Commentaries

Commentary and Rejoinder on Johnson, Cheung, and Donnellan (2014a)

Clean Data: Statistical Artifacts Wash Out Replication Efforts

Simone Schnall

University of Cambridge, UK

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

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

[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).

[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. Further, the study aimed to induce cleanliness but it is unknown how clean the participants' surroundings were while completing the study.

DOI: 10.1027/1864-9335/a000204

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 four 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

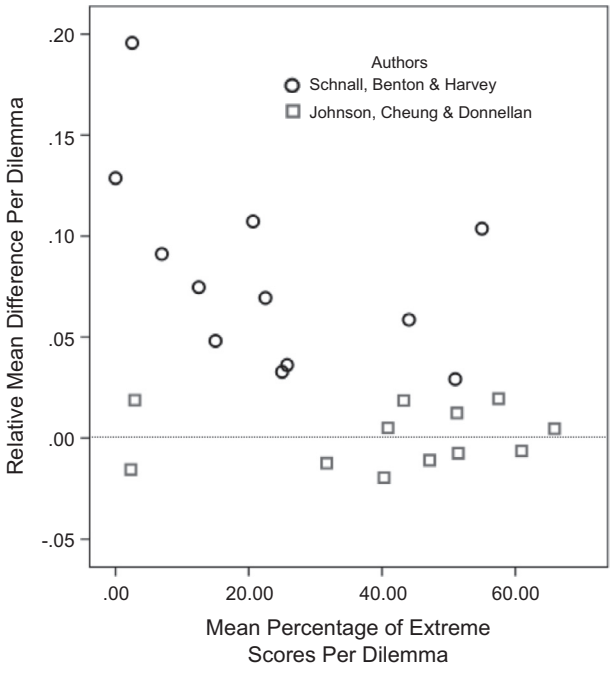

Figure 1. Scatter plot of extreme responses relative to effect size across the 24 moral dilemmas in original vs. replication experiments. For original items effect size was negatively correlated with percentage of extreme scores. For replication items most items had very high percentage of extreme responses.

($r = -.49$, $p = .10$).[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 artifact, the analyses reported by Johnson et al. (2014a) are invalid and allow no conclusions about the reproducibility of the original findings.

[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. (2014a)

[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.

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 preregistration 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.

Acknowledgments

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 from http://www.PsychFileDrawer.org/chart.php?target_article=48&type=success

Besman, M., Dubensky, C., Dunsmore, L., & Daubman, K. (2013, February 23). *Cleanliness primes less severe moral judgments*. Retrieved 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*. Advance online publication. doi 10.1027/1864-9335/a000202

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, 45*, 209–215. doi: 10.1027/1864-9335/a000186

Johnson, D. J., Cheung, F., & Donnellan, M. B. (2014b, January 01). Cleanliness primes do not influence moral judgment. Retrieved 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*. (pp. 107). Thousand Oaks, CA: Sage.

Nosek, B. A., & Lakens, D. (2014). Registered reports: A method to increase the credibility of published results. *Social Psychology, 45*, 137–141.

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.

Published online May 30, 2014

Simone Schnall

University of Cambridge
Department of Psychology
Downing Street
Cambridge, CB2 3EB
United Kingdom
E-mail ss877@cam.ac.uk

Hunting for Artifacts

The Perils of Dismissing Inconsistent Replication Results

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

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

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 preregistered replication studies for advancing research in psychological science.

Keywords: cleanliness, moral judgment, preregistration, replication

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 preregistered 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 [0.08, 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.

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

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	5.65	0.68		

Notes. 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.

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 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 scenario 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, $b = -.04$, $t(731) = -0.52$, $p = .61$. 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 & Mogilner, 2014; Kay, Laurin, Fitzsimons, & 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.

Preregistration and Peer-Review

Schnall (2014) expressed reservations that our preregistered replication did not have sufficient post-data collection peer

[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, Cheung, and Donnellam (2014). We obtained a much larger sample size to test for moderator effects and to detect population effect sizes smaller than the published research.

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 preregistered replication results simply because the results were inconsistent with the original findings.

Acknowledgments

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

References

Gino, F., & Mogilner, C. (2014). Time, money, and morality. *Psychological Science, 25*, 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, 45*, 209–215. doi: 10.1027/1864-9335/a000186

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, 45*, 137–141. doi: 10.1027/1864-9335/a000192

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*. 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.

Published online May 30, 2014

David J. Johnson

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

