BASIC AND APPLIED SOCIAL PSYCHOLOGY, 36:91–98, 2014

ISSN: 0197-3533 print/1532-4834 online
DOI: 10.1080/01973533.2013.856792

Routledge
Taylor & Francis Group

Out, Damned Spot: Can the "Macbeth Effect" Be Replicated?

Brian D. Earp
Yale University

Jim A. C. Everett
University of Oxford

Elizabeth N. Madva
Weill Cornell Medical College

J. Kiley Hamlin
University of British Columbia

Zhong and Liljenquist (2006) reported evidence of a "Macbeth Effect" in social psychology: a threat to people's moral purity leads them to seek, literally, to cleanse themselves. In an attempt to build upon these findings, we conducted a series of direct replications of Study 2 from Z&L's seminal report. We used Z&L's original materials and methods, investigated samples that were more representative of the general population, investigated samples from different countries and cultures, and substantially increased the power of our statistical tests. Despite multiple good-faith efforts, however, we were unable to detect a "Macbeth Effect" in any of our experiments. We discuss these findings in the context of recent concerns about replicability in the field of experimental social psychology.

Over the past two decades, a collection of studies in moral psychology (e.g., Haidt & Graham, 2007; Rozin, Haidt, & McCauley, 2000; Shweder, Much, Mahapatra, & Park, 1990) has shown that in some cultures and groups, concerns about physical purity are associated with people's moral judgments. In these cases, immoral persons and acts are considered physically defiling. For example, some people seek to avoid contamination from certain outgroups (e.g., "dirty" Arabs, Jews) and classes perceived as inferior (e.g., the "untouchable" caste in India). These supposedly contaminating individuals are seen as not just physically disgusting; they are also morally disgusting and therefore less human (Nussbaum, 2004). Thus, although traditional conceptions of morality stress factors such as harm and fairness in arriving at a moral evaluation, Haidt and colleagues have argued that intuitive notions of disgust versus purity may constitute an additional moral foundation. According to this view, disgust evolved to protect the body from such "impure" threats as parasites, germs, and rotten food but then became associated later on with the more abstract domains of social reasoning and moral judgment.

In a now-classic publication, Zhong and Liljenquist (2006) sought to explore the idea that feelings of moral purity are not just associated with but are actually grounded in feelings of physical cleanliness. They began by noting a growing body of work that shows that higher order thoughts and feelings may indeed be scaffolded atop basic bodily experiences—perhaps through a mechanism involving "neural re-use" over evolutionary as well as ontogenetic time (e.g., Anderson, 2010). Zhong and Liljenquist therefore reasoned that any threat to people's moral purity might lead them to seek, literally, to cleanse

Correspondence should be sent to Brian D. Earp, Oxford Centre for Neuroethics, University of Oxford, Suite 8, Littlegate House, 16/17 St Ebbe's Street, Oxford, OX1 1PT, United Kingdom. E-mail: brian.earp@gmail.com

themselves. In the literary canon, this notion traces famously to the dramatic "Out, damned spot!" scene in Shakespeare's *Macbeth*, in which Lady Macbeth seeks to "wash away" her murderous sins by physically scrubbing her hands.

To provide empirical support for the existence of a real-life "Macbeth" effect, Zhong and Liljenquist conducted a series of elegant studies. In one experiment, they asked undergraduates to copy a passage describing an unethical deed as opposed to an ethical deed, and showed that these participants were subsequently likelier to rate cleansing products as more desirable than other consumer products. When asked to remember an unethical vs. ethical deed that they themselves had committed in the past, participants tended to pick an antiseptic wipe over a pen as compensation for their involvement in the study. Finally, when participants were made to cleanse themselves physically by washing their hands after recalling an unethical deed, they ended up being *less* likely to volunteer to help out a "desperate graduate student" at the end of the study. This result was taken to suggest that an act of physical cleansing can unconsciously restore a feeling of moral purity and thus eliminate the need for further moral action.

Zhong and Liljenquist's thought-provoking theory has gained additional support from a number of more recent studies that were carried out to build upon the "Macbeth Effect" foundation. For example, Gollwitzer and Melzer (2012) reported that playing violent video games causes inexperienced players to prefer hygiene-related products over non-hygiene-related products in a subsequent product selection task. Using Zhong and Liljenquist's explanatory framework, the authors speculated that "behaving violently in a virtual environment threatened the moral selves of [the participants], which, in turn, evoked a desire to physically cleanse themselves" (p. 1359). Reuven, Liberman, and Dar (2013) provided experimental evidence that moral cleansing effects may be stronger in individuals with obsessive-compulsive disorder, replicating findings from Zhong and Liljenquist's "volunteerism" task (Study 4) in a patient population. Lee and Schwarz (2010) showed that the Macbeth Effect might even be mode specific: When participants were instructed to perform an immoral action using their *hands* (i.e., typing a malevolent lie into an e-mail message and then actually sending the message), they were more likely to prefer *hand sanitizer* over other products on a consumer product survey, whereas if they performed the same action by using their *mouth* (i.e., by telling the lie over the phone and leaving a voice mail), they were more likely to prefer *mouthwash* over other items. These results suggest that people may form an unconscious motivation to clean the specific part of the body that was used to effectuate an immoral deed. Finally, Schnall, Benton, and Harvey (2008, Study 2) asked a group of participants to watch a disgusting movie and then subsequently form moral judgments concerning others' behavior. Participants who were first instructed to wash their hands before forming the moral judgments evaluated other people's transgressions less harshly compared to participants who did not first wash their hands. Schnall et al. interpreted this finding to mean that handwashing can unconsciously "cleanse" feelings of disgust, thereby dampening the severity of subsequent moral judgments.

As Chapman and Anderson (2013) argued in a recent review, these and other findings provide compelling evidence for the existence of a moral–physical link between purity and disgust in the realm of human cognition. Furthermore, the Macbeth Effect itself—considered as one manifestation of this link—seems to be well supported by the results of several different paradigms and experimental approaches, including the four studies originally reported by Zhong and Liljenquist (see Lee & Schwarz, 2011, for further discussion). Given such a robust experimental foundation for the existence of a real-life Macbeth Effect, we were interested to see whether we could extend Zhong and Liljenquist's theory to still other areas within the moral domain. If manipulating feelings about moral purity can lead to changes in feelings of physical purity, we asked, what exactly is included in the latter concept? Zhong and Liljenquist (2006) focused primarily on *surface cleanliness*, that is, clean skin and clean surroundings. But might the link between morality and physical purity apply to other areas in which purity seems to play a role? For instance, it is often the case that purifying the inside of one's body is as important as purifying the outside, as with some popular diet movements or certain kinds of religious fasting (see, e.g., Eskine, 2013). In addition, it may be the case that one can be pure in mind: Those with high levels of self-control may be less likely to engage in impure thoughts and behaviors (see, e.g., Graham & Haidt, 2010).

To explore these ideas, we conducted a pilot study (Earp, Jarudi, Hamlin, & Madva, 2008) in which we adapted one of Zhong and Liljenquist's (2006) experiments to examine surface cleanliness—for purposes of replication—as well as two additional domains of purity: organic naturalness (i.e., purifying the inside of one's body) and self-control (i.e., purifying one's mind). In Part 1, participants were asked to recall an unethical or ethical deed and to describe in detail any feelings or emotions they experienced (Zhong & Liljenquist, 2006, p. 1451). In Part 2, they then rated the desirability of a variety of consumer products, including surface cleansing items (from Zhong & Liljenquist, 2006, p. 1452), organic items, self-control items, and filler items. We reasoned that if self-control and organic naturalness were additional domains of purity, then morally threatened participants might rate

products associated with these domains higher, as they do cleansing products.

Contrary to predictions, we did not observe any significant differences between the morally affirmed versus threatened groups on ratings for items from our novel dependent categories (Earp et al., 2008). Unexpectedly, however, we also failed to replicate the results from the original "consumer products" experiment (Study 2) reported in Zhong and Liljenquist (2006). Because our sample size for this pilot study was nearly twice as large as the sample used by Zhong and Liljenquist, and because we attempted to hew as closely as possible to their original design, we were certainly surprised by this null finding. Indeed, we had already taken the Macbeth Effect for granted in some of our earlier work (Earp, Dill, Harris, Ackerman, & Bargh, 2010). Nevertheless, as there were some subtle differences between the Zhong and Liljenquist paradigm and our own—including our addition of the "organic" and "self-control" items to the consumer products rating task—we reasoned that our failure to replicate must have been due to design adjustments on our part (see Earp et al., 2008).

Before proceeding further in our research, we decided to review the Macbeth Effect literature to see whether other groups may have experienced similar pitfalls in their own investigations. Despite the fact that unsuccessful replication attempts are typically underrepresented within the published literature—leading to the notorious "file drawer" problem in scientific inference (Rosenthal, 1979; Scargle, 2000)—we did manage to discover two reports of a lack of ability to detect a Macbeth Effect using paradigms quite similar to those employed by Zhong and Liljenquist. In the first set of experiments, Fayard, Bassi, Bernstein, and Roberts (2009) attempted "conceptual" replications of Studies 3 and 4 from Zhong and Liljenquist (2006), adding personality inventories and extra conditions, and making small adjustments to materials and/or procedure. In the second set of experiments, Gámez, Díaz, and Marrero (2011) attempted replication of all four of the paradigms described by Zhong and Liljenquist, but they too added personality measures and made such adjustments as translating the materials into Spanish and changing the number of cleansing items in the consumer product ratings task (see Earp, 2011, for further discussion). In addition, the sample sizes used by Gámez et al. were quite small, meaning that their statistical tests may have been underpowered. For these reasons, although the studies by Fayard et al. and Gámez et al. are certainly interesting, they fall short of being very conclusive.

Given these previous mixed efforts, then, and in light of recent concerns about replicability in the field of experimental social psychology (Francis, 2012; see also the "Special Section on Replicability in Psychological Science: A Crisis of Confidence?" in *Perspectives on Psychological Science*, 2012[1]), we thought it important to attempt a direct (and more adequately powered) replication of Zhong and Liljenquist's Study 2 before carrying out further research in the area.

In this article, we report three such attempts. For these experiments, we used the authors' original materials and methods, we investigated samples that were more representative of the general population than in the original experiments, we investigated samples from different countries and cultures, and we substantially increased the power of our statistical tests. Nevertheless, we still failed to replicate Zhong and Liljenquist's initial reported findings. Our research suggests that more work will be needed to clarify the scope and robustness of the original results.

STUDY 1

For this study, we requested the original materials from the lead author of the 2006 publication, who very graciously provided them to us, along with instructions on how to carry out the experiment. In addition, we enlisted the help of a colleague from the United Kingdom (now the second author of this report)—who had not been involved in any way with the design or execution of our pilot study (Earp et al., 2008)—to attempt replication in his home country. Here we describe this attempt.

Method

Participants. Participants in this study were 153 undergraduate students enrolled at a university in the United Kingdom. Participants were invited to take part via e-mail messages sent to departmental mailing lists and received a chocolate bar in exchange for their time.

Materials and procedure. Using the computer program G*Power 3.1 (Faul, Erdfelder, Buchner, & Lang, 2009), we determined that the required sample size to achieve a power of .85 for a two-tailed t test with two groups and an assumed effect size of Cohen's $d = 0.5$ was 146. Computing the effect size from the original Zhong and Liljenquist Study 2 (for desirability ratings of cleansing items) actually yields a Cohen's d of 1.08; yet, according to Pashler, Coburn, and Harris (2012), "an examination of a small subset of the social/goal priming literature suggests that large effect sizes in the range from .5 to 1.0 are quite typical" (p. 2). This means that even by the standards of the quite ample-seeming effect sizes noted in the goal priming literature (which strike Pashler et al. as "rather curious" indeed; p. 2), the original reported Macbeth Effect is on the higher end of the spectrum. Thus, because initial estimates of

[1] This special section can be seen by visiting http://pps.sagepub.com/content/7/6.toc.

effect sizes for new findings tend to be biased large (Kepes, Banks, McDaniel, & Whetzel, 2012), we elected to assume a "true" effect size of something closer to what is found on the lower end of the social priming literature. In this way, we hoped to be able to compensate for any possible effect size "bias" in the original research, leading us to recruit more than five times as many participants as Zhong and Liljenquist (2006) did in their own Study 2.

In Part 1, participants were randomly assigned to one of two priming conditions: ethical or unethical. In Part 2, participants rated a number of consumer products for their desirability on a scale of 1 to 7. In an attempt to prevent participants' drawing any connections between Parts 1 and 2, they were told that Parts 1 and 2 were two separate experiments. Just as in Study 2 from Zhong and Liljenquist's (2006) original report, participants were told they were taking part in an investigation into handwriting and personality and were asked to hand-copy a short story written in the first person. In the "unethical" condition, the paragraph described an unethical deed from the first-person perspective, as follows:

> Two years ago, when I was a junior partner at a prestigious law firm, I was coming up for promotion against another junior partner, Chris. For several months, Chris had been working on a major case for the city that would make or break his career at the firm. However, he could not locate a key zoning document, without which, it was unlikely that he would have sufficient evidence to successfully argue his case. Late one evening, as I was rummaging through a corner filing cabinet, I happened to come across the zoning document that Chris was in desperate need of. I pulled it from the cabinet and walked over to the office shredder, knowing that my promotion would now be secured.

In the "ethical" condition, the paragraph was exactly the same, except that the last sentence read: "I pulled it from the cabinet and placed it without a note on Chris' desk, knowing that he would be so relieved when he arrived to work the next morning."

Participants were then told that they were taking part in research looking at consumer marketing and were asked to rate the desirability of various products from 1 (*completely undesirable*) to 7 (*completely desirable*) and to say how much they would be willing to pay (£) for each product. The 10 items used were the exact same original items from Zhong and Liljenquist's study, in their original order, with four items adapted slightly for a British sample by replacing unfamiliar American brands with equivalent British brands. The items and their order were specifically as follows: Post-it notes, Dove shower soap, Colgate toothpaste [Crest toothpaste in the original], pressed fruit juice[2] [Nanucket Nectars juice in the original], Energizer batteries, Sony CD cases, Windex glass cleaner, Dettoll disinfectant [Lysol countertop disinfectant in the original], Snickers candy bar, and Surf laundry detergent [Tide laundry detergent in the original].

Upon completion of the consumer products survey, participants were given a chocolate bar to compensate for their time and were thanked for their participation.

Results

Independent samples *t* tests revealed no significant difference of condition on desirability of consumer product, $t(151) = .03$, $p = .97$, 95% CI [–0.29, 0.30], with no significant difference in the mean desirability of the cleansing items between the moral condition ($M = 3.09$) and immoral condition ($M = 3.08$). Similarly, there was no significant difference in how much participants were willing to pay for the consumer products, $t(151) = -.28$, $p = .78$, 95% CI [–0.36, 0.27], with comparable means in both the ethical condition ($M = 2.19$) and the unethical condition ($M = 2.24$). Looking at individual items, there were no significant effects of condition on the desirability of—or willingness to pay for—any individual cleansing item.

Discussion

Contrary to expectations, the present experiment proved unsuccessful in replicating the findings from Study 2 of Zhong and Liljenquist (2006). This is despite using identical instructions, primes, and materials (with four minor adjustments to brand names, discussed next) as well as a much larger sample size ($N = 153$, compared to $N = 27$) and thus sufficient power to detect an effect if one were present. The only difference in materials concerned the specific brand names listed for four of the items in the set of dependent measures. These differences were included to accommodate a British sample. Analyses showed, however, that these altered items did not uniquely influence the results, as indeed the brand names chosen were very close equivalents to the American originals.

[2] This item was called "Innocent Juice" for the first 76 participants used in this study, based on a well-known and popular British brand of health juices. Yet as an anonymous reviewer noted in response to an earlier draft of this article, the word "Innocent" in the brand name could possibly bias subjects (consciously) due to its explicit connotations of purity. For the remaining 77 participants, then, the item was renamed "Pressed Juice." There was no significant difference in desirability ratings for the item based on its name, $t(151) = -1.71$, $p = .09$, suggesting that this factor did not bias our results one way or the other. To confirm this, we then ran separate analyses for participants who completed the study with the item "Innocent Juice" versus those who saw "Pressed Juice." Just as we found for participants' responses overall, there was no effect of condition on desirability or willingness to pay for the cleansing items in either group of participants.

A more general difference, however, between the original study and the replication attempt reported here is that although the original study was conducted in the United States, the replication attempt was carried out in the United Kingdom. Thus, at a minimum, our results may be taken to show that the Macbeth Effect—as measured by the present "consumer products" paradigm—may not be universal in nature but rather culture specific. In line with this interpretation, we note that at least one of the previous failures to replicate the Macbeth Effect (see the Introduction) was also carried out with a non-U.S. sample—i.e., with Spanish participants—leading the authors of the experiments to call their report "The *Uncertain Universality* [emphasis added] of the Macbeth Effect with a Spanish Sample" (Gámez et al., 2011). Given that the United States has an arguably unusual history in terms of its pursuit of purity (see, e.g., "Chasing Dirt: The American Pursuit of Cleanliness" by Hoy, 1995), we decided to run two additional replication attempts. In the first one (Study 2), we used a sample from the United States (just as in the initial Zhong and Liljenquist research) and reverted back to all of the original brand names. In the second study (Study 3), to further investigate the "universality" of the Macbeth Effect, we used an Indian sample. Our aim in this study was to see whether the Macbeth Effect might be found in a non-U.S. culture that is nevertheless similarly known for its emphasis on purity in moral discourse—as in the conflation between moral and physical purity in the Hindu caste system (Shweder, Mahapatra, & Miller, 1987).

STUDY 2

Method

Participants. One hundred fifty-six American participants (83 female, M age = 33), using the Mechanical Turk (MTurk) online interface, participated in exchange for $.30. MTurk is a website that facilitates payment for the completion of tasks posted by researchers. Participant samples recruited through this service have been shown to be more representative of the general population than are student samples, and are known to yield reliable data (Buhrmester, Kwang, & Gosling, 2011). Eight participants were excluded from analyses for failure to complete the questionnaires.

Materials and procedure. As in Study 1, participants were randomly assigned to one of two priming conditions: ethical or unethical. All participants subsequently rated a number of consumer products for their desirability on a scale of 1 to 7, and noted how much they would be willing to pay ($) for each. To adapt the original priming materials from Zhong and Liljenquist for use in an online medium, the passages about helping/sabotaging a coworker were presented on participants' computer screens with all of their punctuation removed. Participants were asked to retype the passage (rather than rewrite it, by hand, as in the original studies), inserting simple punctuation marks such as full stops (periods), commas, and capitalization where appropriate; participants could not advance to the next screen without performing this task, and all participants completed the priming task successfully. Although this design adjustment involved a slight departure from the rewriting task used in Zhong and Liljenquist's original Study 2, we reasoned that our online-friendly prime might actually be more effective than the original. This is because to determine which punctuation marks were needed, participants would presumably have to process the meaning of the passage, whereas to hand-copy a passage exactly as it is written one could work by simple rote.

After participants completed this punctuation priming task, they were shown a screen in which they were told that they were now taking part in research looking at consumer marketing. They were asked to rate the desirability of various products from 1 (*completely undesirable*) to 7 (*completely desirable*) and to say how much they would be willing to pay ($) for each product. The 10 items presented were the original items from Zhong and Liljenquist's study, with no adjustments made to brand names, and were presented in their original order: Post-it notes, Dove shower soap, Crest toothpaste, Nanucket Nectars juice, Energizer batteries, Sony CD cases, Windex glass cleaner, Lysol countertop disinfectant, Snickers candy bar, and Tide laundry detergent.

After completing the consumer products rating task, participants were shown a screen that thanked them for their efforts and were then directed to a link for claiming their small monetary reward.

Results

Independent samples t tests revealed no significant difference of condition on desirability of the cleansing items, $t(146) = -.79$, $p = .43$, 95% CI [–0.62, 0.27] with comparable means in both the ethical ($M = 4.23$) and unethical ($M = 4.41$) conditions. Similarly, there was no significant difference in how much participants were willing to pay for the cleansing items, $t(146) = .17$, $p = .87$, 95% CI [–0.50, 0.59], with comparable means for both the ethical ($M = 3.50$) and unethical conditions ($M = 3.46$).

Analyses were conducted on all individual cleansing items, and revealed no effect of condition on any individual item, with one exception: Consistent with predictions, a significant difference between conditions was found for how much participants were willing to pay for toothpaste, $F(1, 146) = 4.76$, $p = .03$, 95% CI [2.36, 3.03], with participants willing to pay more for the toothpaste in the unethical condition ($M = 2.69$) than in the ethical condition ($M = 2.42$).

Discussion

This study demonstrated no overall relationship between priming condition and ratings of cleansing products, in an online version of the task using the original items and a larger, more representative American sample. There was a significant difference in the expected direction for a single item—the toothpaste—yet this was the only significant difference among all cleansing items, both for desirability and price willing to pay. As this item-specific difference was not seen in any of the other studies reported in this article, it seems unlikely that it will turn out to be consistent or reproducible. More research is needed to confirm this conjecture.

The MTurk sample used in this study is certainly different from the university student sample used in Zhong and Liljenquist's original research (as well as in our own Study 1 carried out in the United Kingdom); however, we believe that this difference allowed us to conduct a potentially stronger (or at least more representative) test of Zhong and Liljenquist's theory than would otherwise be possible. This is because American university students are known to be at the very high end of the "WEIRD people" spectrum—that is, the spectrum of Westernized, Educated people from Industrialized, Rich Democracies (e.g., Henrich, Heine, & Norenzayan, 2010). In other words, American university students are not representative of the general population, nor even of the highly unusual WEIRD population, even within the context of the United States. By contrast, as we just noted, MTurk samples generally reflect a wider range of demographic factors.

The only other difference in this study, compared to the Zhong and Liljenquist original, was that participants were required to type out the priming passage (correcting for punctuation) rather than simply copy it by hand. We are doubtful that our failure to replicate the Macbeth Effect in this experiment can be reasonably attributed to this difference, however, as we think we may have in fact developed a *more* effective prime (i.e., by requiring the participants to engage with the actual meaning of the passage). On the other hand, work by, for example, Bargh and Chartrand (2000) suggests that the effectiveness of some primes can be mitigated by too much conscious awareness of their content. Further research is needed, therefore, to compare these two methods of prime-administration, using manipulation checks.

In sum, we were unsuccessful in replicating the Macbeth Effect using the "consumer products" paradigm, not only in the United Kingdom with slightly different materials (Study 1) but also in the United States with identical dependent measures, greater statistical power, and a more representative sample (Study 2). In our final study, we sought to investigate whether the Macbeth Effect might be detectable in a non-U.S., non-European culture that places a very high value on purity in moral discourse. Accordingly, Study 3 describes a replication effort using an Indian sample.

STUDY 3

Method

Participants. Two hundred eighty-six Indian participants (92 female, M age = 31) using the MTurk online interface participated in exchange for \$.30. Seventeen participants were excluded from analyses for failure to complete the questionnaires.

Materials and procedure. The procedure was identical to that used in Study 2. Just as in Study 1, however, consumer product brand names had to be adjusted to accommodate a non-U.S. sample. In this case, brand names were replaced with generic descriptions of each product. Accordingly, participants were asked to rate their preferences concerning: sticky notes, shower soap, toothpaste, pressed fruit juice, batteries, CD cases, glass cleaner, countertop disinfectant, a candy bar, and laundry detergent.

Results

Independent samples t tests revealed no significant difference of condition on either desirability, $t(260) = -1.83$, $p = .07$, 95% CI [–0.42, 0.02] or how much participants were willing to pay, $t(260) = -.29$, $p = .78$, 95% CI [–1.37, 1.02]. The marginal effect found for desirability of cleansing items ($p = .07$, see earlier) was actually in the opposite direction to what Zhong and Liljenquist found in their original research: Indian participants in the unethical priming condition desired cleansing items (marginally) *less* ($M = 5.25$) than participants in the ethical priming condition ($M = 5.46$). There was no effect of condition on any individual item.

Discussion

In Study 3 we failed to find a relationship between physical and moral purity in an Indian sample, even though Indian culture has been found to place a strong emphasis on purity in moral discourse (Shweder et al., 1987). Of course, it is possible that computer-using Indians who have access to the MTurk interface might be somewhat less concerned with purity than the general Indian population. Further research should be conducted to explore this possibility. Nevertheless, we report one final unsuccessful attempt to detect a Macbeth Effect using materials and methods nearly identical to those described in Study 2 of Zhong and Liljenquist (2006), this time in a

non-U.S., non-European context. We believe we are the first to demonstrate nonreplication in such a sample.

GENERAL DISCUSSION AND CONCLUSION

In 2009, researchers from two independent laboratories published an unsuccessful conceptual replication attempt of both Study 3 and Study 4 from Zhong and Liljenquist's (2006) seminal report, in the appropriately named *Journal of Articles in Support of the Null Hypothesis* (Fayard et al., 2009). This journal is one of the few available resources dedicated to combating the well-known "file drawer" problem in experimental psychology (i.e., "the strong inclination for scientific journals to selectively publish positive findings and their disinclination to publish failures to replicate and null results")—a problem that, as Harris, Coburn, Rohrer, and Pashler (2013) noted, "is increasingly recognized as harmful to the credibility of many scientific fields" (both quotes from p. 6). Nevertheless, in 2011, a second group of researchers reported failure to replicate Studies 1, 2, 3, and 4 (Gámez et al., 2011), although these experiments may have been underpowered. To our knowledge, ours is the first study to show unsuccessful direct replication, using the original materials and methods as well as consistently large samples, in the United States, United Kingdom, and India. Although we confined our own attempt to a single study—Study 2—to follow on from hypotheses in our initial pilot experiment, we would encourage other laboratories to undertake additional direct replications of all of the Macbeth Effect paradigms so that a more robust picture of the underlying effect can begin to take shape.

Before closing, we wish to stress the circumscribed nature of what our (null) findings can reasonably be taken to show. First, we do not suggest that there is no relationship between moral and physical purity in human cognition. As Chapman and Anderson (2013) argued, the body of evidence for such a relationship is large and compelling. Second, we do not claim that the Macbeth Effect, or something very like it, is somehow implausible or does not exist. Indeed, there are good theoretical reasons to posit such an effect—as Zhong and Liljenquiest argued convincingly in their original publication—and there have been a number of studies that have generated evidence that is consistent with their findings, as we noted in the Introduction. At the same time, we must caution against too much reliance on "conceptual" replications to validate an original effect. As Harris et al. (2013) argued, such studies "may [simply] exacerbate the problem of publication bias. [For] when conceptual replications succeed, they have a high likelihood of being published, whereas when they fail, they probably do not result in even so much as private skepticism of the original result" (p. 7). In addition, we caution against too much credulity regarding very large effect sizes based on small numbers of participants—at least within the goal-priming literature, where one would expect to see subtler results. Such large effect sizes, in other words, point to the distinct possibility of a false alarm. As Harris et al. stated, "This could occur if a great number of small underpowered experiments have been conducted, with only those results reaching significance having been published" (p. 2).

Taken together, these considerations call for a careful reassessment of the evidence for a real-life Macbeth Effect within the realm of moral psychology. As Scargle (2000) pointed out, meta-analytic evaluations of new findings can only be trusted "if it is known with certainty that all studies [on the construct being evaluated] that have been carried out [i.e., not just published] are included" in the statistical assessment (p. 91). Hence, well-meaning researchers who choose to leave their unsuccessful replication attempts in the proverbial "file drawer" may be unwittingly undermining the integrity of subsequent meta-analyses and systematic reviews. By resisting the temptation, therefore, to bury our own nonsignificant findings with respect to the Macbeth Effect, we hope to have contributed a small part to the ongoing scientific process.

ACKNOWLEDGMENTS

Brian D. Earp is now affiliated with the Oxford Centre for Neuroethics, University of Oxford. He and Jim A. C. Everett contributed equally to this research and should be considered as co-first authors.

REFERENCES

Anderson, M. L. (2010). Neural reuse: A fundamental organizational principle of the brain. *Behavioral and Brain Sciences, 33*, 245.

Bargh, J. A., & Chartrand, T. L. (2000). The mind in the middle: A practical guide to priming and automaticity research. In H. T. Reis & C. M. Judd (Eds.), *Handbook of research methods in social and personality psychology* (pp. 253–285). New York, NY: Cambridge University Press.

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, *6*, 3–5.

Chapman, H., & Anderson, M. L. (2013). Things rank and gross in nature: A review and synthesis of moral disgust. *Psychological Bulletin, 139*, 300–327.

Earp, B. D. (2011). Do I have more free will than you do? An unexpected asymmetry in intuitions about personal freedom. *New School Psychology Bulletin, 9*, 34–40.

Earp, B. D., Jarudi, I., Hamlin, J. K., & Madva, E. N. (2008). *Unexpected failure to replicate Zhong & Liljenquist (2006): Results from a pilot study*. Unpublished manuscript, Department of Psychology, Yale University, New Haven, CT.

Earp, B. D., Dill, B., Harris, J., Ackerman, J., & Bargh, J. A. (2010). Incidental exposure to no-smoking signs primes craving for cigarettes: An ironic effect of unconscious semantic processing? *Yale Review of Undergraduate Research in Psychology, 2*, 12–23.

Eskine, K. J. (2013). Wholesome foods and wholesome morals? Organic foods reduce prosocial behavior and harshen moral judgments. *Social Psychological and Personality Science, 4*, 251–254.

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

Fayard, J. V., Bassi, A. K., Bernstein, D. M., & Roberts, B. W. (2009). Is cleanliness next to godliness? Dispelling old wives' tales: Failure to replicate Zhong and Liljenquist (2006). *Journal of Articles in Support of the Null Hypothesis, 6*, 21–30.

Francis, G. (2012). Publication bias and the failure of replication in experimental psychology. *Psychonomic Bulletin & Review*, *19*, 975–991.

Gámez, E., Diaz, J., & Marrero, H. (2011). The uncertain universality of the Macbeth Effect with a Spanish sample. *Spanish Journal of Psychology, 14*, 156–162.

Gollwitzer, M., & Melzer A. (2012). Macbeth and the joystick: Evidence for moral cleansing after playing a violent video game. *Journal of Experimental Social Psychology, 48*, 1356–1360.

Graham, J., & Haidt, J. (2010). Beyond beliefs: Religions bind individuals into moral communities. *Personality and Social Psychology Review, 14*, 140–150.

Haidt, J., & Graham, J. (2007). When morality opposes justice: Conservatives have moral intuitions that liberals may not recognize. *Social Justice Research, 20*, 98–116.

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

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world. *Behavioral and Brain Sciences*, *33*, 61–83.

Hoy, S. (1995). *Chasing dirt: The American pursuit of cleanliness*. Oxford, UK: Oxford University Press.

Inbar, Y., Pizarro, D., Iyer, R., & Haidt, J. (2011). Disgust sensitivity, political conservatism, and voting. *Social Psychological and Personality Science, 3,* 537–544. doi:10.1177/1948550611429024

Kepes, S., Banks, G. C., McDaniel, M., & Whetzel, D. L. (2012). Publication bias in the organizational sciences. *Organizational Research Methods, 15*, 624–662.

Lee, S. W., & Schwarz, N. (2010). Dirty hands and dirty mouths: Embodiment of the moral-purity metaphor is specific to the motor modality involved in moral transgression. *Psychological Science, 21*, 1423–1425.

Lee, S. W., & Schwarz, N. (2011). Wiping the slate clean: Psychological consequences of physical cleansing. *Current Directions in Psychological Science, 20*, 307–311.

Nussbaum, M. C. (2004). *Hiding from humanity: Shame, disgust, and the law*. Princeton, NJ: Princeton University Press.

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

Reuven, O., Liberman, N., & Dar, R. (2013). The effect of physical cleaning on threatened morality in individuals with obsessive-compulsive disorder. *Clinical Psychological Science.* doi:10.1177/2167702613485565

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

Rozin, P., Haidt, J., & McCauley, C. (2000). Disgust. In M. Lewis & J. Haviland-Jones (Eds.), *Handbook of emotions* (pp. 637–653). New York, NY: Guilford.

Scargle, J. D. (2000). Publication bias: The "file-drawer" problem in scientific inference. *Journal of Scientific Exploration, 14*, 91–106.

Schnall, S., Benton, J., & Harvey, S. (2008). With a clean conscience: Cleanliness reduces the severity of moral judgments. *Psychological Science, 19*, 1219–1222.

Shweder, R., Much, N., Mahapatra, M., & Park, L. (1990). The "big three" of morality (autonomy, community, and divinity) and the "big three" explanations of suffering. In A. Brandt & P. Rozin (Eds.), *Morality and health* (pp. 119–169). New York, NY: Routledge.

Shweder, R. A., Mahapatra, M., & Miller, J. G. (1987). Culture and moral development. In J. Kagan & S. Lamb (Eds.), *The emergence of morality in young children* (pp. 1–83). Chicago, IL: University of Chicago Press.

Zhong, C., & Liljenquist, K. (2006). Washing away your sins: Threatened morality and physical cleansing. *Science, 313*, 1451–1452.