Clean Data: Statistical Artefacts Wash Out Replication Efforts

Simone Schnall
University of Cambridge

Social Psychology (in Press)

Abstract

Johnson, Cheung and Donnellan (2014a) reported a failure to replicate Schnall, Benton and Harvey (2008)'s effect of cleanliness on moral judgment. However, inspection of the replication data shows that participants provided high numbers of severe moral judgments – a ceiling effect. In the original data percentage of extreme responses per moral dilemma correlated negatively with the effect of the manipulation. In contrast, this correlation was absent in the replications, due to almost all items showing a high percentage of extreme responses. Therefore the parametric statistics reported by Johnson et al. (2014a) are inconclusive regarding the reproducibility of the original effect. Direct replications are prone to error when reviewers only judge similarity of methods, but not resulting data and conclusions. It is my conclusion that preventable problems can arise if publication decisions are made without independent post-data peer evaluation.

Keywords: Cleanliness; Moral Judgment; Registered Replication; Ceiling Effect; Peer Review

Clean Data: Statistical Artefacts Wash Out Replication Efforts

Schnall et al. (2008) demonstrated that primed cleanliness decreases the severity of moral judgments. For each of the 12 moral dilemmas across two experiments the mean for the clean condition was lower than the mean for the neutral condition. Aggregating across dilemmas resulted in effect sizes of Cohen's *d* of .61 (Experiment 1), and .85 (Experiment 2). Two independent direct replications of Experiment 1 (Arbesfeld, Collins, Baldwin, & Daubman, 2014; Besman, Dubensky, Dunsmore, & Daubman, 2013) produced somewhat smaller effects, *d*s = .47 and .48.[1].

Johnson et al. (2014a) carried out registered replications using materials and procedures approved by the first author of the original work and reported non-replication of the effect. To understand the discrepancy between the results from Schnall et al. (2008), Besman et al. (2013) and Arbesfeld et al. (2014) on the one hand and from Johnson et al. (2014a) on the other, the present article provides a comparison of original and replication data.[2] Additional successful replications have been produced recently (Genschow, Loissel & Schnall, 2013).

Inspection of the neutral condition of Experiment 1 across original and replication (Johnson et al., 2014a, Table 1) reveals that item means are generally higher in the replication. Indeed, even at baseline participants gave significantly more severe ratings in the replication study (M = 6.48, SD = 1.14) than the original study (M = 5.81, SD = 1.47), $F(1, 120) = 5.32$, $p = .02$. To further test whether moral responses were more severe even without any manipulation, percentages of extreme responses were compared. Relative to all other responses, the percentage of extreme responses ('9' on a scale from '0' = 'perfectly OK' to '9' = 'extremely wrong') in the neutral condition was significantly greater in Replication Study 1 (37.91%) than in Original Study 1 (28.33%), $\chi^2 = 3.98$, $p = .05$. In the replication by Arbesfeld et al. (2014), the percentage of extreme response in the neutral condition was also 28.33%.Similarly, the percentage of extreme responses ('7' on a scale from '1' = 'nothing wrong at all' to '7' = 'extremely wrong') in the neutral condition was greater in Replication Study 2 (44.20%) than in Original Study 2 (28.03%), $\chi^2 = 10.88$, $p = .001$. This suggests a ceiling effect: Participants may have given higher responses had the scale allowed them to do so.

Because a ceiling effect on a dependent variable can wash out potential effects of an independent variable (Hessling, Traxel & Schmidt, 2004), the relationship between the percentage of extreme responses and the effect of the cleanliness manipulation was examined. First, using all 24 item means from original and replication studies, the effect of the manipulation on each item was quantified. Given the high percentage of extreme responses in the replication data and the resulting severe skew in distributions, effect size measures that assume a parametric distribution (e.g., Cohen's *d*, which uses the standard deviation of the mean in the denominator) cannot be used to effectively compare both original and replication data. Because it makes no assumption about underlying distributions, relative mean difference between neutral and clean condition was used as an effect size measure. For each dilemma the mean of the clean condition was subtracted from the

mean of the neutral condition, and the resulting value was divided by the sum of the two condition means. This provides a normalized measure of effect size per dilemma. Second, for each dilemma the percentage of extreme responses averaged across neutral and clean conditions was computed. This takes into account the extremity of both conditions, and therefore provides an unbiased indicator of ceiling per dilemma. The ceiling indicator was almost twice as high for replication items (M = 41.30, SD = 20.41) as for original items (M = 23.41, SD = 18.21), $F(1, 22) = 5.13$, $p = .03$.[3]

Ceiling for each dilemma was then plotted relative to the effect of the cleanliness manipulation (Figure 1). Across the 24 dilemmas from all 4 experiments, dilemmas with a greater percentage of extreme responses were associated with lower effect sizes (r = -.50, p = .01, two-tailed). This negative correlation was entirely driven by the 12 original items, indicating that the closer responses were to ceiling, the smaller was the effect of the manipulation (r = -.49, p = .10).[4] In contrast, across the 12 replication items there was no correlation (r = .11, p = .74). For 10 out of 12 replication items the modal response was the top value of the rating scale, namely "9" (Experiment 1), or "7" (Experiment 2).

The parametric tests reported by Johnson et al. (2014a) assume a normal distribution of raw scores. Given the excessive number of extreme values and therefore skewed distribution, tests based on means and standard deviations underestimate potential condition differences (Hessling et al., 2004). Although some effects of skew could be ameliorated by transforming the data, even after transformation a null effect is inconclusive: Scores are compressed toward the top end of the scale and therefore show limited determinate variance near ceiling. Because a significance test compares variance due to a manipulation to variance due to error, an observed lack of effect can result merely from a lack in variance that would normally be associated with a manipulation. Given the observed ceiling effect, a statistical artefact, the analyses reported by Johnson et al. (2014a) are invalid and allow no conclusions about the reproducibility of the original findings.

A Cautionary Tale about Replication Efforts in the Absence of Peer-Review?

Direct replications apply methods used in one context in precisely the same manner in a different context. Because of inherent social, cultural and historical differences across testing conditions and subject populations, this can result in inappropriate tests of underlying theoretical constructs (Stroebe & Strack, 2014). The pertinent literature suggests that people draw on a variety of sources when making moral judgments (e.g., Cannon, Schnall, & White, 2011). In particular, politically conservative participants use different moral foundations than liberal participants (Inbar, Pizarro, Iyer, & Haidt, 2012). Participants in the Mid-West of the United States may be more conservative than participants in the United Kingdom, which could result in harsher moral judgments. Given such population differences, stimuli from earlier research have to be used with caution, and data have to be examined to establish acceptable validity and reliability.

As outlined in their editorial, Nosek and Lakens (2014) championed an innovative model of scientific publishing. This model should be commended for the rigorous criteria for pre-registration of methods and open access to data. However, an inherent weakness is that it involved no reviewer input on the final report. Indeed, so far no other journal has accepted manuscripts for publication using a registered replication format that omits independent post-data peer-review. Independent peer evaluation has been the gold standard for assessing research quality because experts are familiar with methods and data and can put specific findings into the context of the broader literature. A reviewer likely would have noticed the higher replication item means for the neutral condition in Table 1 (Johnson et al., 2014a) and requested further information regarding baseline moral judgments. Thus, in the absence of quality control by post-data peer review, it is difficult to assess the validity of replication findings, whether successful or not. It therefore risks throwing out commendable replication efforts with the bath water.

Author's Note

The author declares no conflict-of-interest with the content of this article. I am grateful to Editor-in-Chief Christian Unkelbach for granting me a published response to the replication of my work. I thank Mark Haggard and Mayank R. Mehta for statistical advice, and Norbert Schwarz, Fritz Strack, Jerry Clore, Jon Haidt, Thomas Schubert, Oliver Genschow, Suzanne Brink, Tomas Folke, and Gabriela Pavarini for feedback.

References

Arbesfeld, J. Collins, T., Baldwin, D., & Daubman, K. (2014, February 15). Clean thoughts lead to less severe moral judgment. Retrieved 16:26, February 18, 2014 from http://www.PsychFileDrawer.org/replication.php?attempt=MTc3

Besman, M., Dubensky, C., Dunsmore, L., & Daubman, K (2013, February 23). Cleanliness primes less severe moral judgments. Retrieved 16:52, January 29, 2014 from http://www.PsychFileDrawer.org/replication.php?attempt=MTQ5

Cannon, P. R., Schnall, S., & White, M. (2011). Transgressions and expressions: Affective facial muscle activity predicts moral judgments. *Social Psychological and Personality Science, 2,* 325-331.

Ferguson, M. J., Carter, T. J., & Hassin, R. R. (2014). Commentary on the attempt to replicate the effect of the American flag on increased republican attitudes. *Social Psychology.*

Genschow, O., Loissel, E., & Schnall, S. (2013). [Replications of differential effects of cleanliness on moral judgment]. Unpublished raw data.

Inbar, Y., Pizarro, D., Iyer, R., & Haidt, J. (2012). Disgust sensitivity, political conservatism, and voting. *Social Psychological and Personality Science*, 3, 537-544.

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

Johnson, D. J., Cheung, F., & Donnellan, M. B. (2014b, January 01). Cleanliness primes do not influence moral judgment. Retrieved 08:10, March 12, 2014 from http://www.PsychFileDrawer.org/replication.php?attempt=MTcy

Hessling, R. M., Traxel, N. M., & Schmidt, T. J. (2004). Ceiling effect. In M. S. Lewis-Beck, A. Bryman, & T. F. Liao (Eds.), *SAGE encyclopedia of social science research methods (p. 107).* Thousand Oaks, CA: Sage.

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

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

Srull, T. K., & Wyer, R. S. (1979).The role of category accessibility in the interpretation of information about persons: Some determinants. *Journal of Personality and Social Psychology, 37,* 1660-1672.

Stroebe, W., & Strack, F. (2014). The alleged crisis and the illusion of exact replication. *Perspectives on Psychological Science, 9,* 59-71.

Footnotes

[1] A further online study (Johnson, Cheung & Donnellan, 2014b) was not a direct replication because the manipulation lacked the experimental control of the other studies. The scrambled sentences task (e.g., Srull & Wyer, 1979) involves underlining words on a piece of paper, as in Schnall et al. (2008), Besman et al. (2013) and Arbesfeld et al. (2014). Whereas the paper-based task is completed under the guidance of an experimenter, for online studies it cannot be established whether participants exclusively focus on the priming task. Indeed, results from online versions of priming studies systematically differ from lab-based versions (Ferguson, Carter, & Hassin, 2014). Further, the study aimed to induce cleanliness but it is unknown how clean the participants' surroundings were while completing the study.

[2] SPSS data files are available on the Open Science Framework: osf.io/4j8db. All data exclusions are described in Schnall et al. (2008). No other dependent variables or manipulations beyond those reported were included.

[3] The percentage of extreme responses for Study 1 was 22.08% for Schnall et al. (2008), 26.39% for Arbesfeld et al. (2014) and 38.53% for Johnson et al. (2014).

[4] "Kitten" in Original Study 1 showed a large effect of the manipulation despite high percentage of extreme scores. Without this somewhat unusual item the correlation between effect size and extremity is $r = -.61$, $p = .02$. However, inferences about specific items are inconclusive compared to analyses aggregating across studies that used comparable methods.

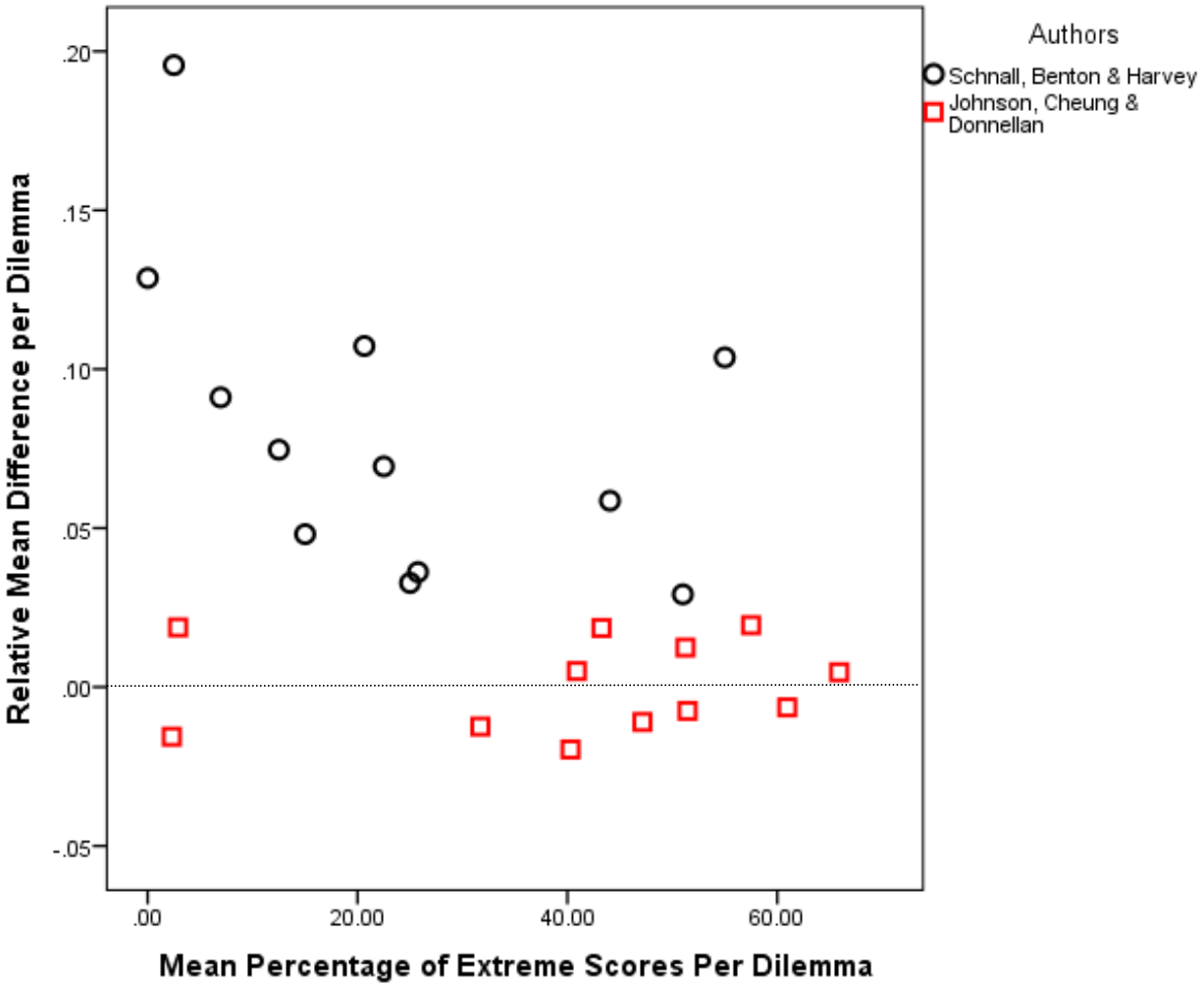

Figure 1. Scatter plot of extreme responses relative to effect size across the 24 moral dilemmas in original vs. replication experiments. For original items effect size was negatively correlated with percentage of extreme scores. For replication items most items had very high percentage of extreme responses.