Does Cleanliness Influence Moral Judgments?

A Direct Replication of Schnall, Benton, and Harvey (2008)

David J. Johnson, Felix Cheung, and M. Brent Donnellan

Department of Psychology, Michigan State University, East Lansing, MI, USA

Abstract. Schnall, Benton, and Harvey (2008) hypothesized that physical cleanliness reduces the severity of moral judgments. In support of this idea, they found that individuals make less severe judgments when they are primed with the concept of cleanliness (Exp. 1) and when they wash their hands after experiencing disgust (Exp. 2). We conducted direct replications of both studies using materials supplied by the original authors. We did not find evidence that physical cleanliness reduced the severity of moral judgments using samples sizes that provided over .99 power to detect the original effect sizes. Our estimates of the overall effect size were much smaller than estimates from Experiment 1 (original $d = -0.60$, 95% CI [−1.23, 0.04], $N = 40$; replication $d = -0.01$, 95% CI [−0.28, 0.26], $N = 208$) and Experiment 2 (original $d = -0.85$, 95% CI [−1.47, −0.22], $N = 43$; replication $d = 0.01$, 95% CI [−.34, 0.36], $N = 126$). These findings suggest that the population effect sizes are probably substantially smaller than the original estimates. Researchers investigating the connections between cleanliness and morality should therefore use large sample sizes to have the necessary power to detect subtle effects.

Keywords: replication, embodiment, cleanliness, morality, effect size

Does cleanliness impact judgments of morality? One intriguing possibility is that individuals make less severe moral judgments when they feel clean. Schnall, Benton, et al. (2008; hereafter SBH) conducted two studies and found that participants primed with cleanliness rated moral vignettes as less wrong than participants in control conditions. They propose that feelings of cleanliness induce a sense of moral purity that is misattributed to the moral judgments following the postulates of the mood as information model (Schwarz & Clore, 1983). The goal of the present research was to replicate the results of SBH.

The research and theory underlying the SBH studies is an extension of previous research suggesting that the experience of disgust causes individuals to increase the severity of their moral judgments (Schnall, Haidt, Clore, & Jordan, 2008). According to this perspective, disgust evolved as a functional emotion for avoiding pathogens and the emotional impact of disgust has since extended to other domains (see Rozin, Haidt, & McCauley, 1999). Put simply, judgments of immorality are often tied to feelings of disgust and disgust itself may impact moral judgments (Schnall, Haidt, et al., 2008). If disgust is linked to moral impurity, this raises the possibility that cleanliness is linked with moral purity. This proposition is based on the idea that feelings of cleanliness generate psychological states that are in the opposite direction as feelings of disgust.

There is now a growing literature pointing to a connection between cleanliness and morality (for a review, see Chapman & Anderson, 2013). Zhong and Liljenquist (2006) found that cleansing oneself after recalling immoral behaviors attenuated feelings of guilt. Likewise, other studies have shown that physical cleansing attenuates post-decisional dissonance (Lee & Schwarz, 2010), reduces task performance after failure (Kaspar, 2013), and can erase feelings of bad luck (Xu, Zwick, & Schwarz, 2012). However, there is evidence that cleansing behaviors sometimes produce *harsher* moral judgments on social issues (Zhong, Strejcek, & Sivanathan, 2010). The idea is that "a clean self may feel virtuous" (Zhong et al., 2010, p. 860) thereby prompting individuals to make more severe moral judgments of others.

In short, there are competing predictions in the literature about the direction of the connection between cleanliness and moral judgments. One attempt to reconcile the results for the impact of cleanliness on moral judgments draws a distinction between *general* cleanliness and *self* cleanliness (Zhong et al., 2010). General cleanliness does not have a clearly identifiable source, making it prone to misattribution. General cleanliness can become attached to others' actions, resulting in less severe moral judgments of those actions. In contrast, when cleanliness is primed through behaviors like hand-washing, it may lead to enhanced personal feelings of virtue and thus more severe judgments of others by contrast effects. However, this explanation runs counter to the results obtained by SBH; participants who washed their hands after experiencing disgust (Exp. 2)

made less severe moral judgments than those who did not. We also point out that other studies have not found evidence for an effect of cleanliness on variables linked with morality (e.g., Earp, Everett, Madva, & Hamlin, 2014; Fayard, Bassi, Bernstein, & Roberts, 2009; Gámez, Díaz, & Marrero, 2011; see Simonsohn, 2013 for a discussion).

On top of the ambiguity surrounding the impact of cleanliness on moral judgments in light of previous research and theorizing, some of the original results from SBH are less convincing upon closer inspection. Cleanliness in Experiment 1 was primed using a scrambled-sentences task. Twenty participants were exposed to words related to cleanliness and purity and 20 participants were exposed to neutral words. Both sets of participants rated six moral vignettes. One contrast out of six reached statistical significance at the conventional $p < .05$ level. Participants in the cleanliness condition rated a vignette about sexual gratification with a kitten as less wrong than participants in the control group ($d = -0.76$, 95% CI [−1.39, −0.11]). The overall composite rating across the six vignettes generated a p value of .064 ($d = -0.60$, 95% CI [−1.23, 0.04]).[1] The results from Experiment 2 were more convincing. Participants were exposed to a disgusting video clip and then randomly assigned to a hand-washing ($n = 21$) or no hand-washing ($n = 22$) condition to manipulate feelings of cleanliness. The same six moral vignettes from Experiment 1 were used. Two of the six comparisons reached statistical significance. Participants in the cleanliness condition rated a vignette about the trolley problem in moral philosophy and a vignette about taking a wallet as less wrong than participants in the control group ($d = -0.78$, 95% CI [−1.39, −0.15] and $d = -0.79$, 95% CI [−1.40, −0.16], respectively). The overall composite was also significantly different for the two groups ($d = -0.85$, 95% CI [−1.47, −0.22]).

In sum, SBH proposed an interesting connection between cleanliness and moral judgments. This paper has attracted considerable scientific interest (the original manuscript has been cited over 150 times as of December 2013) and is part of the larger literature concerning the impact of cleanliness on moral psychology. Accordingly, it is valuable to replicate the original SBH findings in light of the original sample sizes and other studies that have had difficulties replicating the link between cleanliness and moral behaviors (e.g., Earp et al., 2014; Fayard et al., 2009; Gámez et al., 2011). We contacted Dr. Schnall who graciously offered us the materials and procedures used in the two original studies to conduct direct replications. We report all data exclusions, manipulations, and measures, and how we determined our sample sizes. The latter was determined a priori with the goal to obtain power of at least .99 to detect effect sizes for the composite variable from each of the two original experiments. Our replication studies were preregistered and all materials and data are available on the Open Science Framework website (http://osf.io/zwrxc/).

Two general deviations from the original studies are important to note, though we do not believe they have a negative impact on our ability to duplicate the original results. First, our participants were college students from a large public research university in the Midwest region of the United States whereas the participants from SBH were from the University of Plymouth in the United Kingdom. Second, we included the private body consciousness subscale (PBC; Miller, Murphy, & Buss, 1981) after all other experimental procedures (i.e., after participants evaluated the moral vignettes). Schnall, Haidt, and colleagues (2008) demonstrated that the priming effects of disgust on moral judgments were moderated by sensitivity to bodily sensations. Participants with high levels of PBC were more likely to make more severe moral judgments than participants with low levels of PBC. As an extension of this result, we expected that participants primed with cleanliness who had high levels of PBC would make less severe moral judgments than participants with low levels of PBC.

Experiment 1

Power Analysis and Sample Characteristics

We used the point estimate of effect size $d = -0.60$ from the composite to compute statistical power. Assuming equal sized groups, we needed at least 208 participants to achieve .99 power (104 participants in each group). Thus, we collected data from 219 Michigan State University undergraduates, 76.7% of which were females, $M_{age} = 19.5$ years, $SD = 2.4$ (compare to SBH's Exp. 1: 75% female, $M_{age} = 20.0$ years, $SD = 1.9$). Participants received partial fulfillment of course requirements or extra credit for their participation. Eleven participants were removed for admitting to fabricating their answers, failing to correctly complete the scrambled sentence task, or for experimenter error. Analyses were conducted on the remaining 208 participants. These exclusion rules were determined a priori and included in the preregistration materials. Analyses including the 11 participants do not change the results or interpretations reported here.

Procedure

The procedure was identical to SBH's Experiment 1 with the addition of the 5-item PBC scale to the end of the experiment ($\alpha = .46$, $M = 2.62$, $SD = 0.61$). Participants completed the study in individual sessions. Participants provided informed consent and then received a sealed packet that contained all tasks and instructions. They first completed a scrambled sentence task that involved either neutral words (control condition; $n = 102$) or cleanliness

[1] Experiment 1 was conceptually replicated by Besman, Dubensky, Dunsmore, and Daubman (2013) using 60 participants. These researchers used different words in the priming task and their overall results for the composite did not reach conventional levels of significance with a two-tailed test ($p = .08$). Details about specific vignettes were not reported.

Table 1. Mean ratings of moral vignettes in Experiment 1

		Dog		Trolley		Wallet		Plane crash		Resume		Kitten		Total	
		SBH	Rep.	SBH	Rep.	SBH	Rep.	SBH	Rep.	SBH	Rep.	SBH	Rep.	SBH	Rep.
Cleanliness	*M*	5.70	7.37	1.85	2.88	4.95	6.95	6.05	6.87	4.65	6.92	6.70	7.84	4.98	6.47
	SD	2.39	2.27	1.50	2.00	2.35	2.04	2.39	2.57	2.28	2.04	2.49	2.04	1.26	1.12
Neutral	*M*	6.55	7.26	2.75	2.99	5.45	7.02	6.45	7.13	5.40	6.75	8.25	7.74	5.81	6.48
	SD	2.52	2.33	2.38	2.00	2.86	2.81	2.56	2.16	2.26	2.08	1.48	1.84	1.47	1.13
	Cohen's *d*	−0.35	0.04	−0.45	−0.06	−0.19	−0.03	−0.16	−0.11	−0.33	0.08	−0.76*	0.05	−0.60†	−0.01
	d_{LL}	−0.97	−0.23	−1.08	−0.33	−0.81	−0.30	−0.78	−0.38	−0.95	−0.19	−1.39	−0.22	−1.23	−0.28
	d_{UL}	0.28	0.32	0.18	0.22	0.43	0.24	0.46	0.16	0.30	0.35	−0.11	0.33	0.04	0.26

Notes. Response scales ranged from 0 (*perfectly OK*) to 9 (*extremely wrong*). SBH = Experiment 1 (N = 40), Schnall, Benton, et al. (2008). Rep. = Current replication (N = 208), d_{LL} = Lower limit of the 95% CI for Cohen's d, d_{UL} = Upper limit of the 95% CI for Cohen's d. *$p < .05$; $^{\dagger}p < .10$.

related words (cleanliness condition; n = 106). Participants then responded to six vignettes describing moral dilemmas on 10-point scales ranging from 0 (*perfectly OK*) to 9 (*extremely wrong*). A composite score was created by averaging responses to all six dilemmas. Finally, participants gave self-report measures of their current emotions and completed the PBC scale. Research assistants were blind to condition to prevent the possibility of expectancy effects biasing participant responses (Doyen, Klein, Pichon, & Cleeremans, 2012; Klein et al., 2012).

Results

We first tested whether priming influenced participants' self-reported emotions. A series of one-way ANOVAs did not provide evidence that emotions varied based on condition (all ps > .09), consistent with the original experiment. The focal comparisons involved tests of whether the cleanliness prime reduced the severity of participants' judgments of the moral dilemmas, using a series of one-way ANOVAs. We did not find statistically significant effects for the overall composite, $F(1, 206) = 0.004$, $p = .95$, $d = -0.01$, 95% CI [−.28, .26]. Analyses of individual vignettes also yielded null results (see Table 1) including the "kitten" dilemma ($d = 0.05$, $p = .72$, 95% CI [−.22, .33]), the only vignette that yielded a statistically significant difference at $p < .05$ in the original experiment.

We conducted an additional series of analyses to evaluate the role of PBC as a moderator of the cleanliness effect. Moral judgments were regressed onto condition, PBC score (continuous, mean-centered) and their interaction (mean-centered) there was no evidence of a statistical significant interaction at $p < .05$. We also followed the procedures used in Schnall, Haidt, and colleagues (2008) by dividing participants into high and low PBC groups by median splits and conducting ANOVAs on the groups. All main effects and PBC × Prime interactions were nonsignificant for the mean composite and all individual dilemmas; only one PBC × Prime interaction approached significance (the résumé dilemma, $p = .07$). However, this interaction ran counter to predictions as participants low in PBC provided lower ratings than participants high in PBC. Median split approaches have well-known methodological problems (MacCallum, Zhang, Preacher, & Rucker, 2002) and thus we place more emphasis on the regression-based analyses.

Discussion

We found no evidence that participants primed with cleanliness judged morally questionable actions as less wrong than participants primed with neutral words. These null results were consistent across all vignettes (range of ds = −0.11, 95% CI [−.38, .16] to 0.08, 95% CI [−.19, .35]) and regardless of whether participants were filtered based on suspicion.[2] In addition, we found no evidence that PBC, a measure of sensitivity to bodily sensations, moderated the effect of the cleanliness prime on judgments of morality. The one caveat is that this measure has a fairly low level of internal consistency and this may have attenuated our ability to detect these moderator effects. However, neither Schnall and colleagues (2008) nor Miller and colleagues (1981) reported the reliability of their PBC scale scores, making it unclear if our alpha value was unusually low.

In general, we found little evidence linking cleanliness to moral judgments. However, the manipulation in this study was fairly subtle and this may have impacted our ability to detect effects. For example, the manipulation may not have provided substantial enough bodily sensations to make the test of the PBC moderator compelling. These effects might be easier to detect if physical cleanliness were manipulated directly. Indeed, Experiment 2 is arguably a stronger test of SBH's central hypothesis because the act of actual cleansing is manipulated. From an embodied cognition perspective, Experiment 2 is a more direct evaluation of whether there is a strong automatic connection between physical cleanliness and moral judgments.

[2] Before examining the data, we devised three filters of increasing sensitivity for removing participants based on their level of suspicion (syntax files are available on the Open Science Framework website). Analyses were rerun using each filter. Excluding these participants from the analyses does not change the significance of any result.

Experiment 2

Power Analysis and Sample Characteristics

We used the point estimate of effect size $d = -0.85$ from the composite to compute statistical power. Assuming equal sized groups, we needed at least 104 participants to achieve .99 power (52 in each group). Thus, we collected data from 132 Michigan State University undergraduates, 70.5% of which were females, $M_{age} = 20.5$ years, $SD = 3.6$ (compare to SBH's Exp. 2: 73% female, $M_{age} = 22.2$ years, $SD = 4.9$). Participants received partial fulfillment of course requirements or extra credit for their participation. Eight participants were removed for admitting to fabricating their answers or for experimenter error. Analyses were conducted on the remaining 126 participants but results and interpretations are unchanged when these eight participants are included.

Procedure

The procedure followed SBH's Experiment 2 with the addition of the PBC scale to the end of the experiment ($\alpha = .62$, $M = 2.50$, $SD = 0.70$). Participants completed tasks in individual sessions. They first watched a video that invoked disgust (the same clip from *Trainspotting* used by SBH). Participants were randomly asked to either wash their hands (cleanliness condition; $n = 58$) or given no prompt (control condition; $n = 68$). Participants then responded to the same six vignettes describing moral dilemmas on 7-point scales ranging from 1 (*nothing wrong at all*) to 7 (*extremely wrong*). A composite score was created by averaging responses to all six dilemmas. Finally, participants gave self-report measures of the emotions felt directly after watching the disgusting video, and completed the PBC scale. One additional modification was made to the original procedure with respect to the location of physical cleansing. The staff room in our facility did not include a sink. Thus, participants were asked to wash their hands at a sink next to the staff room. Participants in the original study washed their hands in the same room where they responded to the moral vignettes. We do not believe this difference should impact our ability to replicate the original finding as we detail below.

Results

We tested whether participants experienced disgust more than any other emotion after watching the video using repeated-measures ANOVA. No differences were found as a result of condition, $F(1, 122) = 1.75$, $p = .19$, $\eta^2 = .01$, and there was no evidence of a Condition × Emotion interaction, $F(8, 976) = 0.67$, $p = .72$, $\eta^2 = .01$, consistent with the original experiment. Disgust ratings ($M = 18.55$, $SD = 3.63$) were significantly higher than all other emotion ratings, such as anger ($M = 4.26$, $SD = 4.73$) and sadness ($M = 5.58$, $SD = 5.36$), all $ps < .001$. There was also no evidence of differences in self-recalled disgust after watching the video (prior to hand-washing) between individuals in the cleanliness and control conditions, $F(1, 124) = 0.14$, $p = .71$ ($d = -0.07$, 95% CI [−.42, .28]).

The focal comparisons involved tests of whether hand-washing reduced the severity of moral judgments using a series of one-way ANOVAs. We did not find statistically significant effects for the overall composite, $F(1, 124) = 0.001$, $p = .97$, $d = 0.01$, 95% CI [−.34, .36]. Analyses of individual vignettes also yielded null results (see Table 2) including the "trolley" and "wallet" dilemmas ($d = 0.08$, 95% CI [−.27, .43] and −0.11, 95% CI [−.46, .24], respectively), both vignettes that were statistically significant in the original experiment. We also tested whether PBC moderated the experimental effects. As with our analyses for Experiment 1, moral judgments were regressed onto condition, PBC score (continuous, mean-centered) and their interaction (mean-centered). There was no evidence of a statistically significant interaction at $p < .05$. We also followed the median split procedures used in Schnall, Haidt, and colleagues (2008) to supplement these analyses. All main effects and PBC × Prime interactions were nonsignificant for the mean composite and all individual vignettes. One PBC main effect approached significance (the résumé dilemma, $p = .08$) such that individuals with higher PBC tended to rate the résumé dilemma more severely, regardless of the cleanliness manipulation.

Table 2. Mean ratings of moral vignettes in Experiment 2

		Dog		Trolley		Wallet		Plane crash		Resume		Kitten		Total	
		SBH	Rep.	SBH	Rep.	SBH	Rep.	SBH	Rep.	SBH	Rep.	SBH	Rep.	SBH	Rep.
Cleanliness	*M*	5.33	5.97	2.81	3.57	4.62	5.97	5.38	6.05	4.24	5.97	6.00	6.43	4.73	5.66
	SD	1.88	1.49	1.08	1.38	1.53	1.34	1.80	1.38	1.67	1.20	1.18	1.13	0.95	0.59
Neutral	*M*	5.73	5.84	3.64	3.46	5.73	6.12	6.05	6.29	5.09	5.74	6.36	6.49	5.43	5.65
	SD	0.98	1.4	1.05	1.41	1.28	1.36	1.21	1.09	1.15	1.29	1.00	0.87	0.67	0.68
	Cohen's *d*	−0.26	0.09	−0.78*	0.08	−0.79*	−0.11	−0.43	−0.20	−0.60†	0.18	−0.33	−0.05	−0.85**	0.01
	d_{LL}	−0.86	−0.26	−1.39	−0.27	−1.40	−0.46	−1.04	−0.55	−1.21	−0.17	−0.93	−0.40	−1.47	−0.34
	d_{UL}	0.34	0.44	−0.15	0.43	−0.16	0.24	0.17	0.16	0.02	0.54	0.27	0.30	−0.22	0.36

Notes. Response scales ranged from 0 (*nothing wrong at all*) to 7 (*extremely wrong*). SBH = Experiment 2 ($N = 43$), Schnall, Benton, et al. (2008). Rep. = Current replication ($N = 126$), d_{LL} = Lower limit of the 95% CI for Cohen's *d*, d_{UL} = Upper limit of the 95% CI for Cohen's *d*. *$p < .05$; **$p < .01$; $^{\dagger}p < .10$.

Discussion

We found no evidence that hand-washing after experiencing disgust led participants to judge moral vignettes differently than participants who did not wash their hands. These null results were consistent across all vignettes (range of *d*s = −0.20, 95% CI [−.55, .16] to 0.18, 95% CI [−.17, .54) and regardless of whether participants were filtered based on suspicion.[3] Overall, we found virtually no difference between conditions for the composite variable (d = 0.01, 95% CI [−.34, .36]). In short, we found little support for the idea that cleansing behaviors impact moral judgments. These results not only contrast with predictions made by SBH, but also with potentially opposing predictions that physical self-cleansing should lead to more severe moral judgments (Zhong et al., 2011). We also found no indication that private body consciousness moderated the impact of the cleanliness manipulations.

We should emphasize one potential difference between the original study and our replication study in terms of the experimental setting. As we noted, the sink in our study was outside of the room where participants completed the moral vignettes. It is possible that our modification to the original procedure might have attenuated the effects to some degree. Nonetheless, we believe the act of cleansing is the psychologically important ingredient in the manipulation rather than the location of the sink. The one qualifier is that the presence of a sink in the room might also prime cleanliness. In the original SBH procedure, participants in both conditions completed the vignettes in the staff room with the sink. However, SBH observed differences between hand-washing and control groups even though both were exposed to the same sink. If the mere presence of the sink was sufficient to prime cleanliness, it should have reduced the magnitude of the difference between the groups in the original study. If this were the case, the absence of the sink from our staff room should arguably strengthen our manipulation. Furthermore, if the original experimental effects were dependent on the visibility of the sink, it would undermine the idea that the cleansing effects are driven by a purely embodied process.

General Discussion

The idea that cleanliness impacts moral judgments is interesting because of its links to the embodiment literature and with research on the intuitive and nonrational contributors to moral judgments. Cleanliness findings may even have practical applications. These reasons motivated our replication studies of the two experiments reported in Schnall, Benton, and colleagues (2008). We used the same materials and nearly identical procedures with the exception of the location of the sink for the replication of Study 2. Sample size was determined based on the goal of having 99% power to detect the original effect size estimates (for the composite variables) and our attempts were preregistered (see http://osf.io/zwrxc/). Although our results are inconsistent with the results of SBH, they are extremely consistent with one another. Both experiments yielded point estimates of the effect that were centered on zero for the composite variable (d = −0.01, 95% CI [−.28, .26] and d = 0.01, 95% CI [−.34, .36] for Exp. 1 and 2, respectively).

Evaluation of Replication Results

The current results are seemingly compatible with a growing body of research that calls into question the strength of the association of cleanliness manipulations for outcomes in the moral domain (Earp et al., 2014; Fayard et al., 2009; Gámez et al., 2011; Siev, 2012). Nonetheless, we acknowledge that there are ongoing controversies about how researchers should interpret results of replication studies that are inconsistent with original studies (Aspendorf et al., 2013). To help address these sorts of issues, Simonsohn (2013) developed a framework for interpreting replication studies based on the effect size estimates and sample sizes of the original study. Replications that obtain effect sizes significantly smaller than $d_{33\%}$ (i.e., an effect size that the original study would have had only a 33% chance of detecting) are "informative failures" and indicate that the effect size was too small for the original study to have reliably detected. We used this framework to interpret the findings of our replications.

Specifically, we analyzed the effect size estimates from all known replication attempts (including ours) in relation to SBH's Experiments 1 and 2. For Experiment 1, the original point effect size estimate (d = −0.60, 95% CI [−1.23, 0.04]) yields a $d_{33\%}$ of d = −0.50. In other words, the sample size of the original study had 33% power to detect an effect size of d = −0.50 with a sample size of 40. The unpublished Besman, Dubensky, Dunsmore, and Daubman (2013) obtained a point effect size estimate of d = −0.47 (95% CI [−.98, .05]) with a sample size of 60. This result does not significantly differ from $d_{33\%}$ (p = .19). Since the effect size is not smaller than $d_{33\%}$ nor different from 0 (p = .08), the Besman and colleagues replication attempt can be classified as an uninformative replication (Simonsohn, 2013). Our replication point effect size estimate was d = −0.01 (95% CI [−.28, .26]), which is significantly smaller than the referent $d_{33\%}$ (p < .001; see Figure 1) and thus would be considered an informative failure to replicate (Simonsohn, 2013). For Experiment 2, the original point effect size estimate (d = −0.85, 95% CI [−1.47, −0.22]) yields a $d_{33\%}$ of d = 0.47. Our replication point effect size estimate (d = 0.01, 95% CI [−.34, .36]) is significantly smaller than $d_{33\%}$, p = .004 (see Figure 2) and is also considered an informative failure to replicate.

[3] Before examining the data, we again devised three filters of increasing sensitivity for removing participants based on their level of suspicion. Analyses were rerun using each filter. Excluding these participants from the analyses does not change the significance of any result.

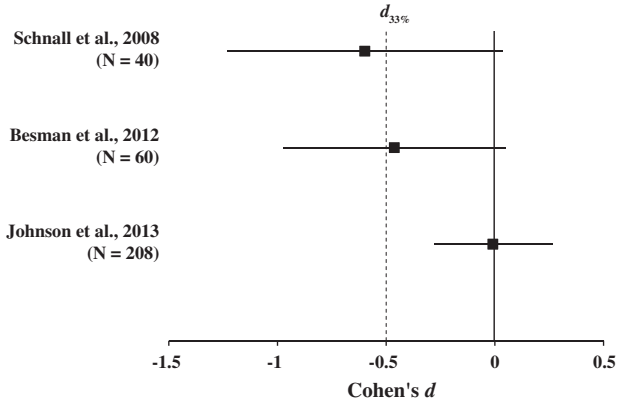

Figure 1. Results from Study 1 by Schnall, Benton, and Harvey and its replications. Markers report effect size (Cohen's *d*) and horizontal bars their 95% confidence intervals. The dashed line indicates the effect size ($d = -0.50$) that would give the original study, with $N = 40$, 33% power. According to Simonsohn's (2013) $d_{33\%}$ standard, our replication is an "informative failure."

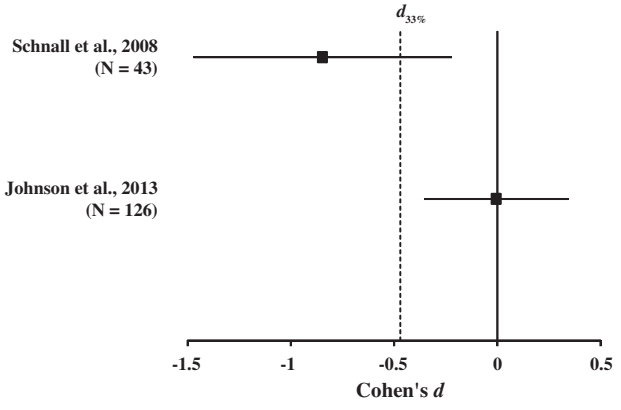

Figure 2. Results from Study 2 by Schnall, Benton, and Harvey and our replication. Markers report effect size (Cohen's *d*) and horizontal bars their 95% confidence intervals. The dashed line indicates the effect size ($d = 0.47$) that would give the original study, with $N = 43$, 33% power. According to Simonsohn's (2013) $d_{33\%}$ standard, our replication is an "informative failure."

In short, the current results suggest that the underlying effect size estimates from these replication experiments are substantially smaller than the estimates generated from the original SBH studies. One possibility is that there are unknown moderators that account for these apparent discrepancies. Perhaps the most salient difference between the current studies and the original SBH studies is the student population. Our participants were undergraduates in United States whereas participants in SBH's studies were undergraduates in the United Kingdom. It is possible that cultural differences in moral judgments or in the meaning and importance of cleanliness may explain any differences. On the other hand, the original authors argued that the automatic connection between disgust and bodily sensation is an evolved adaptation and did not raise the possibility that results would differ across samples drawn from different western populations. The United States and the United Kingdom are similar in terms of language and cultural traditions, and past studies have found a relationship between disgust and moral judgment in samples from the United States (e.g., Schnall, Haidt, et al., 2008). Thus, it seems unlikely that sample differences are a viable explanation for our discrepant results. However, this is ultimately an empirical question and a number of other unknown variables might have impacted the results. Accordingly, future studies should attempt replications of the SBH effects and test for theoretically motivated boundary conditions.

Conclusions

The gold standard of reliability in all sciences is replication. Independent researchers following the same script in different labs should be able to find evidence consistent with the original results of an experiment (Frank & Saxe, 2012). Although replication is an important part of the science of psychology, many of the incentives in the field do not encourage replication studies (e.g., Nosek, Spies, & Motyl, 2012). The purpose of this special issue is to change these incentives. Publication decisions were not predicated on the results of the replication studies per se so there is less motivation to find a particular result. This strikes us as a very positive example for the field.

Regardless of the success or failure of any replication attempt, this kind of research increases the precision of effect size estimates for the field. Thus, although failures to replicate are not always satisfying, they do provide important information to the body of knowledge in psychology. This point about the importance of additional information is the one we wish to emphasize. Our work simply provides more information about an interesting idea. The current studies suggest that the effect sizes surrounding the impact of cleanliness on moral judgments are probably smaller than the estimates provided by the original studies. Researchers attempting future work in this area should use fairly large sample sizes to have the power to detect subtle but perhaps important effects (say a *d* of 0.10 or smaller). It is critical that our work is not considered the last word on the original results in SBH and we hope there are future direct replications of the original results using populations drawn from many different countries. More broadly, we hope that researchers will continue to evaluate the emotional factors that contribute to moral judgments.

Note From the Editors

A commentary and a rejoinder on this paper are available (Johnson, Cheung, & Donnellan, 2014; Schnall, 2014; doi: 10.1027/1864-9335/a000204).

Acknowledgments

This research was supported by a Graduate Research Fellowship from the National Science Foundation awarded to the second author. Designed research: D. J. J., F. C.; Performed research: D. J. J., F. C.; Analyzed Data: D. J. J., F. C., M. B. D.; Wrote paper: D. J. J., F. C., M. B. D. The authors declare no conflict-of-interest with the content of this article. We are also very grateful for the help from Jason Lam, a chemistry doctoral student, who prepared chemical solutions to clean the sink used in Experiment 2. All *d*-metric effect size estimates and confidence intervals were generated using Stata 13.1. The study reported in this article earned *Open Data, Open Materials,* and *Preregistered* badges. All materials, data, and the preregistered design are available at: http://osf.io/zwrxc/.

References

Asendorpf, J. B., Conner, M., De Fruyt, F., De Houwer, J., Denissen, J. J. A., Fiedler, K., … Wicherts, J. M. (2013). Recommendations for increasing replicability in psychology. *European Journal of Personality, 27*, 108–119.

Besman, M., Dubensky, C., Dunsmore, L., & Daubman, K. (2013). *Cleanliness primes less severe moral judgments.* Retrieved from http://www.psychfiledrawer.org/replication.php?attempt=MTQ5

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

Doyen, S., Klein, O., Pichon, C. L., & Cleeremans, A. (2012). Behavioral priming: it's all in the mind, but whose mind? *PloS one, 7*, e29081.

Earp, B. D., Everett, J. A. C., Madva, E. N., & Hamlin, J. K. (2014). Out, damned spot: Can the "Macbeth Effect" be replicated? *Basic and Applied Social Psychology, 36*, 91–98.

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 & Liljenquist (2006). *Journal of Articles in Support of the Null Hypothesis, 6*, 21–30.

Frank, M. C., & Saxe, R. (2012). Teaching replication. *Perspectives on Psychological Science, 7*, 600–604.

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

Johnson, D. J., Cheung, F., & Donnellan, M. B. (2014). Hunting for artifacts: The perils of dismissing inconsistent replication results. Commentary and rejoinder on Johnson, Cheung, and Donnellan (2014). *Social Psychology*. Advance online publication. doi: 10.1027/1864-9335/a000204

Kaspar, K. (2013). Washing one's hands after failure enhances optimism but hampers future performance. *Social Psychological and Personality Science, 4*, 69–73.

Klein, O., Doyen, S., Leys, C., Magalhães de Saldanha da Gama, P. A., Miller, S., Questienne, L., & Cleeremans, A. (2012). Low hopes, high expectations: Expectancy effects and the replicability of behavioral experiments. *Perspectives on Psychological Science, 7*, 572–584. doi: 10.1177/1745691612463704

Lee, S. W., & Schwarz, N. (2010). Washing away postdecisional dissonance. *Science, 328*, 709–709.

MacCallum, R. C., Zhang, S., Preacher, K. J., & Rucker, D. D. (2002). On the practice of dichotomization of quantitative variables. *Psychological Methods, 7*, 19–40.

Miller, L. C., Murphy, R., & Buss, A. (1981). Consciousness of body: Private and public. *Journal of Personality and Social Psychology, 41*, 397–406.

Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific utopia II. Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science, 7*, 615–631.

Rozin, P., Haidt, J., & McCauley, C. R. (1999). Disgust: The body and soul emotion. In T. Dalgleish & M. J. Power (Eds.), *Handbook of cognition and emotion* (pp. 429–445). New York, NY: Wiley.

Schnall, S. (2014). Clean data: Statistical artefacts wash out replication efforts. Commentary and rejoinder on Johnson, Cheung, and Donnellan (2014). *Social Psychology*. Advance online publication. doi: 10.1027/1864-9335/a000204

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

Schnall, S., Haidt, J., Clore, G. L., & Jordan, A. H. (2008). Disgust as embodied moral judgment. *Personality and Social Psychology Bulletin, 34*, 1096–1109.

Schwarz, N., & Clore, G. L. (1983). Mood, misattribution, and judgments of well-being: Informative and directive functions of affective states. *Journal of Personality and Social Psychology, 45*, 513–523.

Siev, J. (2012). *Unpublished experimental results attempting to replicate Zhong & Liljenquist*.

Simonsohn, U. (2013). *Evaluating replication results.* Retrieved from http://dx.doi.org/10.2139/ssrn.2259879

Xu, A. J., Zwick, R., & Schwarz, N. (2012). Washing away your (good or bad) luck: Physical cleansing affects risk-taking behavior. *Journal of Experimental Psychology: General, 141*, 26–30.

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

Zhong, C. B., Strejcek, B., & Sivanathan, N. (2010). A clean self can render harsh moral judgment. *Journal of Experimental Social Psychology, 46*, 859–862.

Received February 28, 2013
Accepted November 27, 2013
Published online May 19, 2014

David J. Johnson

Department of Psychology
316 Physics
Rm 244C
Michigan State University
East Lansing, MI 48824
USA
E-mail djjohnson@smcm.edu