Hunting for Artifacts: The Perils of Dismissing Inconsistent Replication Results

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

Michigan State University

Author Note

David J. Johnson, Department of Psychology, Michigan State University; Felix Cheung, Department of Psychology, Michigan State University; M. Brent Donnellan, Department of Psychology, Michigan State University.

This research was supported by a Graduate Research Fellowship from the National Science Foundation awarded to the second author. All data are available at: https://osf.io/jxad3/. All authors contributed equally to designing and performing the research, analyzing the data, and writing the paper. The authors declare no conflict-of-interest with the content of this article. We would also like to thank Daniël Lakens, Richard Lucas, Hal Pashler, and Daniel Simons for their helpful comments.

Correspondence concerning this article should be addressed to David J. Johnson, Department of Psychology, 316 Physics, Rm 244C, Michigan State University, East Lansing, MI 48824. Email: djjohnson@smcm.edu

Abstract

We attempted high-powered direct replications of the two experiments in Schnall, Benton, and Harvey (2008) and did not duplicate the original results. We therefore concluded that more research was needed to establish the size and robustness of the original effects and to evaluate potential moderators. Schnall (2014) suggests that our conclusions were invalid because of potential psychometric artifacts in our data. We present evidence that undermines concerns about artifacts and defend the utility of pre-registered replication studies for advancing research in psychological science.

Keywords**:** Cleanliness; Moral Judgment, Pre-Registration, Replication

Hunting for Artifacts: The Perils of Dismissing Inconsistent Replication Results

We attempted high-powered direct replications of the two experiments in Schnall, Benton, and Harvey (2008; hereafter SBH) using the same measures with nearly identical procedures. The major difference was that we used larger student samples taken from a different country. We did not duplicate the original results and concluded that more research was needed to establish the size and robustness of the original effects. We also suggested that more work was needed to evaluate potential moderators. Our efforts were pre-registered and no objections about the procedures, measures, or nature of the samples were raised by Dr. Schnall during the proposal review stage. However, Schnall (2014) suggests that our conclusions are invalid. We do not share this pessimistic view and believe that the Schnall (2014) commentary illustrates the pitfalls of criticizing replication studies after the results are known.

Ceiling Effects and Moderators

Schnall (2014) believes that ceiling effects may have prevented us from duplicating the original SBH results. She further suggests that parametric statistical analyses are inappropriate. First, non-parametric tests on each item (i.e., Mann-Whitney U tests) yielded the same conclusions as the parametric analyses reported in our paper. Moreover, we emphasize that there was no a priori reason to suspect that the SBH dependent variables would be inappropriate for use with college students from Michigan because they had been originally been developed for use with college students from Virginia (see Study 2 in Schnall, Haidt, Clore, & Jordan, 2008).[1] The ceiling effect concern is also far less relevant to the summary composite variables (the focal outcome) because extreme item responses tended to be washed out in the aggregate. If the composite variables were at ceiling, we should not have been able to detect gender differences in moral judgments. However, we replicated the effect that women tended to give harsher

judgments than men in both studies (Study 1: $t(206) = 2.47$, $p = .014$, $d = 0.41$, 95% CI [0.09, 0.74]; Study 2: $t(124) = 3.46$, $p < .001$, $d = 0.69$, 95% CI [0.29, 1.09]).

One way to directly address the ceiling concern is to remove participants from both the control and cleanliness conditions who selected the most extreme response for each scenario and to repeat the analyses on an item-by-item basis. If as Schnall (2014) suggests, our null results were due to decreased variance solely because many responses were at ceiling, removing these extreme responses from the analyses should reduce skew, eliminating any bias it may have introduced in our significance tests. Although this approach produced a loss of power, our resulting samples were still larger than the original SBH studies except in one case (the Kitten scenario, Study 2), as demonstrated in Table 1. Importantly, no comparisons attained statistical significance when using this approach, bolstering support that our null results were not simply due to ceiling effects.

We also compared our respective datasets to determine the proportion of extreme responses in the control conditions, as responses in these conditions serve as a baseline for how immoral the scenarios are without experimental manipulation. We focused on scenarios that produced statistically significant effects in the original SBH studies because these are the only relevant comparisons.[2] Significance tests revealed only one relevant scenario with a different distribution between our respective studies (the Wallet scenario, Study 2; $\chi^2(2) = 4.34$, $p = .037$). Moreover, we had a similar proportion of extreme responses ($\chi^2(2) = 1.37$, $p = .242$) in the Kitten scenario for Study 1 (56% of our control participants responded with a "9" compared to 70% in SBH). This was the only scenario that showed a significant difference in Study 1 of SBH. Extreme responding did not prevent SBH from finding supportive evidence for this scenario in

their Study 1 so we are unsure why it would have prevented us from finding similar evidence in our work.

All told, we do not find the psychometric concerns raised by Schnall (2014) compelling. Nonetheless, it is still possible that there are moderators of the original findings in terms of political orientation as suggested by Schnall (2014). To test this possibility, we conducted a large-scale online replication of SBH Study 1 (n = 736) using students drawn from the same student population as our Study 1. We also included a measure of political conservativism.[3] Consistent with our published replication of Study 1, we found no effect of condition on the moral composite, $t(734) = -0.65$, $p = .518$, $d = -0.05$, 95% CI [-0.19, 0.10]. No supportive evidence was found when testing any of the individual scenarios, and these conclusions held when extreme responses from both the control and cleanliness conditions were removed.

As Schnall (2014) predicted, we found that students who identified as conservative were more likely to rate the moral scenarios more harshly ($r = .11$, $p = .002$). However, regressing the moral composite on conservatism (centered), condition, and their product term (centered) produced no indication of a statistical interaction, $b = -.04$, $t(735) = -0.52$, $p = .60$. Moreover, there was little evidence that this student sample was excessively conservative (43.0% identified somewhere on the liberal spectrum whereas 27.9% identified somewhere on the conservative spectrum). While it is possible that the manipulation of cleanliness (as primed by scrambled sentence task) is not effective online as Schnall (2014) suggests, several researchers have successfully found priming effects when using scrambled sentence tasks with online samples (e.g., Preston & Ritter, 2012; Gino & Mogliner, 2014; Kay, Laurin, Fitzsimmons, & Landau, 2013). In sum, this second failure to replicate SBH strengthens our confidence in our published

results and undercuts the suggestion in Schnall (2014) that college students from Michigan are especially conservative.

Pre-Registration and Peer-Review

Schnall (2014) expressed reservations that our pre-registered replication did not have sufficient post-data collection peer review. We do not share this concern. Our proposal was evaluated with respect to the rigor of its methods rather than the actual results of the studies. The innovative procedure used for this special issue (Nosek & Lakens, 2014) is based on the belief that any well-designed study (e.g., an adequately powered study with appropriate measures) provides useful information regardless of the specific findings. In line with this perspective, we see no reason to suppress our results simply because of possible concerns over the distributions of our variables. As investigators, we had no control over the distributions and the distributions themselves provide valuable information for the field about the generalizability of the original findings.

Now that the data from our replication studies have been collected, analyzed, and re-analyzed, the field has gained some additional insights about the robustness of the SBH results. Perhaps our studies might prompt revision of the original measures to make them more sensitive to seemingly subtle priming effects for future investigations. Our studies might even cause some researchers to revise their expectations about the underlying effect sizes. Both of these would be reasonable reactions to our research and neither of these outcomes strikes us as undesirable.

In sum, nothing in Schnall (2014) makes us question our original conclusion that more research with larger sample sizes is needed to determine the precise link between cleanliness and morality examined in the SBH studies. No two studies are perfectly identical so it will always be possible to point to some issue that might explain discrepant results. The relevant question is

whether such post-hoc speculations have merit and we believe this question is best addressed with more research. In the end, we hope the field will not dismiss well-designed and pre-registered replication results simply because the results were inconsistent with the original findings.

References

Gino, F., & Mogilner, C. (2014). Time, money, and morality. *Psychological Science. 25*(2), 414-421.

Johnson, D. J., Cheung, F., & Donnellan, M. B. (2014). Does cleanliness influence moral judgments? A direct replication of Schnall, Benton, and Harvey (2008). *Social Psychology.*

Kay, A. C., Laurin, K., Fitzsimons, G. M., & Landau, M. J. (2013). A functional basis for structure-seeking: Exposure to structure promotes willingness to engage in motivated action. *Journal of Experimental Psychology: General*, 143, 486-491.

Nosek, B. A., & Lakens, D. (2014). Registered reports: A method to increase the credibility of published results. *Social Psychology*.

Preston, J. L., & Ritter, R. S. (2012). Cleanliness and godliness: Mutual association between two kinds of personal purity. *Journal of Experimental Social Psychology*, 48, 1365-1368.

Schnall, S. (2014). Clean data: Statistical artifacts wash out replication efforts. *Social Psychology.*

Schnall, S., Benton, J., & Harvey, S. (2008). With a clean conscience cleanliness reduces the severity of moral judgments. *Psychological Science*, *19*(12), 1219-1222.

Schnall, S., Haidt, J., Clore, G. L., & Jordan, A. H. (2008). Disgust as embodied moral judgment. *Personality and Social Psychology Bulletin*, *34*(8), 1096-1109.

Footnotes

[1] Schnall, Haidt, Clore, and Jordan (2008) conducted an 8 person pilot study to select "six scenarios that generated substantial variance among respondents (i.e., that avoided floor and ceiling effects)" (p. 1100).

[2] The proportion of extreme responses in the control condition between SBH and our replication differed for three scenarios that did not show the predicted effects in the original study at $p < .05$: Wallet (Study 1), Dog (Study 2), and Plane (Study 2).

[3] All data exclusions, manipulations, and measures (with the addition of measures of conservatism, disgust sensitivity, and honesty/humility) were determined using the standards used by Johnson et al. (2014). We obtained a much larger sample size to test for moderator effects and to detect population effect sizes smaller than the published research.

Table 1

Effect of Condition on Severity of Moral Judgments When Removing Extreme Responses

		Study 1					Study 2				
Scenario	Condition	*N*	*M*	*SD*	*t*	*p*	*N*	*M*	*SD*	*t*	*p*
Dog	Neutral	48	5.31	2.09	-0.99	.327	38	4.92	1.26	-0.03	.977
	Cleanliness	53	5.74	2.22			29	4.93	1.51		
Trolley	Neutral	100	2.87	1.83	0.96	.339	67	3.40	1.35	-0.19	.854
	Cleanliness	102	2.64	1.62			56	3.45	1.23		
Wallet	Neutral	57	5.46	1.72	-0.85	.396	29	4.93	1.36	0.01	.999
	Cleanliness	66	5.71	1.61			29	4.93	1.19		
Plane	Neutral	58	5.71	1.86	1.21	.228	27	5.22	1.05	1.03	.310
	Cleanliness	60	5.23	2.35			26	4.88	1.34		
Resume	Neutral	72	5.82	1.77	-0.13	.897	42	4.95	1.03	-0.72	.475
	Cleanliness	70	5.86	1.71			32	5.13	1.01		
Kitten	Neutral	45	6.13	1.75	1.27	.208	22	5.41	0.80	0.71	.482
	Cleanliness	36	5.58	2.14			18	5.17	1.34		
Overall	Neutral	102	6.48	1.13	0.22	.826	68	5.65	0.59	-0.03	.974
	Cleanliness	105	6.45	1.10			58	5.66	0.68		

Note. Response scales in Study 1 ranged from 0 (perfectly OK) to 9 (extremely wrong); participants who responded with "9" were removed from analyses. Response scales in Study 2 ranged from 1 (nothing wrong at all) to 7 (extremely wrong); participants who responded with "7" were removed from analyses.

The study reported in this article earned the *Open Data* badge: https://osf.io/jxad3/

Figure 1. The study reported in this article earned the *Open Data* badge: https://osf.io/jxad3/