohs and Schooler used as a measure of mood.

For the second stage of the study, participants performed a computer-based mental-arithmetic task involving 20 problems after being informed that they would be paid for each problem they solved correctly. They were then told that "the computer had a programming glitch and the correct answer would appear on the screen while they were attempting to solve each problem, but that they could stop the answer from being displayed by pressing the space

bar after the problem appeared" (Vohs and Schooler, 2008, p. 50). Participants were also told that "although the experimenter would not know whether they had pressed the space bar, they should try to solve the problems honestly, on their own" (Vohs & Schooler, 2008, p. 50).

The main findings from Study 1 were as follows: (a) scores on the Free Will subscale from FAD were significantly lower in the Anti-FW condition (d = 1.20),[1] (b) scores on the other three subscales of FAD (Fate, Scientific Causation, and Chance) showed no statistically significant differences between conditions, (c) participants cheated *more frequently* in the Anti-FW condition (d = .88), and (d) mood did not differ significantly across conditions. Vohs and Schooler took these results to provide a partial answer to their initial question.

One shortcoming of Study 1, acknowledged by Vohs and Schooler, is that it relied on a passive cheating paradigm in which cheating was the default option. To overcome this limitation, Vohs and Schooler adopted an active cheating paradigm for Study 2 (see below). They also adopted a new set of prompts. For this round, they included N = 122 participants (distributed across five conditions). Rather than presenting participants with excerpts from Crick's book, they used a series of 15 statements based on the mood induction method developed by Velten (1968)—with each statement being displayed for 1 minute. There were three basic cheating-possible conditions—(a) pro-free will, (b) pro-determinism, and (c) neutral. After reading the 15 statements and being asked to think about their significance, participants in the cheating-possible conditions were once again presented with FAD and PANAS. Then, as in Study 1, participants were told they were going to take part in a second, independent experiment.

[1] Vohs and Schooler did not report effect sizes. We calculated effect sizes for their studies using their reported data.

For Study 2, Vohs and Schooler followed Mazar, Amir, and Ariely (2008) in using an active cheating paradigm. For this part of the study, participants received a set of 15 reading-comprehension, mathematical, and logical reasoning problems that were taken from the Graduate Record Examination (GRE)—a task which had been previously demonstrated to provide participants with a challenging but solvable set of problems (Schmeichel, Vohs, & Baumeister, 2003). Participants were told that that they would receive $1 for each problem they answered correctly. At this point, Vohs and Schooler used a ruse to make it possible for participants to cheat in a context where it was clear that they couldn't be caught.

In addition to this cheating-possible condition, Vohs and Schooler also ran two control conditions that did not allow for cheating. In the first non-cheating condition, participants did not read the Velten statements nor did they respond to FAD or PANAS. Instead, they simply completed the 15 item GRE task (without the ruse and without the opportunity to cheat), providing a neutral baseline for performance on the task. In the second no-cheating condition, participants received the 15 pro-determinism items before completing the 15 item GRE task. The goal of this condition was to enable Vohs and Schooler to determine whether the pro-determinism items by themselves increased performance on the GRE items, thereby controlling for a potential confound.

The results were as follows: (a) scores on the Free Will subscale were higher in the pro-free will condition than in the neutral condition (d = .51),[2] (b) scores on the Free Will subscale were lower in the pro-determinism condition than in the neutral condition (d = 1.46), (c) scores on the Scientific Causation subscale were higher in the pro-determinism condition than in the pro-free will or neutral conditions (d = .80), (d) scores on the Fate and Chance

[2] Here again, we calculated these effect sizes using their reported data.

subscales did not significantly differ by condition, (e) participants in the pro-determinism cheating condition walked away with more money than participants in the other four conditions (*d* = .84), (f) the higher people's scores on the Free Will subscale, the less they paid themselves, and (g) mood did not significantly differ across conditions. Based on these results, Vohs and Schooler concluded that free will beliefs influence moral behavior. Not only was take-home pay higher in the determinism cheating-possible condition than in all of the other cheating-possible conditions, but it was also higher than the pay in the determinism non-cheating condition. Taken together, these results suggest that free will skepticism could influence people's moral behavior.

1.2. Related Work and Recent Findings

Several studies have built upon the Vohs and Schooler (2008) results. In multiple studies, free will beliefs have been associated with increased performance in domains of value. For instance, these beliefs have been associated with higher levels of autonomy and proactivity (Alquist, Ainsworth, & Baumeister, 2013), greater self-efficacy and less perceived helplessness (Baumeister & Brewer, 2012), better academic performance (Feldman, Chandrashekar, & Wong, 2016), increased satisfaction with decision-making and choices (Feldman, Baumeister, & Wong, 2014), endorsement of dispositional explanations of behavior over situational explanations (Genschow, Rigoni, & Brass, 2017), improved feelings of belonging, self-control, and meaningfulness (Moynihan, Igou, & van Tilberg, 2007), improved learning from negative emotions (Stillman & Baumeister, 2010), and more positive ratings from work supervisors and higher job approval ratings (Stillman et al., 2010). As for the relationship between free will beliefs and explicitly moral behavior or attitudes, researchers have reported that these beliefs are associated with decreased aggression and

increased helpfulness towards strangers (Baumeister, Masicampo, & DeWall, 2009), greater epistemic and ethical humility (Earp, Everett, Nadelhoffer, Caruso, Shariff, & Sinnott-Armstrong, 2019), heightened moral judgments and attributions of blame and punishment (Krueger, Hoffman, Walter, & Grafman, 2013; Martin, Rigoni, & Vohs, 2017; Shariff et al., 2014), and more morally responsible behavior, self-control, pursuing and accomplishing goals, and conscious thought and deliberation (Stillman, Baumeister, & Mele, 2011).[3]

Although many of these findings have yet to be replicated, they do appear to bolster the earlier work by Vohs and Schooler in highlighting the various ways that believing in free will may be associated with positive traits or prosocial behavior while disbelieving in free will may be associated with negative traits and antisocial or (otherwise) immoral behavior. However, none of the studies just cited is a direct or very close replication of the work by Vohs and Schooler, which makes it hard to draw firm conclusions about the overall strength of evidence in support of their view. Instead, the above-cited studies are conceptual replications—that is, studies that depart from the original design and materials, often in substantive ways, typically to assess the generalizability of the underlying idea.

Given widespread publication bias against negative findings within psychology and the associated "file drawer" problem (Doyen et al., 2014; Earp, 2017; Greenwald, 1975; Rosenthal, 1979), this raises a problem for how the results of such studies should be taken to bear on the original findings. In a nutshell, extension studies such as conceptual replications tend to be published when they show effects that are theoretically consistent with previously published findings, but tend not to be published when they fail to show such effects (Doyen

[3] It is worth mentioning that Caspar et al. (2017) recently found that *decreased* free will beliefs were associated with *pro-social*, not anti-social, behavior. Moreover, Mercier et al. (n.d.) found that free will beliefs are associated with greater endorsement of economic inequality. These findings cut against the narrative of free will beliefs and pro-social behavior.

et al., 2014). After all, such failure can often plausibly be attributed to the methodological departure in the conceptual replication, rather than to any weakness in the original result or paradigm, leaving their results ambiguous. Therefore, in order to assess the robustness of originally reported phenomena, direct or close replications are needed (Doyen et al., 2014; LeBel et al., 2018).

Until recently, psychologists did not spend as much time as they should thinking about replication (for an overview, see Earp & Trafimow, 2015). But this has started to change in light of a series of critical papers, including a landmark publication in *Science* in 2015 by the Open Science Collaboration project (OSC, 2015). The team selected 100 studies from four high-impact psychology journals for purposes of replication, reporting an overall success rate of just 41% (see Earp, 2016, for discussion). That so few high-profile studies replicated was taken by many to indicate that the field of psychology was experiencing a crisis.[4]

One set of findings that failed to replicate were those by Vohs and Schooler.[5] At the time, the Vohs and Schooler studies were the most widely cited out of the 100 studies the OSC tried to replicate. If this were the only time researchers had a difficult time replicating these results, there might not be much cause for concern. For, just as we should not take a single original report of a finding as conclusive evidence that it exists and is reliable, nor

[4] Given that many people seem to be under the impression that the situation in psychology is uniquely bad with respect to replication, it is worth noting that apparent problems with replication seem to cut across many disciplines, with reports that 40% of economics experiments fail replication (Camerer, Dreber et al. 2016), and that only 11% of findings in haematology and oncology replicate (Belgley & Ellis 2012; note that the ongoing project on the replication of oncology research has produced much more encouraging, though still very mixed, results: Errington, Iorns et al. 2014). Conversely, in cognitive psychology, tiny p values (e.g. $p < .001$) and better replicability are not uncommon (Scholl, 2017). Some fields, such as experimental philosophy, seem to have fared better, with a 73% success rate reported by Cova et al. (2018). But there are ongoing debates about the criteria that should be used to judge whether (and in what sense) a replication attempt has been successful (for a philosophical analysis, see Earp, in press), so all of these estimates must be taken with a grain of salt.

[5] Researchers were unable to manipulate free will beliefs using the prompts used by Vohs and Schooler and they did not find a relationship between diminished free will beliefs and cheating. See Embley, Johnson, and Giner-Sorolla (2015).

should we take a single apparently unsuccessful replication attempt as invalidating the original result. Especially when using Null Hypothesis Significance Testing and p-values, a *series* of high-quality, close or direct replications is typically required to get a meaningful sense of the underlying evidence (LeBel et al., 2018; Earp & Trafimow, 2015; Trafimow & Earp, 2016; 2017). It is therefore worth emphasizing that, to date, several other attempts to conceptually replicate and extend the basic findings of Vohs and Schooler have yielded inconsistent findings—either failing to successfully manipulate free will beliefs or failing to find that successfully manipulating them consistently influences moral behavior (e.g., Crone & Levy, 2018; Monroe, Brady, & Malle, 2016; Schooler et al., 2014).[6]

So, where does that leave us? As we have seen, a number of studies have conceptually replicated and extended certain aspects of the findings by Vohs and Schooler (2008), at least when it comes to the relationship between free will beliefs and moral or prosocial behavior (of one sort or another). However, attempts to more directly or closely replicate the original findings have yielded mixed results. Even Vohs and Schooler have subsequently admitted that the relationship between free will beliefs and moral behavior is more elusive than they originally assumed (Schooler et al., 2014, p. 89). Given the outsized influence that the Vohs and Schooler findings have had—having been cited 850 times in the academic literature since 2008 and widely referenced in the popular press in venues like *The Atlantic*, *The Guardian*, *New York Times*, *The Smithsonian*, and *Scientific American* (to name a few)—we believe that these findings merit more critical attention.

2. Present Research

[6] An unpublished failed replication of Vohs and Schooler can be found at: https://rolfzwaan.blogspot.com/2013/03/the-value-of-believing-in-free-will.html

We tried to faithfully replicate the results of Vohs and Schooler (2008) with a series of five studies, four of these being both high-powered and pre-registered. For Studies 1 and 4, we used the Velten statements from Vohs and Schooler (2008) and attempted only to influence free will beliefs (not moral behavior). For Studies 2 and 3, we used the Crick statements from Vohs and Schooler (2008) and attempted to replicate the behavioral findings using an online version of the active cheating paradigm from Vohs and Schooler (2008). Finally, for Study 5, we combined the Velten statements with the active cheating paradigm. In all studies, we measured free will beliefs using FAD+ and mood using either PANAS or the Brief Mood Introspection Scale (BMIS: Mayer & Gaschke, 1988).[7] Moreover, in presenting all studies, we follow best scientific practices by reporting how our sample sizes were determined, which data were excluded and why, all manipulations, and all measures included in the study (Simmons, Nelson, & Simonsohn, 2012). Our complete data sets and our Supplemental Materials (including measures, stimuli, and further analyses that we did not have space to include in the paper) for all five studies—as well as the AsPredicted preregistration forms for Studies 2 through 5—can be found at the following OSF page: https://osf.io/d53um/.

2.1. Study 1

2.1.1. Participants and Experimental Design

This study was approved by the College of Charleston Institutional Review Board. Based on Vohs and Schooler's original sample size for Study 2 (which employed the Velten paradigm used here for Study 1) of N = 122 (roughly n = 25 per condition, given that they had 5 conditions) and based on the effect size of their Anti-FW manipulation (d = 1.46), we aimed for 225 participants—a target that, given our plan to have only 2 conditions, was in keeping

[7] While Vohs and Schooler (2008) used FAD, the unpublished version of FAD+, we opted to use the latter since it has been validated and published. When Vohs and Schooler published their findings, FAD+ had yet to be published.

with standard protocols that suggest having at least three times the original sample size for a replication study (and which gave us roughly 80% power to detect an effect of *d* = .39 or greater). Participants were recruited via the Amazon Mechanical Turk (MTurk) service and paid $1 for completing the survey. Participants were limited to those in the United States, who had a 98% successful completion rate (or higher) on MTurk and had successfully completed 500 tasks on MTurk. 221 participants completed the study. From our initial sample, 22 were excluded for failing an attention check[8] and 37 for reporting having potentially completed a similar study in the past (or failing to report whether or not this was the case).[9] This left us with a sample of *N* = 162 (58% male, 41% female, 1% other or unspecified; age: *M* = 33.97, *SD* = 10.56).

Our goal was to test one of the key conceptual and explanatory claims made by Vohs and Schooler. On their view, prompts that challenge participants' free will beliefs temporarily influence these beliefs (as measured by FAD scores). It is this temporary influence on belief—coupled with an inference on the part of participants that the justification of moral beliefs and behaviors is in some way dependent on free will or with a sense that free will is required to make a difference to how we act—that Vohs and Schooler use to explain the cheating scores. After all, if the prompts did not influence beliefs in a detectable way, how could one

[8] The attention check was randomly presented when participants were completing FAD+—which is a long stretch of similar sounding items which might induce inattention. The item we used was: "For this item, you must select "disagree." If you don't select "disagree," you will not get credit for completing this HIT." We also told participants before they read the primes that they would have to write a brief note about the content of the prime and state how difficult they found it to comprehend. We did not exclude participants based on how they responded to these items—but we included them, following Vohs and Schooler (2008), to increase the likelihood that participants paid attention. We used these same three methods for improving attention across all four studies.

[9] Previous research has shown non-naivete to undermine the effectiveness of certain experimental manipulations (Chandler, Mueller, & Paolacci, 2014; Rand, 2018). Moreover, previous research found that 30% of participants report prior experience taking part in research related to free will beliefs (Crone & Levy, 2018). So, to be on the safe side, for all four of our studies, we excluded participants who reported that they had taken a similar study in the past. We were explicit that they would be paid even if they reported having taken a similar study.

know whether it is these altered beliefs that are driving the behavioral results? Given this, the first thing we wanted to do was to simply focus on whether we could manipulate participants' scores on FAD+ with one of the prompts used by Vohs and Schooler.

Participants first completed a Velten-style free will belief manipulation (Vohs & Schooler, 2008) modelled on the original Velten mood induction procedure (Velten, 1968). In this task, participants were required to read and carefully think about 15 statements (with each statement presented for 30 seconds).[10] Participants were randomly assigned to one of two conditions (Control vs. Anti-FW), with the content of the statements varying across conditions. In the Anti-FW condition (*n* = 79), the statements were intended to undermine participants' belief in free will (e.g., "Science has demonstrated that free will is an illusion"), whereas in the Control condition (*n* = 83), the statements were simple factual statements with no relation to free will (e.g., "Oceans cover 71% of the earth's surface").

For our primary measure of free will beliefs we used FAD+ (Paulhus & Carey, 2011). As noted, FAD+ is a 27-item self-report measure of belief in free will in which participants rate the extent to which they agree with each statement on a scale from 1 (Strongly disagree) to 5 (Strongly agree) with the statements forming four subscales: Free Will (e.g., "People have complete free will"; α = .85), Scientific Determinism (e.g., "Your genes determine your future"; α = .80), Fatalistic Determinism (e.g., "I believe that the future has already been determined by fate"; α = .86), Unpredictability (e.g., "People's futures cannot be predicted"; α = .79).

[10] In Vohs and Schooler (2008) participants were presented with each statement for 1 minute. Since Vohs and Schooler did not offer an a priori justification for using the 1 minute threshold, for this study we limited exposure to 30 seconds. As we discuss below, in Study 4 we presented the Velten statements for 1 minute.

Participants also completed the 16-item BMIS (Mayer & Gaschke, 1988) to measure current mood.[11] The BMIS was administered with a prompt probing the extent to which participants felt 16 different emotions "right now." Scores for the positive-tired (α = .88) and negative-relaxed (α = .87) subscales were computed as outlined by Mayer and Gaschke.

Following Vohs and Schooler (2008), our prediction was that condition would differentially impact FAD+ scores, with scores in the Anti-FW condition being lower than scores in the Control condition. We assumed that if this prediction was not borne out, then this would cast some preliminary doubt on one of the key elements of the earlier findings by Vohs and Schooler (since they found a difference along this very dimension).

2.1.2. Results

2.1.3. Effect of manipulation on free will and related beliefs

For our primary analyses, we used the conventional alpha criterion of .05 for significance testing (as used by Vohs and Schooler, 2008). We performed an independent samples Welch t-test comparing free will belief using the free will subscale of FAD+ across conditions, failing to find evidence of an effect on free will beliefs ($t_{(156.43)}$ = 1.44, p = .153, d = -0.23). To the extent that there was any trend in the data, it was in the *opposite direction* of the hypothesized effect: participants in the Anti-FW condition (M = 3.82, 95% CI [3.66, 3.98], SD = 0.74) reported *stronger* Free will beliefs than participants in the Control condition (M = 3.66 [3.52, 3.80], SD = 0.67).

Turning to the other subscales of FAD+, we found that the Anti-FW manipulation elicited a significant increase in Scientific Determinism (Control M = 2.77 [2.63, 2.91], SD =

[11] Vohs and Schooler used PANAS as a measure of mood. We weren't confident that PANAS is a good measure for this task given that PANAS asks about mood *over the past two weeks*. So, we chose to use BMIS for Study 1—which asks participants how they feel *right now*.

0.63; Anti-FW M = 3.14 [3.00, 3.28], SD = 0.72; $t_{(154.45)}$ = 3.45, p = .001, d = -0.54), but no statistically significant effects on Fatalistic Determinism (Control M = 2.26 [2.08, 2.44], SD = 0.82; Anti-FW M = 2.47 [2.28, 2.66], SD = 0.86; $t_{(158.4)}$ = 1.63, p = .105, d = -0.26) or Unpredictability (Control M = 3.23 [3.09, 3.37], SD = 0.63; Anti-FW M = 3.22 [3.05, 3.39], SD = 0.77; $t_{(151.58)}$ = 0.04, p = .969, d = 0.01).

2.1.4. Effect of manipulation on mood

The original study by Vohs and Schooler (2008) raised the possibility that the effect of their manipulation could be mediated by mood (i.e., participants in the Anti-FW condition cheated more because they experienced heightened negative affect and/or diminished positive affect). In the online Supplementary Materials, we present analyses exploring the possibility that the effect of the manipulation on Free will beliefs was mediated by positive or negative affect, finding no support for either mediation effect.

2.1.5. Discussion

In short, our attempt in Study 1 to decrease free will beliefs with the Velten statements used by Vohs and Schooler (2008) was unsuccessful. Indeed, if anything, we found that scores on the Free Will subscale of FAD+ were slightly higher in the Anti-FW condition than in the Control condition. More puzzling still, scores on the Scientific Determinism subscale of FAD+ were also higher—which should have corresponded with lower scores on the Free Will subscale given the explanation Vohs and Schooler provided for their findings. Given this, we decided to make three changes and run a follow up study. First, we would roughly double our sample size to substantially increase power. Second, we would try the Crick paradigm rather than the Velten paradigm. Finally, we would add a behavioral task to allow for and measure cheating behavior.

2.2. Study 2

2.2.1. Participants and Experimental Design

For this preregistered study—which was approved by the College of Charleston Institutional Review Board—participants were recruited via MTurk. Participants were limited to those in the United States, who had at least a 98% success rate and had successfully completed at least 500 previous tasks on Mturk. Study 2 of Vohs and Schooler (2008) had *N* = 119 participants divided among 5 conditions. To prioritize avoidance of a Type II over a Type I error, since this is a replication attempt, we decided to dramatically increase our power relative to the original studies by Vohs and Schooler: We ran two conditions, and planned to recruit 175 participants per cell, or 350 participants overall. A sample of 350 would give us > .999 power (two-tailed) to detect an effect of the size detected by Vohs and Schooler in their study (*d* = .84).

We successfully recruited 342 participants for the study. Each participant was paid $3 for completing the study. Ten participants received a $10 bonus (see below for details). Of the participants who began the study, 20 did not complete the study, 3 failed the attention check, and 36 reported that they had previously taken or might have previously taken a similar study. We excluded these participants from all analysis, resulting in a final sample of *N* = 283 (53% male, 47% female, <%1 other or unspecified; *age*: *M* = 36.7, *SD* = 11.60 ; Anti-FW condition, *n* = 144, Control condition, *n* = 139). Based on the sample size for all participants whose data were included in the analysis, the resulting power to detect an effect the size detected by Vohs and Schooler in their study remained > .999. Given the resulting sample size, we had 80% power to detect an effect size of *d* = .34.

The purpose of Study 2 was to replicate a study from Vohs and Schooler (2008) which focused on the influence (or lack thereof) of Anti-FW prompts on people's beliefs, moods, and behaviors. Vohs and Schooler ran two studies, which involved undergraduates who

participated in their respective labs. While our replication study was conducted on-line and is thus not an exact replication, we tried our best to ensure that our study methodologically reflected their earlier work as much as possible despite some inescapable differences (see Earp, in press for theoretical support).

Participants in Study 2 were first presented with a consent form where they were informed that they would be participating in two completely independent studies—a salient feature of Vohs and Schooler's earlier studies that was not reported. Based on private conversations with Vohs and Schooler, it appears that splitting the study in this way made a difference to the results. This is a standard form of harmless deception in the experimental literature. We adopted the same strategy for Studies 2 and 3.

In Study 2, participants were told that, as part of their first study, they would be presented with a passage to read. They were also told that they would be expected to briefly summarize the passage and say how difficult it was to understand. Participants were then told that we would have them provide some information about their concepts, beliefs, and mood. This is the same approach used by Vohs and Schooler (including the same materials).

At this point, participants were told that in their second study, they would be completing an intelligence test that contains three sections and fifteen items: 5 logical reasoning puzzles, 5 verbal synonym questions, and 5 verbal sentence completion tasks. This technique—namely, telling participants that their performance on a task will be reflective of some positive personal trait such as IQ, intelligence, or life success—has been used successfully in previous cheating paradigms (see, e.g., Deinstbier et al., 1980).

We also told participants that the 15-item test was designed to take no more than 15 minutes—so, they should try not to spend more than 1 minute on any given item. We included a timer on the screen so they could track how long they had been working.

Participants were told that the 10 participants who received the highest scores would receive a $10 bonus. They were also told that in the event that more than ten participants got the highest possible score—i.e., 15 out of 15—we would select 10 people randomly from this pool to determine which participants got the reward.

For the first part of the study, participants read one of two different excerpts by Francis Crick: In the Anti-FW condition, the excerpt was a passage about free will. In the control condition, the excerpt was a passage about consciousness that did not mention free will. These are the same two prompts used by Vohs and Schooler (2008)—including a picture of Crick, a brief description of who he is, and the excerpt from his book *The Astonishing Hypothesis*. After reading their prompt, participants were asked to briefly summarize the Crick piece and state how difficult they found it to comprehend. These two steps were also part of the original Vohs and Schooler studies.

After participants read the excerpt, they filled out FAD+ (Paulhus & Carey, 2011). Reliabilities were again adequate for all four subscales (Free Will: α = .75; Scientific Determinism: α = .68; Fatalistic Determinism: α = .86; Unpredictability: α = .72). In their original studies, Vohs and Schooler next had participants fill out PANAS in an effort to make sure that their behavioral findings were produced by the specific content of the Crick prompts, rather that whatever emotions the Crick prompt may have induced in the participants. However, for the reasons we mentioned in Study 1 (namely, for this type of study, we are interested in whatever impact the manipulation has on participants' mood *at the time* not on participants' reported mood from the *past two weeks*), we once again used BMIS as a measure of current mood instead. Scores for the positive-tired (α = .74) and negative-relaxed (α = .85) subscales were computed as outlined by Mayer and Gaschke (1988).

Participants were then informed that they had completed the first study and that they would now take part in the second study. This is where we introduced participants to the aforementioned cheating paradigm involving the intelligence test. For our replication, we decided to adopt a similar active cheating paradigm (similar to Study 2 of Vohs & Schooler, 2008), because this rules out that the cheating is simply due to de-motivating effects of the Anti-FW prompt. Vohs and Schooler had participants work on 15 word, mathematical, logical, and reasoning items, taken from the GRE (which we confirmed via personal correspondence). While we do not know which 15 items Vohs and Schooler used—as they do not recollect (based on personal correspondence)—we adopted a similar paradigm which involves 15 items taken from the Law School Admissions Test (LSAT) and the GRE.

Participants in both conditions were told to do their work with pencil and paper (so we could not possibly see how well they performed). We then showed them the correct answers and gave them the chance to self-report how many items they answered correctly (which gave them an opportunity to overreport). Because they did their work with pencil and paper, they knew we had no way of knowing whether they were cheating or overreporting. Of course, this means that we did not, in fact, have any way to know whether any particular individual overreported. Instead, the collective answers for each *condition* shed light on whether people in a given condition were more likely to overreport than people in different conditions. This was a feature of the active cheating paradigm used by Vohs and Schooler (2008). Based on the findings reported by Vohs and Schooler (2008), our prediction was that participants in the Anti-FW condition would cheat more—that is, have higher self-reported scores—than the individuals in the Control condition.

2.2.2. Results

First, independent samples t-test were conducted to test whether there was a difference between the conditions on FAD+ or BMIS. There was no evidence of an effect of condition on any of the FAD+ subscales (uncorrected *p*s >.228), or on mood (uncorrected *p*s >.05). See Table 1 for a summary. Second, to test whether cheating behavior differed between Anti-FW and Control conditions, we ran an independent samples Welch t-test. With a power of .80 to detect an effect of d = .34 and a power > .999 to detect an effect of the size Vohs and Schooler detected in their original study (d = .84), we failed to find any evidence of a difference in cheating behavior between the Anti-FW (M = 9.52 SD = 3.60, 95% CI [8.93, 10.11]) and Control conditions (M = 10.34, SD = 3.73, 95% CI [9.72, 10.96]), $t_{(319.3)}$ = -1.873, p = .062, d = .18. To the extent there was a trend in the data, it was in the *opposite* direction of what was reported in Vohs and Schooler (2008). See Figure 1.

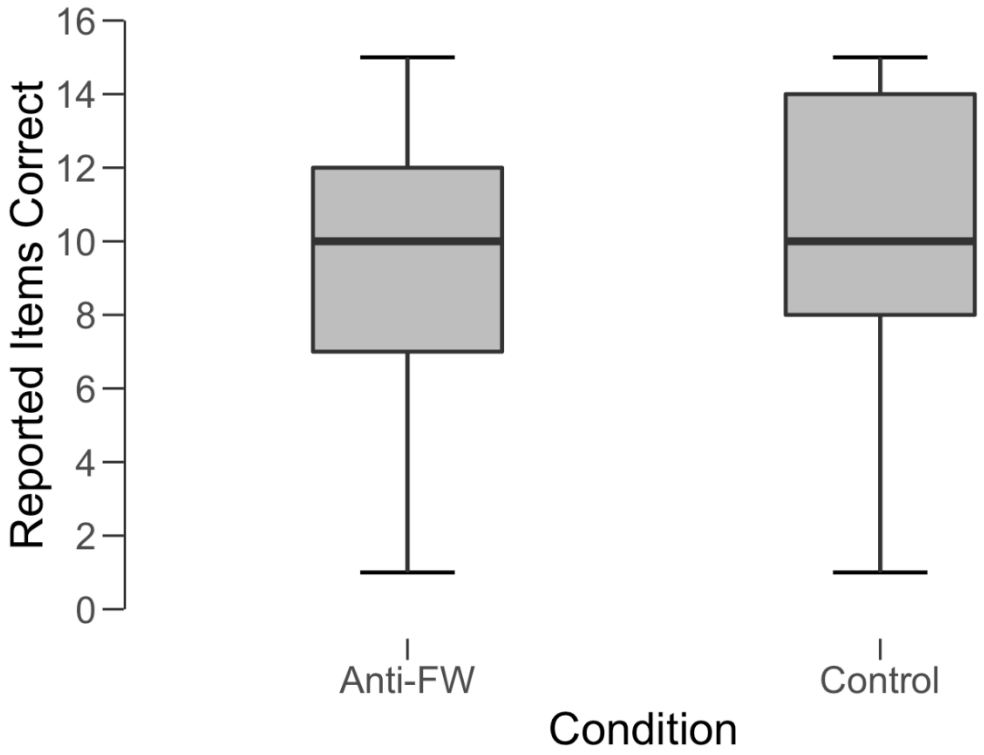
Table 1.

Independent Samples t-Test comparing Anti-FW and Control conditions on FAD+ Subscales

	t	df	p	Mean Difference	SE Difference	Cohen's d
Free Will	-0.057	280.9	0.955	-0.005	0.085	-0.01
Scientific Determinism	0.114	280.9	0.909	0.009	0.077	0.01
Fatalistic Determinism	-1.044	266	0.297	-0.112	0.108	-0.13
Randomness	-0.848	278.5	0.397	-0.069	0.082	-0.10

Note. Welch's t-test. P-values have not been corrected for multiple comparisons. Positive values indicate higher valu
condition

Figure 1. Box Plot of Reported Items Correct for Anti-FW and Control Conditions



2.2.3. Discussion

Contrary to our prediction and contrary to the original findings reported by Vohs and Schooler (2008), the Crick paradigm did not influence scores on free will beliefs and did not influence cheating behavior (if anything, the data trended in the opposite direction). As such, we failed to replicate these findings. We had one hypothesis for why we might have come up empty handed. Namely, responding to FAD+ immediately after reading the prompt might diminish the potential impact of the prompt by the time participants begin the cheating task. To test this hypothesis, we made a few basic changes to the paradigm and reran Study 2 with the goal of trying once again to replicate the earlier behavioral findings from Vohs and Schooler.

2.3. Study 3

2.3.1 Participants and Experimental Design

For this preregistered study—which was approved by the College of Charleston Institutional Review Board—participants were recruited via MTurk. Participants were limited to those in the United States, who had at least a 98% success rate and had successfully completed at least 500 previous tasks on Mturk. For Study 2 of Vohs and Schooler (2008), they had 119 participants split across 5 conditions (roughly 20 participants per cell). For Study 3, we once again ran two conditions, and planned to recruit 150 participants per cell, or 300 participants overall. This would have given us > .999 power (two-tailed) to detect an effect of the size detected by Vohs and Schooler in their study (d = .84).

We successfully recruited 302 participants for the study. Each participant was paid $3 for completing the study. Ten participants received a $10 bonus. Of the participants who began the study, 3 did not complete the study, 4 failed the attention check, and 27 reported that they had previously taken or might have previously taken a similar study. We excluded these participants from all analyses, resulting in a sample of N = 268 (58% male, 41% female, <%1 other or unspecified; *age*: M = 38.2, SD = 12.66; Anti-FW condition, n = 142, Control condition, n = 126). Based on the sample size for all participants whose data were included in the analysis, the resulting power to detect an effect of the size detected by Vohs and Schooler remained > .999. Given the resulting sample size, we had 80% power to detect an effect size of d = .35. Each was paid $3 for completing the study. Ten participants received a $10 bonus (see below for details). Participants took on average 32 minutes and 38 seconds to complete the study.

For our second behavioral study, we wanted to give the original findings a fair day in court. As noted, we thought that one possible explanation for why we failed to replicate the findings in our previous study was that giving people FAD+ and BMIS *between* reading the

prompts and participating in the cheating paradigm might dampen the influence of the prompts on cheating behavior. For this study, we once again had two conditions. The prompts were the same two Crick prompts from Study 2. However, Study 3 reversed the order of key parts of Study 2 and added one additional measure of moral belief. The order of presentation was as follows:

1. Prompt (Anti-FW vs. Control)
2. Participants answered two multiple choice questions, and then, just as in Study 2, they wrote a brief summary of the passage and described how hard they found it to comprehend.
3. Participants completed the cheating paradigm (precisely as before).
4. Participants responded to a six-item Business Ethics measure (Cooper & Pullig, 2013).[12]
5. Participants responded to FAD+.[13]
6. Participants responded to BMIS.[14]
7. Participants provided demographics.
8. Participants were debriefed, just as before.

2.3.2. Results

First, independent samples t-test were conducted to test whether there was a difference between the conditions on FAD+ or BMIS. There was no evidence of an effect of condition on any of the FAD+ subscales (uncorrected *p*s >.157), or on mood (uncorrected *p*s >.065). See Table 2 for a summary. Second, to test whether cheating behavior differed between Anti-FW condition and Control condition, we performed an independent samples Welch t-test. With a power of .80 to detect a an effect of d = .35, and power > .999 to detect an effect of the size detected by Vohs and Schooler in their original study (d = .84), we failed to find any evidence of a difference in cheating behavior between the Anti-FW condition (M = 9.78, SD = 3.32, 95%

[12] We included this so we had an indirect measure of immoral behavior.
[13] Reliabilities were again adequate for all four subscales of FAD+ (Free Will: α = .68; Scientific Determinism: α = .66; Fatalistic Determinism: α = .85; Unpredictability: α = .71).
[14] Reliabilities for BMIS positive-tired (α = .66) and negative-relaxed (α = .84) subscales were adequate.

CI [9.23, 10.33]) and the Control condition (*M* = 10.08, *SD* = 3.10, 95% CI [9.54, 10.62]), $t_{(265.3)}$ = -0.777, *p* = .438, *d* = -.10. To the extent there was a trend in the data, it was, again, in the *opposite* direction of what was found in Vohs and Schooler (2008). See Figure 2. Additionally, there was no evidence of an effect of condition on any of the six items from the Business Ethics questionnaire, *p*s > .230. See Table 3 for a summary.

Table 2.

Independent Samples t-Test comparing Anti-FW and Control Conditions on FAD+ Subscales

	t	df	p	Mean Difference	SE Difference	Cohen's d
Free Will	-1.418	266	0.157	-0.12	0.085	-0.17
Scientific Determinism	0.249	266	0.804	0.019	0.078	0.03
Fatalistic Determinism	0.800	265.7	0.424	0.086	0.107	0.10
Randomness	0.703	261.8	0.482	0.062	0.087	0.09

Note. Welch's t-test. P-values have not been corrected for multiple comparisons. Positive values indicate higher valu
condition

Figure 2. Box Plot of Reported Items Correct for Anti-FW and Control Conditions.

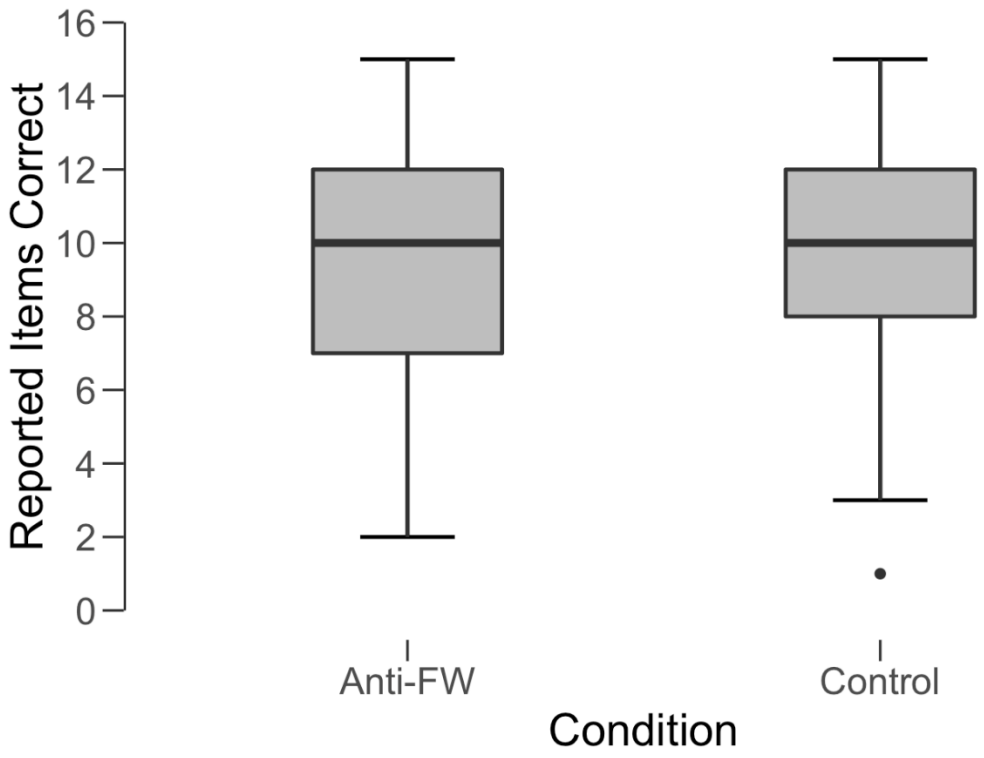


Table 3.

Independent Samples t-Test Comparing Anti-FW and Control Conditions on Business Ethics Items

	t	df	p	Mean Difference	SE Difference	Cohen's d
Paid expense account	-0.0458	254.9	0.647	-0.056	0.121	-0.01
Quid pro quo	-0.652	262.5	0.515	-0.137	0.210	-0.08
Nepotism	0.087	262.5	0.931	0.017	0.192	0.01
Misleading marketing	1.204	265.8	0.230	0.212	0.176	0.15
Violate company policy	0.794	262.3	0.428	0.12	0.151	0.10
Environmental harm	-0.296	261.8	-0.064	-0.064	0.217	-0.04

Note. Welch's t-test. P-values have not been corrected for multiple comparisons. Positive values indicate higher valu
condition

2.3.3. Discussion

Despite our best efforts, we once again failed to replicate the earlier findings by Vohs and Schooler (2008). First, the Crick prime did not influence free will beliefs. Second, participants in the Anti-FW condition were no more likely to cheat than participants in the Control condition. Finally, scores on a measure of moral beliefs were no different between the two conditions. Given that previous research suggests that it can be difficult to replicate these findings, we take our findings to cast further doubt on the claim that challenging free will beliefs in the way done by Vohs and Schooler (2008) increases cheating behavior (as they operationalized this) in a reliable manner.

However, before throwing in the towel in our efforts to replicate the findings from Vohs and Schooler (2008), we thought we would run another study that might correct for some potential explanations for why we failed to successfully manipulate free will beliefs in our first three studies. For instance, one reason the prime we used in Study 1 may not have influenced free will beliefs is that we only exposed participants to the Velten statements for 30 seconds rather than 1 minute, which was the approach used by Vohs and Schooler (2008). Another possible explanation is that while our studies were very highly powered relative to the original studies by Vohs and Schooler, they might still have been underpowered to detect an effect that is smaller than what was detected by Vohs and Schooler (2008), but still potentially of scientific interest. So, we decided to run a preregistered follow-up study that addressed these two concerns. Also, rather than departing from Vohs and Schooler in using BMIS to measure mood as we did in Study 1, for Study 4 we decided to use a modified version of PANAS that indexed responses to present and not past emotions and mood.

2.4. Study 4

2.4.1. Participants and Experimental Design

For this preregistered study—which was approved by the College of Charleston Institutional Review Board—a power analysis using G*Power suggested that with two experimental conditions, we would need a minimum of 858 participants to detect a small effect of d = .20, given an alpha of .05 and power of .90 for a one-tailed significance test. Note that we decided to use a one-tailed, rather than two-tailed, significance test so as to maximize the chance of a successful replication. We decided in advance that d = .20 would be the smallest effect size of interest (even though the effect size in Vohs and Schooler was much larger).[15] We recruited 967 participants via MTurk (compared to N = 30 in the original Vohs and Schooler study using the Velten paradigm, that is, a more than 30 times greater sample size).

Participants were paid $1 for completing the survey and were limited to those in the United States who had at least a 98% successful completion rate on MTurk and had successfully completed 500 tasks. As outlined in the pre-registration, of these participants, 35 were excluded for failing an attention check, 123 for reporting having potentially completed a similar study in the past, and 5 for having missing data on one or more of the key outcome measures (the four FAD+ subscales or the FW slider) or PANAS. This left us with a final sample of N = 804 for our primary analyses (53% male, 46% female, 1% other or unspecified; age: M = 38.88, SD = 12.07).

Since we did not have any success manipulating free will beliefs in Studies 2 and 3 using the Crick paradigm despite having what we took to be adequate power, we thought it made sense to try a higher-powered study using the 15-item Velten paradigm for Study 4. Participants were randomly assigned to one of two conditions: the Control condition (n = 375)

[15] See, e.g., http://daniellakens.blogspot.com/2017/05/how-power-analysis-implicitly-reveals.html

or the Anti-FW condition (*n* = 429). The manipulation was identical to Study 1 with the exception that each statement was displayed for 60 seconds rather than 30, following the original methods of Vohs and Schooler (2008).

As in Study 1, we measured free will beliefs using FAD+ (Paulhus & Carey, 2011). Reliabilities were again adequate for all four subscales (Free Will α = .85; Scientific Determinism α = .78; Fatalistic Determinism α = .85; Unpredictability α = .81).

As an additional measure of free will beliefs, we included a single-item slider in which participants rated their agreement with the statement "I have free will" on a slider with values ranging from 0 (Strongly disagree) to 100 (Strongly agree).

Participants also completed a modified version of PANAS (Watson, Clark, & Tellegen, 1988) to measure *current* mood. The PANAS was administered with a modified prompt geared towards enabling us to explore the extent to which participants felt 20 different emotions *right now*. These 20 items were then used to create measures of positive affect (PA; α = .92), and negative affect (NA; α = .91).

2.4.2. Results

2.4.2.1. Effect of manipulation on free will belief

As outlined in the pre-registration, we performed one-tailed significance tests for the effects of the manipulation on free will beliefs as measured by the slider and the free will subscale of FAD+. The manipulation produced statistically significant decreases in both: slider β = -0.41, 95% CI [-0.56, -0.28], one-tailed $p < .001$; FAD+ β = -0.14 [-0.28, -0.00], one-tailed $p = .045$. Please see the following footnote for an alternative way of reporting these analyses using

conventional t-tests, accompanied by effect size estimates expressed as Cohen's *d*.[16] Notably the effect sizes are close to zero, and thus far smaller than what we had pre-designated as the "smallest effect size of interest."

2.4.2.2. Effect of manipulation on related beliefs

We also performed a set of parallel analyses on the effect of the manipulation on the three other subscales of FAD+ (with these significance tests being two-tailed). We found no effect of the manipulation on Fatalistic Determinism (β = 0.03 [-0.10, 0.17], p = .642) or Unpredictability (β = 0.05 [-0.09, 0.19], p = .474); however, we did observe a statistically significant, but very small in terms of effect size, increase in the endorsement of Scientific Determinism (β = 0.15 [0.02, 0.29], p = .028).[17]

2.4.2.3. Exploratory test for potential mediation by mood

As in Study 1, we tested separate mediation models for each of the four combinations of free will beliefs (FAD+ and slider) and mood (PA and NA) measures, finding neither direct effects of condition on mood, nor indirect effects of condition on free will beliefs via mood. These models are summarized in the online Supplementary Materials.

2.4.3. Discussion

[16] Because our analysis plan resulted in regressing our outcome measures on a single binary variable, we note here that it would have been more conventional to perform t-tests, so we report these here (with the results being identical aside from the slightly different degrees of freedom calculation in the Welch t-test). For the FAD+ measure of free will belief, participants in the Anti-FW condition (M = 3.69, SD = 0.70) scored slightly but significantly lower than those in the Control condition (M = 3.78, SD = 0.70), $t_{(913.89)}$ = 1.99, p = .023 (one-tailed), d = -0.13. Similarly, for the slider measure, participants in the Anti-FW condition (M = 72.93, SD = 25.93) scored significantly lower than those in the Control condition (M = 82.91, SD = 20.78), t(913.98) = 6.49, p < .001 (one-tailed), d = -0.43.

[17] Again, for completeness, the results of Welch t-tests addressing the same questions were as follows. For Scientific Determinism, participants in the Control condition (M = 2.98, SD = 0.64) scored significantly lower than those in the Anti-FW condition (M = 3.08, SD = 0.69), $t_{(923.97)}$ = 2.33, p = .020, d = 0.15. For Fatalistic Determinism, there were no significant differences between conditions (Control M = 2.47, SD = 0.86; Anti-FW M = 2.53, SD = 0.84; $t_{(909.9)}$ = 1.05, p = .293, d = 0.07). Similarly for Unpredictability, there were no significant differences between conditions (Control M = 3.36, SD = 0.64; Anti-FW M = 3.39, SD = 0.68, $t_{(921.77)}$ = 0.70, p = .486, d = 0.05).

Our goal with Study 4 was to once again try to replicate the impact of primes used by Vohs and Schooler (2008) on free will beliefs. Given the two main modifications we made compared to Study 1—changing exposure from 30 seconds to 1 minute and making sure our study was very highly powered—we were able, with a generous one-tailed t-test, to decrease free will beliefs (as measured by the Free Will subscale of FAD+ and a new Free Will slider) using the Velten paradigm that was used in one of the original studies by Vohs and Schooler. Moreover, mood did not seem to explain these results. However, the Cohen's effect size estimate for the decreases in FAD+ scores, while significant, was only *d* = -0.13, which is far smaller than the pre-designated "smallest effect size of interest" (*d* = .20) and incomparably smaller than the original effect size detected by Vohs and Schooler (*d* = .84). Thus, we are not sure what to make of the findings from Study 4. While we were finally able to detect at least some movement in free will beliefs as a result of administering an Anti-FW prompt used by Vohs and Schooler (2008), our "sledgehammer" approach (very large sample size, one-tailed t-test, and so on) yielded effects so small in size it is hard to know what their practical or theoretical importance could be. To put this another way, if Vohs and Schooler (2008) were able to manipulate both free will beliefs and moral behavior with roughly 20 participants per cell—using many of the very same materials we used in our studies—it is surprising that we should need 400+ participants per cell to find any measurable effect of the manipulation.

All told, it seems much more difficult to manipulate free will beliefs than a number of researchers have suggested in the wake of Vohs and Schooler's original findings. However, as we were technically able to manipulate such beliefs (by some amount) using the Velten paradigm with a very large sample, we wanted to run a similarly high-powered follow-up study that added back in the cheating paradigm we used in Studies 2 and 3. After all, the best way of testing the purported relationship between decreased free will beliefs and increased

cheating behavior requires that we first successfully manipulate free will beliefs (which we were unable to do in Studies 1 through 3). In other words, we wanted to close the loop between manipulated free will beliefs and behavior, now that we had found a way to measurably achieve the manipulation, however weak.

2.5. Study 5

2.5.1. Participants and Experimental Design

For this preregistered study—which was approved by the College of Charleston Institutional Review Board—a power analysis using G*Power suggested that with two experimental conditions, we would need a minimum of 858 participants to detect a small effect of d = .20, given an alpha of .05 and power of .90 for a one-tailed significance test. We recruited 1040 participants via MTurk. Participants were limited to those in the United States, who had at least a 98% success rate and had successfully completed at least 500 previous tasks on Mturk.

Of the 1040 participants recruited, 32 did not complete the study, 7 failed at least one of the attention checks, and 19 reported they had previously taken a similar study. We excluded these participants from all analyses, resulting in a sample of N = 982 (52% male, 48% female, <%1 other or unspecified; *age*: M = 40.9, SD = 12.82; Anti-FW condition, n = 491 , Control condition, n = 491). Each was paid $3 for completing the study. Ten participants received a $10 bonus (see below for details). Participants took on average 39 minutes and 27 seconds to complete the study.

In our previous two behavioral studies we did not find any effect of prime (Anti-FW vs. Control) on cheating behavior or on free will beliefs. However, in Study 4, we were able to achieve a very small effect of prime on free will beliefs. The main difference between Study 4 and the previous behavioral studies was that Study 4 used the Velten sentences instead of the Crick passages and had a much larger sample size. For this study we therefore increased

the sample size to be comparable to, and even larger than, Study 4; we used the Velten sentences as the prime (Anti-FW vs. Control); and we included the same cheating paradigm we used in Studies 2 and 3. We also once again used the Free Will slider (as in Study 4) as an additional measure of free will beliefs. The order of presentation was as follows:

1. Velten sentences (Anti-FW vs. Control)
2. Participants responded to FAD+[18]
3. Participants responded to the free will slider
4. Participants completed the cheating paradigm (precisely as before).
5. Participants provided demographics
6. Participants were debriefed

2.5.2. Results

First, independent samples Welch's t-tests were conducted to test whether there was a difference between the conditions on FAD+ subscales and the Free Will slider. Using unadjusted *p*-values and a one-tailed test, there was a detectable effect of condition on the FAD+ Free Will subscale, $t_{(979.73)} = -2.368$, $p = .018$, $d = -.15$, and the Free Will slider, $t_{(971.51)} = -4.482$, $p < .001$, $d = -.29$, such that those in the Anti-FW condition reported lower beliefs in free will than those in the Control condition. We also found an effect of condition on the FAD+ Scientific Determinism subscale, $t_{(970.59)} = 2.737$, $p = .006$, $d = .18$, such that those in the Anti-FW condition reported higher beliefs in scientific determinism than those in the Control condition. There was no evidence of an effect on the remaining FAD+ subscales, *p*s > .246. See Table 4 for a summary.

Second, to test whether cheating behavior differed between the Anti-FW condition and Control condition, we performed an independent samples Welch's t-test. Even using a

[18] Reliabilities were again adequate for all four subscales of FAD+ (Free Will: $\alpha = .84$, Scientific Determinism: $\alpha = .77$; Fatalistic Determinism: $\alpha = .85$; Unpredictability: $\alpha = .78$).

one-tailed test, we failed to find evidence of a difference in cheating behavior between the Anti-FW condition (M = 9.88, SD = 3.27, 95% CI [9.59 , 10.2]) and the Control condition (*M* = 10.11, *SD* = 3.38, 95% CI [9.81, 10.40]), $t_{(265.3)} = 1.066$, $p = .287$, $d = .13$. In fact, in absolute terms, the mean cheating score was lower in the Control condition than in the Anti-FW condition, contrary to predictions. See Figure 3.

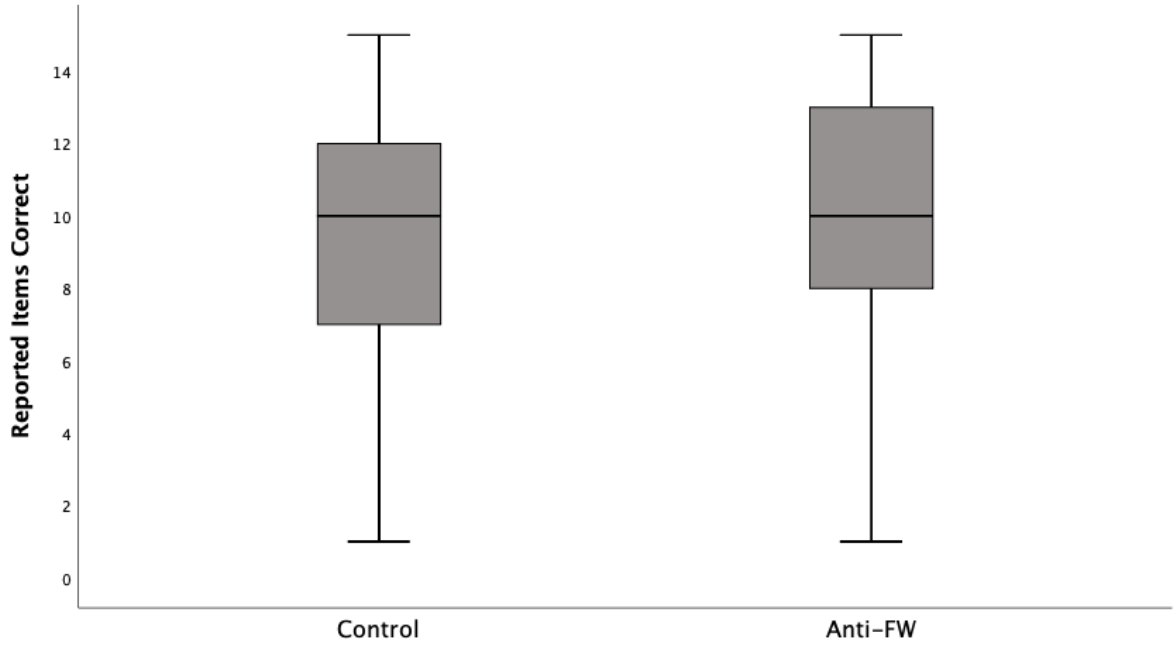
Table 4.

Independent Samples t-Test comparing Anti-FW and Control Conditions on FAD+ Subscales

	t	df	p	Mean Difference	SE Difference	Cohen's d
FAD+ Free Will	-2.368	979.7	0.018	-0.766	0.323	-0.15
FAD+ Scientific Determinism	2.737	970.6	0.006	0.831	0.304	0.18
FAD+ Fatalistic Determinism	1.160	980.0	0.246	0.324	0.279	0.07
FAD+ Randomness	-0.968	980.0	0.333	-0.244	0.253	-0.06
Free Will Slider	-4.482	971.5	< .001	-7.289	1.626	-0.29

Note. Welch's t-test. P-values have not been corrected for multiple comparisons. Positive values indicate higher valu
condition

Figure 3. Box Plot of Reported Items Correct for Anti-FW and Control Conditions.



2.5.3. Discussion

For Study 5, our primary goal was to extend what we did in Study 4 by adding a behavioral measure. We thought that if we could once again successfully manipulate free will beliefs, however weakly, using another very large sample, we would be well-placed to better test the purported relationship between these beliefs and cheating behavior. As we saw it, there were three ways the findings could turn out. First, we might successfully manipulate both free will beliefs and cheating behavior (which would support the original findings by Vohs and Schooler). Second, we could fail to successfully manipulate either free will beliefs or cheating behavior (which would not, by itself, undermine the relationship between the two). Finally, we could successfully manipulate free will beliefs but nevertheless fail to manipulate cheating behavior (which would call the purported link between the two into question).

The results from Study 5 provide compelling evidence for the third option—that is, while we were able to measurably weaken participants' beliefs in free will, this did not make them any more likely to cheat. Whereas Vohs and Schooler detected a very large effect size

regarding the purported relationship between free will beliefs and cheating behavior using a very small sample, we failed to find any evidence of this relationship using a very large sample. Moreover, while we were able to manipulate free will beliefs as measured by FAD+, the Cohen's effect size estimate for the decreases in FAD+ scores, while significant, was only *d* = -0.15, which is once again far smaller than the pre-designated "smallest effect size of interest" (*d* = .20) and much smaller than the original effect size detected by Vohs and Schooler (*d* = .84).

We think this is bad news for their original findings. On the one hand, we failed to manipulate free will beliefs in three studies using much larger samples sizes than the original studies by Vohs and Schooler (Studies 1–3). On the other hand, while we were able to weakly manipulate free will beliefs in two very high-powered studies (Studies 4 and 5), this did not influence how likely participants were to cheat (Study 5). So, not only is manipulating free will beliefs much more difficult than Vohs and Schooler originally suggested, but even when one succeeds in this regard, using a "sledgehammer" approach to maximize the chance of success, it does not appear to have any impact on cheating behavior.

3. Pooled Data Analysis

To provide a more precise estimate of the effects of the two manipulations used across the five studies, we conducted a set of exploratory regression analyses using data pooled across Studies 1, 4, and 5 (using the Velten task) and across Studies 2 and 3 (using the Crick manipulation). For the first set of analyses, given that Studies 1 and 4 only attempted to manipulate free will beliefs (with no accompanying measure of ethical behavior), we assessed the effects of the Velten manipulation on all four subscales of the FAD+.[19] All four analyses

[19] We did not assess the effect of the manipulation on mood given that the two studies used different measures.

entailed regressing the outcome variable (FAD+ subscale) on dummy variables representing condition (Control vs. Anti-FW) and study number. Results of these regressions are presented in Table 5.

Contrasting with the analyses of Studies 1 and 4, and consistent with Study 5, we observed a small, significantly negative effect of the Velten manipulation on the Free Will subscale, with the coefficient suggesting that the manipulation decreased free will beliefs by around one tenth of a standard deviation, or just one twelfth of a scale point (on a five point scale). Interestingly, however, the manipulation also induced a significant increase in scores on the Scientific Determinism subscale (also small, but close to twice the magnitude of the effect of the manipulation on free will beliefs). Additionally, the manipulation produced a non-significant increase in Fatalistic Determinism, with an effect size of similar magnitude to the effect of the manipulation on free will beliefs. There was no effect of the manipulation on scores on the Unpredictability subscale.[20] An additional regression model pooling data from Studies 2 and 5 to estimate the effect of the manipulation on the Free Will slider measure revealed a statistically significant but negative effect of the manipulation on free will beliefs (β = -0.34 [-0.43 – -0.26], $p < .001$). See the Table in the Pooled Analysis folder on our OSF page for this project for additional details: https://osf.io/d53um/

[20] We also observed a significant effect of study number on scores on the Unpredictability subscale. Specifically, participants in Study 1 reported significantly lower Unpredictability scores. We are unsure how to interpret this.

Table 5.
Pooled data analysis of effects of Velten manipulation on FAD+ subscales

	Free Will		Scientific Determinism		Fatalistic Determinism		Unpredictability	
Predictors	*B [95% CI]*	*p*	*B [95% CI]*	*p*	*B [95% CI]*	*p*	*B [95% CI]*	*p*
Intercept	0.11 [-0.05 – 0.27]	.190	-0.18 [-0.33 – -0.02]	**.030**	-0.17 [-0.33 – -0.01]	**.034**	-0.23 [-0.39 – -0.07]	**.005**
Condition	-0.11 [-0.20 – 0.03]	**.010**	0.19 [0.11 – 0.28]	**<.001**	0.09 [-0.00 – 0.17]	.051	-.02 [-0.10 – 0.07]	.705
Study 4	-0.01 [-0.17 – 0.16]	.951	0.11 [-0.05 – 0.28]	.175	0.16 [-0.01 – 0.33]	.062	0.23 [0.07 – 0.40]	**.006**
Study 5	-0.10 [-0.26 – 0.07]	.246	0.05 [-0.11 – 0.22]	.515	0.12 [-0.04 – 0.29]	.151	0.27 [0.11 – 0.44]	**.001**
R^2 / R^2 adjusted	.005 / .004		.011 / .009		.004 / .002		.005 / .004	

Note: All Ns = 2071; All coefficients are standardized; *p* values < .05 in boldface.

For the second set of analyses using the Crick manipulation in Studies 2 and 3, we again assessed the effects of the Crick manipulation on all four subscales of FAD+, as well as on BMIS subscales and our behavioral measure of cheating. Analyses were structured identically to those described above for Studies 1 and 2, regressing the outcome variable on dummy variables for condition and study number. Results of these analyses are presented in Table 6. Again, consistent with the analyses of the two studies in isolation, we observed no effects of the Crick manipulation on any FAD+ of BMIS subscale (though we note that the coefficient for the effect of the manipulation on free will beliefs was similar in magnitude to that of the Velten manipulation reported in Table 5). We did, however, observe a non-significant effect of the manipulation on our cheating measure in the *opposite* direction to that reported in the literature, such that participants in the Anti-FW condition reported slightly fewer items answered correctly. In other words, the Anti-FW manipulation produced

slightly less cheating (albeit without statistically significantly shifting participants' free will beliefs).

Table 5.
Pooled data analysis of effects of Crick manipulation on FAD+ and BMIS subscales and cheating behavior.

	Free Will		Scientific Determinism		Fatalistic Determinism		Unpredictability		Reported Items Correct		Positive Tired		Negative Relaxed	
Predictors	*B [95% CI]*	*p*	*B [95% CI]*	*p*	*B [95% CI]*	*p*	*B [95% CI]*	*p*	*B [95% CI]*	*p*	*B [95% CI]*	*p*	*B [95% CI]*	*p*
(Intercept)	0.03 [-0.12 – 0.17]	.693	0.05 [-0.10 – 0.19]	.507	0.08 [-0.07 – 0.22]	.308	-0.02 [-0.16 – 0.13]	.797	0.08 [-0.06 – 0.23]	.252	-0.04 [-0.19 – 0.10]	.577	0.06 [-0.09 – 0.20]	.437
Condition	-0.09 [-0.25 – 0.08]	.313	0.02 [-0.15 – 0.19]	.800	-0.02 [-0.19 – 0.15]	.832	-0.01 [-0.18 – 0.16]	.922	-0.16 [-0.33 – 0.00]	.054	0.03 [-0.14 – 0.19]	.769	-0.02 [-0.18 – 0.15]	.845
Study	0.03 [-0.14 – 0.20]	.706	-0.12 [-0.29 – 0.04]	.148	-0.14 [-0.30 – 0.03]	.114	0.05 [-0.12 – 0.22]	.575	0.00 [-0.17 – 0.17]	.979	0.06 [-0.11 – 0.23]	.498	-0.10 [-0.27 – 0.07]	.241
R^2 / R^2 adjusted	0.002 / -0.002		0.004 / 0.000		0.005 / 0.001		0.001 / -0.003		0.007 / 0.003		0.001 / -0.003		0.003 / -0.001	

Note: All Ns = 551; All coefficients are standardized.

4. General Discussion

The free will debate has gone mainstream in recent years in the wake of scientific advances that on some accounts seem to undermine free will. Given the traditional associations between free will and moral responsibility, a great deal may hang on this debate. In a high-profile paper on the relationship between free will beliefs and moral behavior, Vohs and Schooler (2008) cautioned against public pronouncements disputing the existence of free will, based on their findings concerning the relationship between free will beliefs and cheating. Our goal in this paper was to replicate their landmark findings. Across five studies, we were unable to detect any effects consistent with their findings greater than d = |.17|when using FAD+ as belief in free will measure or greater than d = |.29| when using free will slider as belief in free will measure. Specifically, while we were able to measurably, but very weakly, influence people's beliefs in free will in two of the five studies, we failed in our efforts to find any relationship between free will beliefs and cheating behavior. When coupled with the work of other researchers who have had difficulty replicating the original findings by Vohs and Schooler, we think this should give us further cause for concern.

There are a few other noteworthy findings that merit discussion before closing. First, our findings may suggest that the Velten statements are more effective in manipulating free will beliefs, however slightly, than the excerpt from Crick (even if neither seem to be effective in manipulating cheating behavior). It is possible that this is an artifact not of the respective manipulations, however, but of relative sample size (i.e., we had a much larger total sample size for the studies using Velten statements than the studies using the Crick excerpts). Nevertheless, there are at least two possible reasons why the Velten statements might be more effective. First, the Velten statements are overtly about both determinism and free will

while the Crick passage argues for a more esoteric kind of epiphenomenalism that people may find more confusing (or less compelling). And second, the control condition in the Velten paradigm is more appropriate than the control condition for the Crick paradigm—which involves an excerpt about consciousness from Crick's book, a topic that is closely related to the passage used in the experimental condition. The second thing worth mentioning is that our pooled analysis suggests that the Velten paradigm can (again, quite weakly) influence both free will beliefs and beliefs about scientific determinism (and sometimes fatalistic determinism). So, even when the Velten paradigm is effective in manipulating people's beliefs about free will, it lacks specificity. As such, it might be more fruitful moving forward to design manipulations that target either free will or determinism.[21]

All of that said, there are three potential limitations to our studies. First, we ran our studies online rather than using a convenience sample, as Vohs and Schooler did. While we tried to ensure that we mimicked their original work as much as possible, follow up studies with a convenience sample would certainly be valuable in terms of exact replication. However, the differences in sampling method should not detract from the upshot of our replication attempts. After all, the original effect (and its societal implications) are claimed to be pervasive and of potential practical importance. If *directly* communicating skepticism about free will barely weakened people's free will beliefs and (going beyond our own data) at most resulted in only a trivial increase in bad behavior (or affected behavior in a very limited range of contexts), then the basic effect is arguably unimportant and unworthy of the substantial attention it has received so far. Moreover, in terms of ecological validity, most

[21] In light of our findings, we thought we might make the following suggestion for researchers interested in trying to manipulate free will beliefs: Given the estimated effect size of the two manipulations in the pooled data analysis was around one tenth of a SD (i.e., $d = 0.10$), this implies that a simple two condition study would require over 1,500 participants *per condition* to achieve 80% power for a significant two-tailed manipulation check using the Free Will subscale of FAD+.

ordinary people who read about scientific findings that might bear on free will likely do so on their computers at home, as in our study, rather than in a psychology department laboratory context with limited demographic diversity.

A second limitation of our research is that we only used American participants. However, this limitation is an artifact of our goal of trying to replicate the work by Vohs and Schooler. Because they used an American sample, we used an American sample. Figuring out whether their work replicates in a non-American sample is a task for another day. That said, we would obviously welcome cross-cultural studies that implemented our paradigms to see whether our findings (or lack thereof) are cross-culturally stable.

The third and final limitation of our experimental design concerns the possibility that MTurk participants may not be as attentive as in-lab participants, such that the manipulations we used may be less effective for the former than the latter. To guard against this possibility, we used an attention check in all of our studies and excluded any participants who failed it. We also used two items designed to encourage participants to pay attention by reminding them that they would be asked to write about the content of the vignette they read. While these measures can obviously not guarantee that participants are paying attention, we'd like to think that they reduce the likelihood of inattention. Additionally, many lab tasks that are particularly susceptible to lapses in attention have been replicated using MTurk populations, including tasks that depend on differences in reaction times on the scale of milliseconds (e.g., Erikson Flanker tasks) and memory tasks that are heavily attention-dependent (see Woods, et al., 2015 for a review).

Setting these potential limitations aside, we nevertheless think we have made a valuable contribution to the literature on the relationship between free will beliefs and moral behavior. Minimally, our findings serve as a cautionary tale for those who fret that challenging

free will beliefs might undermine public morality. Future research on this front will have to take into consideration the difficulty of replicating both standard manipulations of belief in free will and the purported link between free will skepticism and morality. Contrary to our initial expectations, the association between free will beliefs and moral behavior appears to be highly elusive. As such, worries about the purported erosion of societal mores in the wake of recent advances in neuroscience are likely to be misplaced. The belief in free will appears to be more stable, robust, and resistant to challenge than earlier work suggests. While some scientists may think that their research undermines the traditional picture of agency and responsibility, public beliefs on this front are likely to be relatively slow to change. Moreover, even if beliefs about free will were to incrementally change, given the lack of association between dispositional free will beliefs and moral behavior reported by Crone and Levy (2018), it is unclear that people would have difficulty integrating such beliefs into a coherent worldview that permits the same level of moral behavior.

References

Alquist, J. L., Ainsworth, S. E., & Baumeister, R. F. (2013). Determined to conform: Disbelief in free will increases conformity. *Journal of Experimental Social Psychology*, *49*, 80–86.

Aquino, K. F., & Reed, A. I. (2002). The self-importance of moral identity. *Journal of Personality and Social Psychology*, *83*, 1423–1440.

Bargh, J. (2008). Free will is unnatural. In J. Baer, J. Kaufman, & R. Baumeister (Eds.), *Are we free? Psychology and free will* (pp. 128-154). New York: Oxford University Press.

Baumeister, R. F., & Brewer, L. E. (2012). Believing versus disbelieving in free will: Correlates and consequences. *Social and Personality Psychology Compass*, *6*(10), 736–745.

Baumeister, R. F., Masicampo, E. J., & DeWall, C. N. (2009). Prosocial benefits of feeling free: Disbelief in free will increases aggression and reduces helpfulness. *Personality and Social Psychology Bulletin, 35*, 260–268.

Baumeister, R. F., Sparks, E. A., Stillman, T. F., & Vohs, K. D. (2008). Free will in consumer behavior: Self-control, ego depletion, and choice. *Journal of Consumer Psychology, 18*, 4–13.

Baumeister, R. (2008). Free will, consciousness, and cultural animals. In J. Baer, J. Kaufman, and R. Baumeister, (Eds.) *Are we free? Psychology and free will* (pp. 65–85). New York: Oxford University Press.

Camerer, C. F., Dreber, A. et al. (2016). Evaluating replicability of laboratory experiments in economics. *Science*, *351*(6280),1433–6.

Caspar, E. A., Vuillaume, L., Magalhães de Saldanha da Gama, P.A., & Cleeremans, A. (2017). The influence of (dis)belief in free will on immoral behavior. *Frontiers in Psychology*, *17*, https://doi.org/10.3389/fpsyg.2017.00020

Cashmore, A. (2010). The Lucretian Swerve: The biological basis of human behavior and the criminal justice system. *Proceedings of the National Academy of Sciences of the United States of America*, *107*, 4499–504.

Chandler, J. J., Mueller, P., & Paolacci, G. (2014). Nonnaïveté among Amazon Mechanical Turk workers: consequences and solutions for behavioral researchers. *Behavior Research Methods*, *46*(1), 112–30. http://doi.org/10.3758/s13428-013-0365-7

Coyne, J. (2012). Why you don't really have free will. *USA Today* (January 1st)

Cova, F., Strickland, B., Abatista, A., Allard, A., Andow, J., et al. (2018). Estimating the reproducibility of experimental philosophy. *Review of Philosophy and Psychology*, 1–36.

Crescioni, A. W., Baumeister, R. F., Ainsworth, S. E., Ent, M., & Lambert, N. M. (2016). Subjective correlates and consequences of belief in free will. *Philosophical Psychology*, *5089*(July), 1–23. http://doi.org/10.1080/09515089.2014.996285

Crick, F. (1994). *The astonishing hypothesis*. New York: Scribner's.

Crone, D. & Levy, N. (2018). Are free will believers nicer people? Four studies suggest not. *Social Psychological and Personality Science*. https://doi.org/10.1177/1948550618780732

Dienstbier, R. A., Kahle, L. R., Willis, K. A., & Tunnell, G. B. (1980). The impact of moral theories on cheating. *Motivation and Emotion*, *4*(3), 193–216.

Doyen, S., Klein, O., Simons, D. J., & Cleeremans, A. (2014). On the other side of the mirror: Priming in cognitive and social psychology. *Social Cognition*, *32*(Supplement), 12–32.

Earp, B. D., Everett, J. A. C., Nadelhoffer, T., Caruso, G. D., Sharif, A., & Sinnott-Armstrong, W. (2019). Determined to be humble? Exploring the relationship between belief in free will and humility. *PsyArXiv.* https://psyarxiv.com/3bxra/.

Earp, B. D., & Trafimow, D. (2015). Replication, falsification, and the crisis of confidence in social psychology. *Frontiers in Psychology*, *6*(621), 1–11.

Earp, B. D. (2016). What did the OSC replication initiative reveal about the crisis in psychology? *BMC Psychology*, *4*(28), 1–19.

Earp, B. D. (2017). The need for reporting negative results—a 90 year update. *Journal of Clinical and Translational Research*, *3*(S2), 1–4.

Earp, B. D. (in press). Falsification: How does it relate to reproducibility? In J.-F. Morin, C. Olsson, & E. O. Atikcan (Eds.), *Research Methods in the Social Sciences: An A-Z of Key Concepts*. Oxford: Oxford University Press.

Embley, J., Johnson, L. G., & Giner-Sorolla, R. (2015). *Replication of Study 1 by Vohs & Schooler (2008, Psychological Science)*. Retrieved from https://osf.io/2nf3u/

Errington, T. M., Iorns, E., Gunn, W., Tan, F. E., Lomax, J., & Nosek, B. A. (2014). Science Forum: An open investigation of the reproducibility of cancer biology research. *eLife 2014, 3:e04333 DOI: 10.7554/eLife.04333*

Feldman, G., Baumeister, R. F., & Wong, K. F. E. (2014). Free will is about choosing: The link between choice and the belief in free will. *The Journal of Experimental Social Psychology*, *55*, 239–245.

Feldman, G., Chandrashekar, S. P., & Wong, K. F. E. (2016). The freedom to excel: Belief in free will predicts better academic performance. *Personality and Individual Differences*, *90*, 377–383.

Genschow, O., Rigoni, D., & Brass, M. (2017). Belief in free will affects causal attributions when judging others' behavior. *Proceedings of the National Academy of Sciences*, *114*(38), 10071–10076.

Greene, J., & Cohen, J. (2004). For the law neuroscience changes nothing and everything. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *359*, 1775–85.

Greenwald, A. G. (1975). Consequences of prejudice against the null hypothesis. *Psychological bulletin*, *82*(1), 1.

Harris, S. (2012). *Free Will.* New York City: Free Press.

Haynes, J. D. (2011). Beyond Libet: Long-term prediction of free choices from neuroimaging signals." In W. Sinnott-Armstrong & L. Nadel (eds.), *Conscious Will and Responsibility* (pp. 85–96). Oxford: Oxford University Press.

Hertz, S. G., & Krettenauer, T. (2016). Does moral identity effectively predict moral behavior?: A meta-analysis. *Review of General Psychology*, *20*(2), 129–140. http://doi.org/10.1037/gpr0000062

Ioannidis, J. P. (2008). Why most discovered true associations are inflated. *Epidemiology*, *19*(5), 640–648.

Krueger, F., Hoffman, M., Walter, H., & Grafman, J. (2014). An fMRI investigation of the effects of belief in free will on third-party punishment. *Social Cognitive and Affective Neuroscience, 9*(8), 1143–1149.

LeBel, E. P., McCarthy, R. J., Earp, B. D., Elson, M., & Vanpaemel, W. (2018). A unified framework to quantify the credibility of scientific findings. *Advances in Methods and Practices in Psychological Science*, *1*(3), 389–402.

Martin, N. D., Rigoni, D., & Vohs, K. D. (2017). Free will beliefs predict attitudes toward unethical behavior and criminal punishment. *PNAS*, *114*(28), 7325–7330.

Mayer, J. D., & Gaschke, Y. N. (1988). The experience and meta-experience of mood. *Journal of Personality and Social Psychology*, *55*, 102–111.

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

Mercier, B., Wiwad, D., Piff, P., Aknin, L. B., Robinson, A. R., & Shariff, A. (n.d.) Does belief in free will increase support for economic inequality? https://osf.io/pgyzw/?fbclid=IwAR1hsicR5nmcKmOTCgeFTjRNbdxmVHPyPoggZtNxyBHOrLW6WD997oKmUrw

Monroe, A. E., Brady, G., & Malle, B. F. (2016). This isn't the free will worth looking for: General free will beliefs do not influence moral judgments, agent-specific choice ascriptions do. *Social Psychological and Personality Science*, 1–9.

Montague, P. R. (2008). Free will. *Current Biology,* 18, R584–5.

Moynihan, A. B., Igou, E. R., & van Tilberg, V. A. P. (2017). Free, connected, and meaningful" Free will beliefs promote meaningfulness through belongingness. *Personality and Individual Differences*, *107*, 54–65.

Murphy, R. O., Ackermann, K. A., & Handgraaf, M. J. J. (2011). Measuring social value orientation. *Judgment and Decision Making*, *6*, 771–781.

Ogletree, S. M., & Oberle, C. D. (2008). The nature, common usage, and implications of free will and determinism. *Behavior and Philosophy*, *36*, 97–111.

Open Science Collaboration: Nosek, B., Aarts, A., Alexander, A., Anderson, J., Kappes, J. E., Barry, H. et al. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716: DOI: 10.1126/science.aac4716

Overbye, D. (2007). Free will: Now you have it, now you don't. *The New York Times*. January, 2nd.

Paulhus, D. L. & Carey, J. M. (2011). The FAD+: Measuring lay beliefs regarding free will and related constructs. *Journal of Personality Assessment, 93*, 96–104.

Paulhus, D.L., & Margesson, A. (1994). Free Will and Determinism (FAD) scale. Unpublished manuscript, University of British Columbia, Vancouver, British Columbia, Canada.

Rand, D. G. (2018). Non-naïvety may reduce the effect of intuition manipulations. *Nature Human Behaviour*, *81*(2017), 190560. http://doi.org/10.1038/s41562-018-0404-6

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological bulletin*, *86*(3), 638.

Schmeichel, B. J., Vohs, K. D., & Baumeister, R. F. (2003). Intellectual performance and ego depletion: Role of the self in logical reasoning and other information processing. *Journal of Personality and Social Psychology*, *85*, 33–46.

Scholl, B. (2017). Reliability in psychology: Means versus ends. *APS Observer, 30*(9), 38–39.

Schooler, J., Nadelhoffer, T., Nahmias, E., & Vohs, K. (2014). Measuring and manipulating beliefs and behaviors associated with free will: The good, the bad, and the ugly. In A. Mele (Ed.), *Surrounding free will* (pp. 72-94). New York: Oxford University Press.

Shariff, A. F., Greene, J. D., Karremans, J. C., Luguri, J. B., Clark, C. J., Schooler, J. W., et al. (2014). Free will and punishment a mechanistic view of human nature reduces retribution. *Psychological Science*, *25*, 1563–1570.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2012). A 21 word solution. *Dialogue: The Official Newsletter of the Society for Personality and Social Psychology*, *26*, 4–7. http://dx.doi.org/10.2139/ssrn.2160588

Stillman, T. F. & Baumeister, R. F. (2010). Guilty, free, and wise: Determinism and psychopathy diminish learning from negative emotions. *Journal of Experimental Social Psychology*, *46*, 951–960.

Stillman, T. F., Baumeister, R. F., Vohs, K. D., Lambert, N. M., Fincham, F. D., & Brewer, L. E. (2010). Personal philosophy and personnel achievement: Belief in free will predicts better job performance. *Social Psychological and Personality Science*, *1*, 43–50.

Stillman, T. F., Baumeister, R. F., & Mele, A. R. (2011). Free will in everyday life: Autobiographical accounts of free and unfree actions. *Philosophical Psychology, 24*, 381–394.

Stillman, T. F., Baumeister, R. F., Vohs, K. D., Lambert, N. M., Fincham, F. D., & Brewer, L. E. (2010). Personal philosophy and personnel achievement: Belief in free will predicts better job performance. *Social Psychological and Personality Science, 1,* 43–50.

Trafimow, D., & Earp, B. D. (2016). Badly specified theories are not responsible for the replication crisis in social psychology: Comment on Klein. *Theory & Psychology*, *26*(4), 540–548.

Trafimow, D., & Earp, B. D. (2017). Null hypothesis significance testing and Type I error: The domain problem. *New Ideas in Psychology*, *45*, 19–27.

Velten, E. (1968). A laboratory task for the induction of mood states. *Behavioral Research and Therapy*, *6,* 607–617.

Vohs, K.D., & Schooler, J. W. (2008). The value of believing in free will: Encouraging a belief in determinism increases cheating. *Psychological Science*, *19*, 49–54.

Walter, H. (2001). Neurophilosophy of free will: From libertarianism illusions to a concept of natural autonomy. Cambridge, MA: MIT Press

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology*, *54*, 1063–1070.

Wegner, D. (2002). *The illusion of conscious will*. MIT Press.

Woods, A. T., Velasco, C., Levitan, C. A., Wan, X., Spence, C. (2015). Conducting perception research over the internet: A tutorial review. *PeerJ*, *3*, e1058.