Volume 45 / Number 3 / 2014
ISSN-L 1864-9335, ISSN-Print 1864-9335, ISSN-Online 2151-2590

Social Psychology

3/14

www.hogrefe.com/journals/sp

Special Issue
Replications of Important
Results in Social Psychology

Guest Editors
Brian A. Nosek
Daniël Lakens

HOGREFE

Social Psychology

Publisher Hogrefe Publishing, Merkelstr. 3, 37085 Göttingen, Germany,
Tel. +49 551 99950-0, Fax +49 551 99950-111, E-mail publishing@hogrefe.com
North America: Hogrefe Publishing, 38 Chauncy Street, Suite 1002, Boston, MA 02111, USA,
Tel. (866) 823-4726, Fax (617) 354-6875, E-mail publishing@hogrefe.com

Production Regina Pinks-Freybott, Hogrefe Publishing, Merkelstr. 3, 37085 Göttingen, Germany,
Tel. +49 551 99950-0, Fax +49 551 99950-111, E-mail journalsproduction@hogrefe.com

Subscriptions Hogrefe Publishing, Herbert-Quandt-Str. 4, D-37081 Göttingen, Germany,
Tel. +49 551 99950-900, Fax +49 551 90050-998

Advertising/Inserts Melanie Beck, Hogrefe Publishing, Merkelstr. 3, D-37085 Göttingen, Germany,
Tel. +49 551 99950-423, Fax +49 551 99950-111, E-mail marketing@hogrefe.com

ISSN ISSN-L 1864-9335, ISSN-Print 1864-9335, ISSN-Online 2151-2590

Publication Published in 6 issues per annual volume. *Social Psychology* is the continuation of *Zeitschrift für Sozialpsychologie* (ISSN 0044-3514), the last annual volume of which (Volume 38) was published in 2007.

Subscription Prices Calendar year subscriptions only. Rates for 2014: Institutions €314.00/US $424.00/£ 254.00; Individuals €157.00/US $212.00/£ 127.00 (all plus US $24.00/€18.00/£ 15.00 shipping & handling; Germany: €5.00).

Payment Payment may be made by check, international money order, or credit card, to Hogrefe Publishing, Merkelstr. 3, D-37085 Göttingen, Germany. US and Canadian subscriptions can also be ordered from Hogrefe Publishing, 38 Chauncy Street, Suite 1002, Boston, MA 02111, USA.

Electronic Full Text The full text of *Social Psychology* is available online at www.psyjournals.com and in PsycARTICLES.

Abstracting Services Abstracted/indexed in Current Contents/Social and Behavioral Sciences (CC/S&BS), Social Sciences Citation Index (SSCI), PsycINFO, PASCAL, PSYNDEX, ERIH, Scopus, and EM Care.
Impact Factor (2012): 1.892

Social Psychology

Volume 45, No. 3, 2014

HOGREFE

Contents

Contents

Registered Reports

A Method to Increase the Credibility of Published Results

Brian A. Nosek[1] and Daniël Lakens[2]

[1]University of Virginia and Center for Open Science, Charlottesville, VA, USA,
[2]Eindhoven University of Technology, The Netherlands

Ignoring Replications and Negative Results Is Bad for Science

The published journal article is the primary means of communicating scientific ideas, methods, and empirical data. Not all ideas and data get published. In the present scientific culture, novel and positive results are considered more publishable than replications and negative results. This creates incentives to avoid or ignore replications and negative results, even at the expense of accuracy (Giner-Sorolla, 2012; Nosek, Spies, & Motyl, 2012). As a consequence, replications (Makel, Plucker, & Hegarty, 2012) and negative results (Fanelli, 2010; Sterling, 1959) are rare in the published literature. This insight is not new, but the culture is resistant to change. This article introduces the first known journal issue in any discipline consisting exclusively of preregistered replication studies. It demonstrates that replications have substantial value, and that incentives can be changed.

There are a number of advantages of performing direct replications, and publishing the results irrespective of the outcome. First, direct replications add data to increase precision of the effect size estimate via meta-analysis. Under some circumstances, this can lead to the identification of false positive research findings. Without direct replication, there is no way to confidently identify false positives. Conceptual replications have been more popular than direct replications because they abstract a phenomenon from its original operationalization and contribute to our theoretical understanding of an effect. However, conceptual replication are not best suited to clarify the truth of any particular effect because nonsignificant findings are attributable to changes in the research design, and rarely lead researchers to question the phenomenon (LeBel & Peters, 2011; Nosek, Spies, & Motyl, 2012).

Second, direct replication can establish generalizability of effects. There is no such thing as an exact replication. Any replication will differ in innumerable ways from the original. A direct replication is the attempt to duplicate the conditions and procedure that existing theory and evidence anticipate as necessary for obtaining the effect (Open Science Collaboration, 2012, 2013; Schmidt, 2009). Successful replication bolsters evidence that all of the sample, setting, and procedural differences presumed to be irrelevant are, in fact, irrelevant.

Third, direct replications that produce negative results facilitate the identification of boundary conditions for real effects. If existing theory anticipates the same result should occur and, with a high-powered test, it does not, then something in the presumed irrelevant differences between original and replication could be the basis for identifying constraints on the effect. In other words, understanding any effect requires knowing when it does and does not occur. Therefore, replications and negative results are consequential for theory development.

Registered Reports Are a Partial Solution

Despite their theoretical and empirical value, the existing scientific culture provides few incentives for researchers to conduct replications or report negative results (Greenwald, 1975; Koole & Lakens, 2012). Editors and reviewers of psychology journals often recommend against the publication of replications (Neuliep & Crandall, 1990, 1993). If journals will not publish replications, why would researchers bother doing them?

This special issue of *Social Psychology* presents 15 articles with replications of important results in social psychology. Moreover, these articles demonstrate a novel publishing format – Registered Reports. By reviewing and accepting preregistered proposals prior to data collection, Registered Reports are an efficient way to change incentive

Social Psychology 2014; Vol. 45(3):137–141
DOI: 10.1027/1864-9335/a000192

structures for conducting replications and reporting results irrespective of their statistical significance.

In 2013, the guest editors issued calls for submissions of proposals to replicate published studies in social psychology (Nosek & Lakens, 2013). Prospective authors proposed a study or studies for replication and articulated (1) why the result is important to replicate, and (2) the design and analysis plan for a high-powered replication effort. Proposals that passed initial editorial review went out for peer review. Reviewers evaluated the importance of conducting a replication and the quality of the methodology. At least one author of the original article was invited to be a reviewer if any were still alive. Most invited original authors provided a review. Authors incorporated feedback from peer review in their designs and, if the proposal had not been accepted initially, resubmitted for review and acceptance (or rejection) based on reviewer feedback.

We received 36 pre-proposals of which 24 were encouraged to submit full proposals. Ultimately 14 proposals were accepted. A 15th article (Moon & Roeder, 2014) was solicited as a second replication of one of the peer reviewed, accepted proposals (Gibson, Losee, & Vitiello, 2014) because reviewers suggests that the effect may not occur among Asian women at southern US universities (Gibson et al.'s sample).

Accepted proposals were registered at the Open Science Framework (OSF; http://osf.io/) prior to data collection along with the study materials. Authors proceeded with the data collection with assurance that the results would be published irrespective of the outcome, as long they followed the registered plans or provided reasonable, explicit justifications for deviating from the plan. The infrequent deviations were assessed by the action editor as whether they sacrificed integrity of the confirmatory plan before acceptance. For example, in two cases, the sample size was far short of the registered plan. The editors required additional data collection prior to acceptance. In the published articles, authors report results according to the registered confirmatory analysis plan, disclose any deviation from the plan, and sometimes provide additional exploratory analyses – clearly designated as such.

Successful proposals were designs that peer reviewers considered to be high-powered, high-quality, faithful replication designs. Peer review prior to data collection lowered the barrier to conduct replications because authors received editorial feedback about publication likelihood before much of the work was done. Furthermore, authors could focus on reporting a confirmatory analysis (Wagenmakers, Wetzels, Borsboom, van der Maas, & Kievit, 2012), without the need to hunt for positive and clean results (Simmons, Nelson, & Simonsohn, 2011).

Registered reports also shift the incentives for reviewers. When the results are known, evaluation of quality is likely influenced by preexisting beliefs (Bastardi, Uhlmann, & Ross, 2011). Motivated reasoning makes it easy to generate stories for why results differed from expectations. Following Kerr's (1998) observation of hypothesizing about one's own research outcomes post facto, this might be termed, CARKing, critiquing after the results are known.

When reviewing a study proposal, only the design is available as a basis for critique. Reviewers' motivation is to make sure that the design provides a fair test. Reviewers could insist that there are innumerable conditions and moderating influences that must be met. However, each of these constrains the scope of the original effect and risks trivializing the result. So, reviewers may have competing interests – just as they do in theorizing – providing just enough constraint to ensure a fair test, but not so much to make the effect uninteresting or inapplicable.

In sum, review prior to data collection focused researchers and reviewers to evaluate the methodological quality of the research, rather than the results.

What Registered Reports Do Not Do

Preregistration and peer review in advance of data collection or analysis do not lead to definitive results. Even highly powered designs – like those in this issue – leave room for Type 1 and Type 2 errors. Furthermore, when registered reports are used for replication studies, different results between original and replication research could mean that there are unknown moderators or boundary conditions that differentiate the two studies. As such, the replication can raise more questions than it answers. At the same time, effects size estimates in small samples – common in original research – can vary considerably and are more likely to elicit an exaggerated effect size than results from larger sample sizes (Schönbrodt & Perugini, 2013). Therefore, *not finding* a predicted effect in a *large* study may indicate more about the likelihood that an effect is true than *finding* a predicted effect in a *small study*, because the former is statistically less likely if an effect is true (Button et al., 2013; Lakens & Evers, in press).

Registered Reports do not prevent or discourage exploratory analysis. Rather, they make clear the distinction between confirmatory and exploratory analysis. This applies to registered reports whether they are conducted for replications or original research. Confirmatory results follow a preregistered analysis plan and thereby ensure interpretability of the reported *p*-values (Wagenmakers et al., 2012). In exploratory analysis, *p*-values lose their meaning due to an unknown inflation of the alpha-level. That does not mean that exploratory analysis is not valuable; it is just more tentative.

Open Science Practices in This Special Issue

The articles published in this special issue adopted transparency practices that further enhance the credibility of the published results. These practices make explicit how the research was conducted and make all the relevant materials and data available to facilitate reanalysis, reuse, and replication. The practices include:

- For all articles, original proposals, anonymized data, and study materials are registered and available at the Open Science Framework (OSF; http://osf.io/). Each article earned badges acknowledging preregistration, open data, and open materials (Miguel et al., 2014) that are maintained by the Open Science Collaboration (https://osf.io/tvyxz/). Badges and links to the OSF projects appear in the acknowledgments section of each article.
- Some OSF projects have additional material such as photos or video simulations of the procedures.
- All articles specify the contributions of each author.
- All articles specify funding sources.
- All articles disclosed whether authors had conflicts of interest (Greenwald, 2009).
- All articles make explicit all conditions, measures, data exclusions, and how samples sizes were determined (LeBel et al., 2013; Simmons, Nelson, & Simonsohn, 2012). This disclosure standard has been introduced at *Psychological Science* starting in January 2014 as an expectation for all reviewed submissions (Eich, 2013).

This Special Issue

The articles in this special issue demonstrate a variety of ways in which published findings can be important enough to replicate. For one, every discipline has a number of classic, textbook studies that exemplify a research area. These studies are worth revisiting, both to assure their robustness and sometimes to analyze the data with modern statistical techniques. This special issue contains several replications of textbook studies, sometimes with surprising results (Nauts, Langner, Huijsmans, Vonk, & Wigboldus, 2014; Sinclair, Hood, & Wright, 2014; Vermeulen, Batenburg, Beukeboom, & Smits, 2014; Wesselmann et al., 2014).

Second, several teams (Brandt, IJzerman, & Blanken, 2014; Calin-Jageman & Caldwell, 2014; Johnson, Cheung, & Donnellan, 2014; Lynott et al., 2014) replicated recent work that has received substantial attention and citation. Given their high impact on contemporary research trajectories, it is important to investigate these effects and the conditions necessary to elicit them to ensure efficient development of theory, evidence, and implications.

Third, replication studies might provide a way to validate results when previous research lines have reached opposite conclusions (Žeželj & Joki, 2014), or provide more certainty about the presence and mechanisms of the original effect by performing direct replications while simultaneously testing theorized moderators (Gibson, Losee, & Foxwell, 2014; Moon & Roeder, 2014; Müller & Rothermund, 2014).

Fourth, replications can reveal boundary conditions, for example by showing how sex differences in distress from infidelity is reliably observed in a young sample, but not in an older sample (IJzerman et al., 2014). Performing direct replications can be especially insightful when a previous meta-analysis suggests the effect is much smaller than suggested by the published findings (Blanken, Van de Ven, Zeelenberg, & Meijers, 2014).

Finally, Many Labs replication project (Klein et al., 2014) was a large international collaboration that amassed 36 samples and 6,344 participants to assess variation in replicability across samples and settings of 13 effects. It revealed relatively little variation in effect sizes across samples and settings, and demonstrated that crowdsourcing offers a feasible way to collect very large sample sizes and gain substantial knowledge about replicability.

No single replication provides the definitive word for or against the reality of an effect, just as no original study provides definitive evidence for it. Original and replication research each provides a piece of accumulating evidence for understanding an effect and the conditions necessary to obtain it. Following this special issue, *Social Psychology* will publish some commentaries and responses by original and replication authors of their reflections on the inferences from the accumulated data, and questions that could be addressed in follow-up research.

Closing

Registered Reports are new model for publishing that incorporates preregistration of designs and peer review before data collection. The approach nudges incentives for research accuracy to be more aligned with research success. As a result, the model may increase the credibility of the published results. Some pioneering journals in psychology and neuroscience have adopted Registered Reports offering substantial opportunity to evaluate and improve this publishing format (e.g., Chambers, 2013; Simons & Holcombe, 2013; Wolfe, 2013). Further, through the OSF (http://osf.io/), the Center for Open Science (http://cos.io/) provides free services to researchers and journals to facilitate Registered Reports and other transparency practices including badges, disclosure standards, and private or public archiving of research materials and data.

This special issue shows that the incentive structures to perform and publish replication studies and negative results can change. However, it is just a demonstration. Many cultural barriers remain. For example, when judging the importance of replication proposals, some reviewers judged a replication as unimportant because as expert "insiders" they already knew that the original result was not robust, even though this knowledge is not shared in the scientific literature. The irreproducibility of certain effects may be informally communicated among particular insiders, but never become common knowledge. Knowledge accumulation will be much more efficient if insider knowledge is accessible and discoverable. Registered Reports are just one step for addressing that challenge.

Two central values of science are openness and reproducibility. In principle, the evidence supporting scientific knowledge can be reproduced by following the original methodologies. This differentiates science from other ways of knowing – confidence in claims is not based on trusting the source, but in evaluating the evidence itself.

Acknowledgments

This article was supported by the Center for Open Science. B. N. and D. L. both conceived and wrote the paper. The authors declare no conflict of interest in the preparation of this article.

References

Bastardi, A., Uhlmann, E. L., & Ross, L. (2011). Wishful thinking, belief, desire, and the motivated evaluation of scientific evidence. *Psychological Science, 22*, 731–732.

Blanken, I., Van de Ven, N., Zeelenberg, M., & Meijers, M. H. C. (2014). Three attempts to replicate the moral licensing effect. *Social Psychology, 45*, 232–238. doi: 10.1027/1864-9335/a000189

Brandt, M. J., IJzerman, H., & Blanken, I. (2014). Does recalling moral behavior change the perception of brightness? A replication and meta-analysis of Banerjee, Chatterjee, and Sinha (2012). *Social Psychology, 45*, 246–252. doi: 10.1027/1864-9335/a000191

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafo, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience, 14*, 1–12. doi: 10.1038/nrn3475

Calin-Jageman, R. J., & Caldwell, T. L. (2014). Replication of the superstition, performance study by Damisch, Stoberock, Mussweiler (2010). *Social Psychology, 45*, 239–245. doi: 10.1027/1864-9335/a000190

Chambers, C. D. (2013). Registered reports: A new publishing initiative at Cortex. *Cortex, 49*, 609–610.

Eich, E. (2013). Business not as usual. *Psychological Science, 25*, 3–6. doi: 10.1177/0956797613512465

Fanelli, D. (2010). ''Positive'' results increase down the hierarchy of the sciences. *PLoS One, 5*, e10068.

Gibson, C. E., Losee, J., & Foxwell, C. (2014). A replication of stereotype susceptibility (Shih, Pittinsky, & Ambady, 1999): Identity salience and shifts in quantitative performance. *Social Psychology, 45*, 194–198. doi: 10.1027/1864-9335/a000184

Gibson, C. E., Losee, J., & Vitiello, C. (2014). A replication attempt of stereotype susceptibility: Identity salience and shifts in quantitative performance. *Social Psychology, 45*, 1XX–1XX. doi: 10.1027/1864-9335/a000184

Giner-Sorolla, R. (2012). Science or art? How aesthetic standards grease the way through the publication bottleneck but undermine science. *Perspectives on Psychological Science, 7*, 562–571.

Greenwald, A. G. (1975). Consequences of prejudice against the null hypothesis. *Psychological Bulletin, 82*, 1–20. doi: 10.1037/h0076157

Greenwald, A. G. (2009). What (and where) is the ethical code concerning researcher conflict of interest? *Perspectives on Psychological Science, 4*, 32–35.

IJzerman, H., Blanken, I., Brandt, M. J., Oerlemans, J. M., Van den Hoogenhof, M. M. W., Franken, S. J. M., & Oerlemans, M. W. G. (2014). Sex differences in distress from infidelity in early adulthood and in later life: A replication and meta-analysis of Shackelford et al. (2004). *Social Psychology, 45*, 202–208. doi: 10.1027/1864-9335/a000185

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

Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review, 2*, 196–217.

Klein, R. A., Ratliff, K., Vianello, M., Adams, A. B. Jr., Bahník, S., Bernstein, N. B., ... Nosek, B. A. (2014). Investigating variation in replicability: A ''Many Labs'' Replication Project. *Social Psychology, 45*, 142–152. doi: 10.1027/1864-9335/a000178

Koole, S. L., & Lakens, D. (2012). Rewarding replications: A sure and simple way to improve psychological science. *Perspectives on Psychological Science, 7*, 608–614. doi: 10.1177/1745691612462586

Lakens, D., & Evers, E. (in press). Sailing from the seas of chaos into the corridor of stability: Practical recommendations to increase the informational value of studies. *Perspectives on Psychological Science*.

LeBel, E., Borsboom, D., Giner-Sorolla, R., Hasselman, F., Peters, K., Ratliff, K., & Smith, C. (2013). Psychdisclosure.Org: Grassroots support for reforming reporting standards in psychology. *Perspectives on Psychological Science, 8*, 424–432.

LeBel, E. P., & Peters, K. R. (2011). Fearing the future of empirical psychology: Bem's (2011) evidence of psi as a case study of deficiencies in modal research practice. *Review of General Psychology, 15*, 371–379. doi: 10.1037/a0025172

Lynott, D., Corker, K. S., Wortman, J., Connell, L., Donnellan, M. J., Lucas, R. E., & O'Brien, K. (2014). Replication of ''Experiencing physical warmth promotes interpersonal warmth'' by Williams and Bargh (2008). *Social Psychology, 45*, 216–222. doi: 10.1027/1864-9335/a000187

Makel, M., Plucker, J., & Hegarty, B. (2012). Replications in psychology research: How often do they really occur? *Perspectives in Psychological Science, 7*, 537–542. doi: 10.1177/1745691612460688

Miguel, E., Camerer, C., Casey, K., Cohen, J., Esterling, K. M., Gerber, A., ... Van der Laan, M. (2014). Promoting transparency in social science research. *Science, 343*, 30–31. doi: 10.1126/science.1245317

Moon, A., & Roeder, S. S. (2014). A secondary replication attempt of stereotype susceptibility (Shih, Pittinsky, & Ambady, 1999). *Social Psychology, 45*, 199–201. doi: 10.1027/1864-9335/a000193

Müller, F., & Rothermund, K. (2014). Replication of stereotype activation (Banaji & Hardin, 1996; Blair & Banaji, 1996). *Social Psychology, 45*, 187–193. doi: 10.1027/1864-9335/a000183

Nauts, S., Langner, O., Huijsmans, I., Vonk, R., & Wigboldus, D. H. J. (2014). Forming impressions of personality: A replication and review of Asch's (1946) evidence for a primacy-of-warmth effect in impression formation. *Social Psychology, 45*, 153–163. doi: 10.1027/1864-9335/a000179

Neuliep, J. W., & Crandall, R. (1990). Editorial bias against replication research. *Journal of Social Behavior and Personality, 5*, 85–90.

Neuliep, J. W., & Crandall, R. (1993). Reviewer bias against replication research. *Journal of Social Behavior and Personality, 8*, 21–29.

Nosek, B. A., & Lakens, D. (2013). Call for proposals: Special issue of social psychology on ''Replications of important results in social psychology''. *Social Psychology, 44*, 59–60. doi: 10.1027/1864-9335/a000143

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

Open Science Collaboration. (2012). An open, large-scale, collaborative effort to estimate the reproducibility of psychological science. *Perspectives on Psychological Science, 7*, 657–660. doi: 10.1177/1745691612462588

Open Science Collaboration. (2013). The reproducibility project: A model of large-scale collaboration for empirical research on reproducibility. In V. Stodden, F. Leisch, & R. Peng (Eds.), *Implementing reproducible computational research (a volume in the r series)*. New York, NY: Taylor & Francis.

Schmidt, S. (2009). Shall we really do it again? The powerful concept of replication is neglected in the social sciences. *Review of General Psychology, 13*, 90–100. doi: 10.1037/a0015108

Schönbrodt, F. D., & Perugini, M. (2013). At what sample size do correlations stabilize? *Journal of Research in Personality, 47*, 609–612. doi: 10.1016/j.jrp.2013.05.009

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22*, 1359–1366.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2012). *A 21 word solution.* Retrieved from http://ssrn.com/abstract=2160588

Simons, D., & Holcombe, A. (2013). *Registered replication reports.* Retrieved from http://www.psychologicalscience.org/index.php/replication

Sinclair, H. C., Hood, K., & Wright, B. (2014). Revisiting Romeo and Juliet (Driscoll, Davis, & Lipetz, 1972): Reexamining the links between social network opinions and romantic relationship outcomes. *Social Psychology, 45*, 170–178. doi: 10.1027/1864-9335/a000181

Sterling, T. D. (1959). Publication decisions and their possible effects on inferences drawn from tests of significance – Or vice versa. *Journal of the American Statistical Association, 54*, 30–34.

Vermeulen, I., Batenburg, A., Beukeboom, C., & Smits, T. (2014). Breakthrough or one-hit wonder? Replicating effects of single-exposure musical conditioning on choice behavior (Gorn, 1982). *Social Psychology, 45*, 179–186. doi: 10.1027/1864-9335/a000182

Wagenmakers, E. J., Wetzels, R., Borsboom, D., van der Maas, H. L., & Kievit, R. A. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science, 7*, 632–638.

Wesselmann, E. D., Williams, K. D., Pryor, J. B., Eichler, F. A., Gill, D., & Hogue, J. D. (2014). Revisiting Schachter's research on rejection, deviance, and communication (1951). *Social Psychology, 45*, 164–169. doi: 10.1027/1864-9335/a000180

Wolfe, J. M. (2013). Registered reports and replications in attention, perception, & psychophysics. *Attention, Perception, & Psychophysics, 75*, 781–783. doi: 10.3758/s13414-013-0502-5

Žeželj, I. L., & Jokić, B. R. (2014). Replication of experiments evaluating impact of psychological distance on moral judgment (Eyal, Liberman & Trope, 2008; Gong & Medin, 2012). *Social Psychology, 45*, 223–231. doi: 10.1027/1864-9335/a000188

Published online May 19, 2014

Brian A. Nosek

Department of Psychology
University of Virginia
102 Gilmer Hall, Box 400400
Charlottesville, VA 22904
USA
E-mail nosek@virginia.edu

Investigating Variation in Replicability

A "Many Labs" Replication Project

Richard A. Klein,[1] Kate A. Ratliff,[1] Michelangelo Vianello,[2] Reginald B. Adams Jr.,[3] Štěpán Bahník,[4] Michael J. Bernstein,[5] Konrad Bocian,[6] Mark J. Brandt,[7] Beach Brooks,[1] Claudia Chloe Brumbaugh,[8] Zeynep Cemalcilar,[9] Jesse Chandler,[10,36] Winnee Cheong,[11] William E. Davis,[12] Thierry Devos,[13] Matthew Eisner,[10] Natalia Frankowska,[6] David Furrow,[15] Elisa Maria Galliani,[2] Fred Hasselman,[16,37] Joshua A. Hicks,[12] James F. Hovermale,[17] S. Jane Hunt,[18] Jeffrey R. Huntsinger,[19] Hans IJzerman,[7] Melissa-Sue John,[20] Jennifer A. Joy-Gaba,[17] Heather Barry Kappes,[21] Lacy E. Krueger,[18] Jaime Kurtz,[22] Carmel A. Levitan,[23] Robyn K. Mallett,[19] Wendy L. Morris,[24] Anthony J. Nelson,[3] Jason A. Nier,[25] Grant Packard,[26] Ronaldo Pilati,[27] Abraham M. Rutchick,[28] Kathleen Schmidt,[29] Jeanine L. Skorinko,[20] Robert Smith,[14] Troy G. Steiner,[3] Justin Storbeck,[8] Lyn M. Van Swol,[30] Donna Thompson,[15] A. E. van 't Veer,[7] Leigh Ann Vaughn,[31] Marek Vranka,[32] Aaron L. Wichman,[33] Julie A. Woodzicka,[34] and Brian A. Nosek[29,35]

[1]University of Florida, Gainesville, FL, USA, [2]University of Padua, Italy, [3]The Pennsylvania State University, University Park, PA, USA, [4]University of Würzburg, Germany, [5]Pennsylvania State University Abington, PA, USA, [6]University of Social Sciences and Humanities Campus Sopot, Poland, [7]Tilburg University, The Netherlands, [8]City University of New York, USA, [9]Koç University, Istanbul, Turkey, [10]University of Michigan, Ann Arbor, MI, USA, [11]HELP University, Kuala Lumpur, Malaysia, [12]Texas A&M University, College Station, TX, USA, [13]San Diego State University, CA, USA, [14]Ohio State University, Columbus, OH, USA, [15]Mount Saint Vincent University, Nova Scotia, Canada, [16]Radboud University Nijmegen, The Netherlands, [17]Virginia Commonwealth University, Richmond, VA, USA, [18]Texas A&M University-Commerce, TX, USA, [19]Loyola University Chicago, IL, USA, [20]Worcester Polytechnic Institute, MA, USA, [21]London School of Economics and Political Science, London, UK, [22]James Madison University, Harrisonburg, VA, USA, [23]Occidental College, Los Angeles, CA, USA, [24]McDaniel College, Westminster, MD, USA, [25]Connecticut College, New London, CT, USA, [26]Wilfrid Laurier University, Waterloo, ON, Canada, [27]University of Brasilia, DF, Brazil, [28]California State University, Northridge, CA, USA, [29]University of Virginia, Charlottesville, VA, USA, [30]University of Wisconsin-Madison, WI, USA, [31]Ithaca College, NY, USA, [32]Charles University, Prague, Czech Republic, [33]Western Kentucky University, Bowling Green, KY, USA, [34]Washington and Lee University, Lexington, VA, USA, [35]Center for Open Science, Charlottesville, VA, USA, [36]PRIME Research, Ann Arbor, MI, USA, [37]University Nijmegen, The Netherlands

Abstract. Although replication is a central tenet of science, direct replications are rare in psychology. This research tested variation in the replicability of 13 classic and contemporary effects across 36 independent samples totaling 6,344 participants. In the aggregate, 10 effects replicated consistently. One effect – imagined contact reducing prejudice – showed weak support for replicability. And two effects – flag priming influencing conservatism and currency priming influencing system justification – did not replicate. We compared whether the conditions such as lab versus online or US versus international sample predicted effect magnitudes. By and large they did not. The results of this small sample of effects suggest that replicability is more dependent on the effect itself than on the sample and setting used to investigate the effect.

Keywords: replication, reproducibility, generalizability, cross-cultural, variation

Replication is a central tenet of science; its purpose is to confirm the accuracy of empirical findings, clarify the conditions under which an effect can be observed, and estimate the true effect size (Brandt et al., 2013; Open Science Collaboration, 2012, 2014). Successful replication of an experiment requires the recreation of the essential conditions of the initial experiment. This is often easier said than done. There may be an enormous number of variables

DOI: 10.1027/1864-9335/a000178

influencing experimental results, and yet only a few tested. In the behavioral sciences, many effects have been observed in one cultural context, but not observed in others. Likewise, individuals within the same society, or even the same individual at different times (Bodenhausen, 1990), may differ in ways that moderate any particular result.

Direct replication is infrequent, resulting in a published literature that sustains spurious findings (Ioannidis, 2005) and a lack of identification of the eliciting conditions for an effect. While there are good epistemological reasons for assuming that observed phenomena generalize across individuals and contexts in the absence of contrary evidence, the failure to directly replicate findings is problematic for theoretical and practical reasons. Failure to identify moderators and boundary conditions of an effect may result in overly broad generalizations of true effects across situations (Cesario, 2014) or across individuals (Henrich, Heine, & Norenzayan, 2010). Similarly, overgeneralization may lead observations made under laboratory observations to be inappropriately extended to ecological contexts that differ in important ways (Henry, MacLeod, Phillips, & Crawford, 2004). Practically, attempts to closely replicate research findings can reveal important differences in what is considered a direct replication (Schmidt, 2009), thus leading to refinements of the initial theory (e.g., Aronson, 1992; Greenwald, Pratkanis, Leippe, & Baumgardner, 1986). Close replication can also lead to the clarification of tacit methodological knowledge that is necessary to elicit the effect of interest (Collins, 1974).

Overview of the Present Research

Little attempt has been made to assess the variation in replicability of findings across samples and research contexts. This project examines the variation in replicability of 13 classic and contemporary psychological effects across 36 samples and settings. Some of the selected effects are known to be highly replicable; for others, replicability is unknown. Some may depend on social context or participant sample, others may not. We bundled the selected studies together into a brief, easy-to-administer experiment that was delivered to each participating sample through a single infrastructure (http://projectimplicit.net/).

There are many factors that can influence the replicability of an effect such as sample, setting, statistical power, and procedural variations. The present design standardizes procedural characteristics and ensures appropriate statistical power in order to examine the effects of sample and setting on replicability. At one extreme, sample and situational characteristics might have little effect on the tested effects – variation in effect magnitudes may not exceed expected random error. At the other extreme, effects might be highly contextualized – for example, replicating only with sample and situational characteristics that are highly consistent with the original circumstances. The primary contribution of this investigation is to establish a paradigm for testing replicability across samples and settings and provide a rich data set that allows the determinants of replicability to be explored. A secondary purpose is to demonstrate support for replicability for the 13 chosen effects. Ideally, the results will stimulate theoretical developments about the conditions under which replication will be robust to the inevitable variation in circumstances of data collection.

Method

Researcher Recruitment and Data Collection Sites

Project leads posted a call for collaborators to the online forum of the Open Science Collaboration on February 21, 2013 and to the SPSP Discussion List on July 13, 2013. Other colleagues were contacted personally. For inclusion, each replication team had to: (1) follow local ethical procedures, (2) administer the protocol as specified, (3) collect data from at least 80 participants,[1] (4) post a video simulation of the setting and administration procedure, and (5) document key features of recruiting, sample, and any changes to the standard protocol. In total, there were 36 samples and settings that collected data from a total of 6,344 participants (27 data collections in a laboratory and 9 conducted online; 25 from the US, 11 from other countries; see Table 1 for a brief description of sites and for a full descriptions of sites, site characteristics, and participant characteristics by site).

Selection of Replication Studies

Twelve studies producing 13 effects were chosen based on the following criteria:

1. *Suitability for online presentation.* Our primary concern was to give each study a "fair" replication that was true to the original design. By administering the study through a web browser, we were able to ensure procedural consistency across sites.
2. *Length of study.* We selected studies that could be administered quickly so that we could examine many of them in a single study session.
3. *Simple design.* With the exception of one correlational study, we selected studies that featured a simple, two-condition design.

[1] One sample fell short of this requirement ($N = 79$) but was still included in the analysis. All sites were encouraged to collect as many participants as possible beyond the required 80, but the decision to end data collection was determined independently by each site. Researchers had no access to the data prior to completing data collection.

Table 1. Data collection sites

Site identifier	Location	*N*	Online (O) or laboratory (L)	US or international (I)
Abington	Penn State Abington, Abington, PA	84	L	US
Brasilia	University of Brasilia, Brasilia, Brazil	120	L	I
Charles	Charles University, Prague, Czech Republic	84	L	I
Conncoll	Connecticut College, New London, CT	95	L	US
CSUN	California State University, Northridge, LA, CA	96	O	US
Help	HELP University, Malaysia	102	L	I
Ithaca	Ithaca College, Ithaca, NY	90	L	US
JMU	James Madison University, Harrisonburg, VA	174	O	US
KU	Koç University, Istanbul, Turkey	113	O	I
Laurier	Wilfrid Laurier University, Waterloo, Ontario, Canada	112	L	I
LSE	London School of Economics and Political Science, London, UK	277	L	I
Luc	Loyola University Chicago, Chicago, IL	146	L	US
McDaniel	McDaniel College, Westminster, MD	98	O	US
MSVU	Mount Saint Vincent University, Halifax, Nova Scotia, Canada	85	L	I
MTURK	Amazon Mechanical Turk (US workers only)	1,000	O	US
OSU	Ohio State University, Columbus, OH	107	L	US
Oxy	Occidental College, LA, CA	123	L	US
PI	Project Implicit Volunteers (US citizens/residents only)	1,329	O	US
PSU	Penn State University, University Park, PA	95	L	US
QCCUNY	Queens College, City University of New York, NY	103	L	US
QCCUNY2	Queens College, City University of New York, NY	86	L	US
SDSU	SDSU, San Diego, CA	162	L	US
SWPS	University of Social Sciences and Humanities Campus Sopot, Sopot, Poland	79	L	I
SWPSON	Volunteers visiting www.badania.net	169	O	I
TAMU	Texas A&M University, College Station, TX	187	L	US
TAMUC	Texas A&M University-Commerce, Commerce, TX	87	L	US
TAMUON	Texas A&M University, College Station, TX (Online participants)	225	O	US
Tilburg	Tilburg University, Tilburg, Netherlands	80	L	I
UFL	University of Florida, Gainesville, FL	127	L	US
UNIPD	University of Padua, Padua, Italy	144	O	I
UVA	University of Virginia, Charlottesville, VA	81	L	US
VCU	VCU, Richmond, VA	108	L	US
Wisc	University of Wisconsin-Madison, Madison, WI	96	L	US
WKU	Western Kentucky University, Bowling Green, KY	103	L	US
WL	Washington & Lee University, Lexington, VA	90	L	US
WPI	Worcester Polytechnic Institute, Worcester, MA	87	L	US

4. *Diversity of effects.* We sought to diversify the sample of effects by topic, time period of original investigation, and differing levels of certainty and existing impact. Justification for study inclusion is described in the registered proposal (http://osf.io/project/aBEsQ/).

The Replication Studies

All replication studies were translated into the dominant language of the country of data collection (*N* = 7 languages total; 3/6 translations from English were back-translated). Next, we provide a brief description of each experiment, original finding, and known differences between original and replication studies. Most original studies were conducted with paper and pencil, all replications were conducted via computer. Exact wording for each study, including a link to the study, can be found in the supplementary materials. The relevant findings from the original studies can be found in the original proposal.

1. *Sunk costs (Oppenheimer, Meyvis, & Davidenko, 2009).* Sunk costs are those that have already been incurred and cannot be recovered (Knox & Inkster, 1968). Oppenheimer et al. (2009; adapted from Thaler, 1985) asked participants to imagine that they have tickets to see their favorite football team play an important game, but that it is freezing cold on the day of the game. Participants rated their likelihood of attending the game on a 9-point scale (1 = *definitely stay at home*, 9 = *definitely go to the game*). Participants were marginally more likely to go to the game if they had paid for the ticket than if the ticket had been free.

2. *Gain versus loss framing (Tversky & Kahneman, 1981).* The original research showed that changing the focus from losses to gains decreases participants' willingness to take risks – that is, gamble to get a better outcome rather than take a guaranteed result. Participants imagined that the US was preparing for the outbreak of an unusual Asian disease, which is expected to kill 600 people. Participants were then asked to select a course of action to combat the disease from logically identical sets of alternatives framed in terms of gains as follows: Program A will save 200 people (400 people will die), or Program B which has a 1/3 probability that 600 people will be saved (nobody will die) and 2/3 probability that no people will be saved (600 people will die). In the "gain" framing condition, participants are more likely to adopt Program A, while this effect reverses in the loss framing condition. The replication replaced the phrase "the United States" with the country of data collection, and the word "Asian" was omitted from "an unusual Asian disease."
3. *Anchoring (Jacowitz & Kahneman, 1995).* Jacowitz and Kahneman (1995) presented a number of scenarios in which participants estimated size or distance after first receiving a number that was clearly too large or too small. In the original study, participants answered 3 questions about each of 15 topics for which they estimated a quantity. First, they indicated if the quantity was greater or less than an anchor value. Second, they estimated the quantity. Third, they indicated their confidence in their estimate. The original number served as an anchor, biasing estimates to be closer to it. For the purposes of the replication we provided anchoring information before asking just for the estimated quantity for four of the topics from the original study – distance from San Francisco to New York City, population of Chicago, height of Mt. Everest, and babies born per day in the US for countries that use the metric system, we converted anchors to metric units and rounded them.
4. *Retrospective gambler's fallacy (Oppenheimer & Monin, 2009).* Oppenheimer and Monin (2009) investigated whether the rarity of an independent, chance observation influenced beliefs about what occurred before that event. Participants imagined that they saw a man rolling dice in a casino. In one condition, participants imagined witnessing three dice being rolled and all came up 6's. In a second condition two came up 6's and one came up 3. In a third condition, two dice were rolled and both came up 6's. All participants then estimated, in an open-ended format, how many times the man had rolled the dice before they entered the room to watch him. Participants estimated that the man rolled dice more times when they had seen him roll three 6's than when they had seen him roll two 6's or two 6's and a 3. For the replication, the condition in which the man rolls two 6's was removed leaving two conditions.
5. *Low-versus-high category scales (Schwarz, Hippler, Deutsch, & Strack, 1985).* Schwarz and colleagues (1985) demonstrated that people infer from response options what are low and high frequencies of a behavior, and self-assess accordingly. In the original demonstration, participants were asked how much TV they watch daily on a low-frequency scale ranging from "up to half an hour" to "more than two and a half hours," or a high-frequency scale ranging from "up to two and a half hours" to "more than four and a half hours." In the low-frequency condition, fewer participants reported watching TV for more than two and a half hours than in the high-frequency condition.
6. *Norm of reciprocity (Hyman & Sheatsley, 1950).* When confronted with a decision about allowing or denying the same behavior to an ingroup and outgroup, people may feel an obligation to reciprocity, or consistency in their evaluation of the behaviors (Hyman & Sheatsley, 1950). In the original study, American participants answered two questions: whether communist countries should allow American reporters in and allow them to report the news back to American papers and whether America should allow communist reporters into the United States and allow them to report back to their papers. Participants reported more support for allowing communist reporters into America when that question was asked after the question about allowing American reporters into the communist countries. In the replication, we changed the question slightly to ensure the "other country" was a suitable, modern target (North Korea). For international replication, the target country was determined by the researcher heading that replication to ensure suitability (see supplementary materials).
7. *Allowed/Forbidden (Rugg, 1941).* Question phrasing can influence responses. Rugg (1941) found that respondents were less likely to endorse forbidding speeches against democracy than they were to not endorse allowing speeches against democracy. Respondents in the United States were asked, in one condition, if the US should allow speeches against democracy or, in another condition, whether the US should forbid speeches against democracy. Sixty-two percent of participants indicated "No" when asked if speeches against democracy should be allowed, but only 46% indicated "Yes" when asked if these speeches should be forbidden. In the replication, the words "The United States" were replaced with the name of the country the study was administered in.
8. *Quote Attribution (Lorge & Curtiss, 1936).* The source of information has a great impact on how that information is perceived and evaluated. Lorge and Curtiss

(1936) examined how an identical quote would be perceived if it was attributed to a liked or disliked individual. Participants were asked to rate their agreement with a list of quotations. The quotation of interest was, "I hold it that a little rebellion, now and then, is a good thing, and as necessary in the political world as storms are in the physical world." In one condition the quote was attributed to Thomas Jefferson, a liked individual, and in the other it was attributed to Vladimir Lenin, a disliked individual. More agreement was observed when the quote was attributed to Jefferson than Lenin (reported in Moskowitz, 2004). In the replication, we used a quote attributed to either George Washington (liked individual) or Osama Bin Laden (disliked individual).

9. *Flag Priming (Carter, Ferguson, & Hassin, 2011; Study 2)*. The American flag is a powerful symbol in American culture. Carter et al. (2011) examined how subtle exposure to the flag may increase conservatism among US participants. Participants were presented with four photos and asked to estimate the time of day at which they were taken. In the flag-prime condition, the American flag appeared in two of these photos. In the control condition, the same photos were presented without flags. Following the manipulation, participants completed an 8-item questionnaire assessing views toward various political issues (e.g., abortion, gun control, affirmative action). Participants in the flag-primed condition indicated significantly more conservative positions than those in the control condition. The priming stimuli used to replicate this finding were obtained from the authors and identical to those used in the original study. Because it was impractical to edit the images with unique national flags, the American flag was always used as a prime. As a consequence, the replications in the United States were the only ones considered as direct replications. For international replications, the survey questions were adapted slightly to ensure they were appropriate for the political climate of the country, as judged by the researcher heading that particular replication (see supplementary materials). Further, the original authors suggested possible moderators that they have considered since publication of the original study. We included three items at the very end of the replication study to test these moderators: (1) How much do you identify with being American? (1 = *not at all*; 11 = *very much*), (2) To what extent do you think the typical American is a Republican or Democrat? (1 = *Democrat*; 7 = *Republican*), (3) To what extent do you think the typical American is conservative or liberal? (1 = *Liberal*; 7 = *Conservative*).
10. *Currency priming (Caruso, Vohs, Baxter, & Waytz, 2013)*. Money is a powerful symbol. Caruso et al. (2013) provide evidence that merely exposing participants to money increases their endorsement of the current social system. Participants were first presented with demographic questions, with the background of the page manipulated between subjects. In one condition the background showed a faint picture of US$100 bills; in the other condition the background was a blurred, unidentifiable version of the same picture. Next, participants completed an 8-question "system justification scale" (Kay & Jost, 2003). Participants in the money-prime condition scored higher on the system justification scale than those in the control condition. The authors provided the original materials allowing us to construct a near identical replication for US participants. However, the stimuli were modified for international replications in two ways: First, the US dollar was usually replaced with the relevant country's currency (see supplementary materials); Second, the system justification questions were adapted to reflect the name of the relevant country.
11. *Imagined contact (Husnu & Crisp, 2010; Study 1)*. Recent evidence suggests that merely imagining contact with members of ethnic outgroups is sufficient to reduce prejudice toward those groups (Turner, Crisp, & Lambert, 2007). In Husnu and Crisp (2010), British non-Muslim participants were assigned to either imagine interacting with a British Muslim stranger or to imagine that they were walking outdoors (control condition). Participants imagined the scene for one minute, and then described their thoughts for an additional minute before indicating their interest and willingness to interact with British Muslims on a four-item scale. Participants in the "imagined contact" group had significantly higher contact intentions than participants in the control group. In the replication, the word "British" was removed from all references to "British Muslims." Additionally, for the predominately Muslim sample from Turkey the items were adapted so Christians were the outgroup target.
12. *Sex differences in implicit math attitudes (Nosek, Banaji, & Greenwald, 2002)*. As a possible account for the sex gap in participation in science and math, Nosek and colleagues (2002) found that women had more negative implicit attitudes toward math compared to arts than men did in two studies of Yale undergraduates. Participants completed four Implicit Association Tests (IATs) in random order, one of which measured associations of math and arts with positivity and negativity. The replication simplified the design for length to be just a single IAT.
13. *Implicit math attitudes relations with self-reported attitudes (Nosek et al., 2002)*. In the same study as Effect 12, self-reported math attitudes were measured with a composite of feeling thermometers and semantic differential ratings, and the composite was positively related with the implicit measure. The replication used a subset of the explicit items (see supplementary materials).

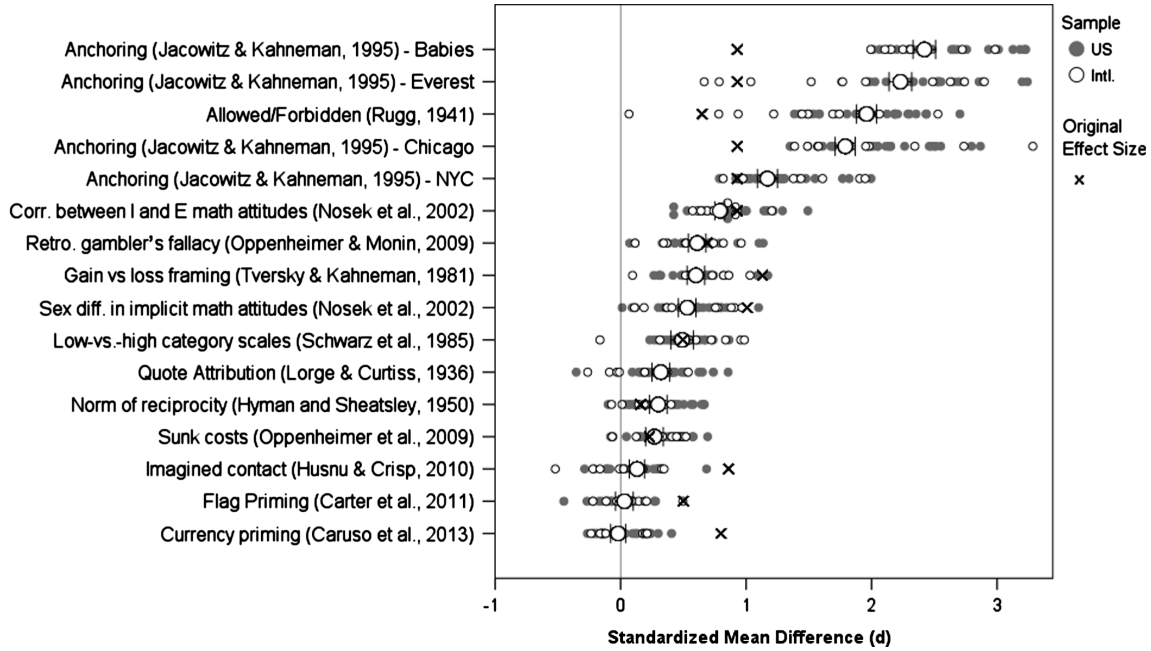

Figure 1. Replication results organized by effect. "X" indicates the effect size obtained in the original study. Large circles represent the aggregate effect size obtained across all participants. Error bars represent 99% noncentral confidence intervals around the effects. Small circles represent the effect sizes obtained within each site (black and white circles for US and international replications, respectively).

Procedure

The experiments were implemented on the Project Implicit infrastructure and all data were automatically recorded in a central database with a code identifying the sample source. After a paragraph of introduction, the studies were presented in a randomized order, except that the math IAT and associated explicit measures were always the final study. After the studies, participants completed an instructional manipulation check (IMC; Oppenheimer et al., 2009), a short demographic questionnaire, and then the moderator measures for flag priming. See Table S1[2] for IMC and summary demographic information by site. The IMC was not analyzed further for this report. Each replication team had a private link for their participants, and they coordinated their own data collection. Experimenters in laboratory studies were not aware of participant condition for each task, and did not interact with participants during data collection unless participants had questions. Investigators who led replications at specific sites completed a questionnaire about the experimental setting (responses summarized in Table S1), and details and videos of each setting along with the actual materials, links to run the study, supplemental tables, datasets, and original proposal are available at https://osf.io/ydpbf/.

Confirmatory Analysis Plan

Prior to data collection we specified a confirmatory analysis plan. All confirmatory analyses are reported either in text or in supplementary materials. A few of the tasks produced highly erratic distributions (particularly anchoring) requiring revisions to those analysis plans. A summary of differences between the original plans and actual analysis is reported in the supplementary materials.

Results

Summary Results

Figure 1 presents an aggregate summary of replications of the 13 effects, presenting each of the four anchoring effects separately. Table 2 presents the original effect size, median effect size, weighted and unweighted effect size and 99% confidence intervals, and proportion of samples that rejected the null hypothesis in the expected and unexpected direction. In the aggregate, 10 of the 13 studies replicated the original results with varying distance from the original

[2] Table names that begin with the prefix "S" (e.g., Table S1) refer to tables that can be found in the supplementary materials. Tables with no prefix are in this paper.

Table 2. Summary confirmatory results for original and replicated effects

	Original study			Unweighted		Weighted		Null hypothesis significance tests by sample (N = 36)			Null hypothesis significance tests of aggregate			
Effect	ES	95% CI lower, upper	Median replication ES	Replication ES	99% CI lower, upper	Replication ES	99% CI lower, upper	Proportion p < .05, opposite direction	Proportion p < .05, same direction	Proportion *ns*	Key statistics	*df*	*N*	*p*
Anchoring – babies born	0.93	.51, 1.33	2.43	2.60	2.41, 2.79	2.42	2.33, 2.51	0.00	1.00	0.00	t = 90.49	5,607	5,609	<.001
Anchoring – Mt. Everest	0.93	.51, 1.33	2.00	2.45	2.12, 2.77	2.23	2.14, 2.32	0.00	1.00	0.00	t = 83.66	5,625	5,627	<.001
Allowed/forbidden	0.65	.57, .73	1.88	1.87	1.58, 2.16	1.96	1.88, 2.04	0.00	0.97	0.03	χ^2 = 3,088.7	1	6,292	<.001
Anchoring – Chicago	0.93	.51, 1.33	1.88	2.05	1.84, 2.25	1.79	1.71, 1.87	0.00	1.00	0.00	t = 65.00	5,282	5,284	<.001
Anchoring – distance to NYC	0.93	.51, 1.33	1.18	1.27	1.13, 1.40	1.17	1.09, 1.25	0.00	1.00	0.00	t = 42.86	5,360	5,362	<.001
Relations between I and E math attitudes	0.93	.77, 1.08	0.84	0.79	0.63, 0.96	0.79	0.75, 0.83	0.00	0.94	0.06	r = .38		5,623	<.001
Retrospective gambler fallacy	0.69	.16, 1.21	0.61	0.59	0.49, 0.70	0.61	0.54, 0.68	0.00	0.83	0.17	t = 24.01	5,940	5,942	<.001
Gain vs. loss framing	1.13	.89, 1.37	0.58	0.62	0.52, 0.71	0.60	0.53, 0.67	0.00	0.86	0.14	χ^2 = 516.4	1	6,271	<.001
Sex differences in implicit math attitudes	1.01	.54, 1.48	0.59	0.56	0.45, 0.68	0.53	0.46, 0.60	0.00	0.71	0.29	t = 19.28	5,840	5,842	<.001
Low vs. high category scales	0.50	.15, .84	0.50	0.51	0.42, 0.61	0.49	0.40, 0.58	0.00	0.67	0.33	χ^2 = 342.4	1	5,899	<.001
Quote attribution	na		0.30	0.31	0.19, 0.42	0.32	0.25, 0.39	0.00	0.47	0.53	t = 12.79	6,323	6,325	<.001
Norm of reciprocity	0.16	.06, .27	0.27	0.27	0.18, 0.36	0.30	0.23, 0.37	0.00	0.36	0.64	χ^2 = 135.3	1	6,276	<.001
Sunk costs	0.23	−.04, .50	0.32	0.31	0.22, 0.39	0.27	0.20, 0.34	0.00	0.50	0.50	t = 10.83	6,328	6,330	<.001
Imagined contact	0.86	.14, 1.57	0.12	0.10	0.00, 0.19	0.13	0.07, 0.19	0.03	0.11	0.86	t = 5.05	6,334	6,336	<.001
Flag priming	0.50	.01, .99	0.02	0.01	−0.07, 0.08	0.03	−0.04, 0.10	0.04	0.00	0.96	t = 0.88	4,894	4,896	0.38
Currency priming	0.80	.05, 1.54	0.00	0.01	−0.06, 0.09	−0.02	−0.08, 0.04	0.00	0.03	0.97	t = −0.79	6,331	6,333	0.83

Notes. All effect sizes (ES) presented in Cohen's d units. Weighted statistics are computed on the whole aggregated dataset (N > 6,000); Unweighted statistics are computed on the disaggregated dataset (N = 36). 95% CI's for original effect sizes used cell sample sizes when available and assumed equal distribution across conditions when not available. The original anchoring article did not provide sufficient information to calculate effect sizes for individual scenarios, therefore an overall effect size is reported. The Anchoring original effect size is a mean point-biserial correlation computed across 15 different questions in a test-retest design, whereas the present replication adopted a between-subjects design with random assignments. One sample was removed from sex difference and relations between implicit and explicit math attitudes because of a systemic error in that laboratory's recording of reaction times. Flag priming includes only US samples. Confidence intervals around the unweighted mean are based on the central normal distribution. Confidence intervals around the weighted effect size are based on noncentral distributions.

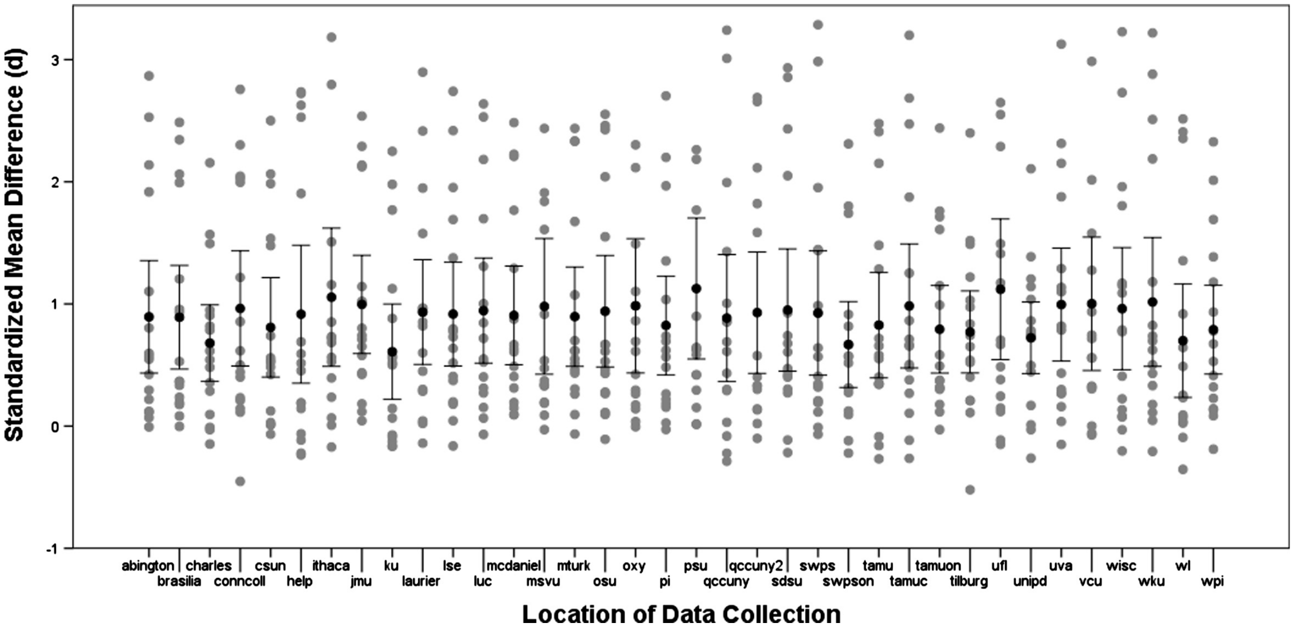

Figure 2. Replication results organized by site. Gray circles represent the effect size obtained for each effect within a site. Black circles represent the mean effect size obtained within a site. Error bars represent 95% confidence interval around the mean.

effect size. One study, imagined contact, showed a significant effect in the expected direction in just 4 of the 36 samples (and once in the wrong direction), but the confidence intervals for the aggregate effect size suggest that it is slightly different than zero. Two studies – flag priming and currency priming – did not replicate the original effects. Each of these had just one p-value < .05 and it was in the wrong direction for flag priming. The aggregate effect size was near zero whether using the median, weighted mean, or unweighted mean. All confidence intervals included zero. Figure 1 presents all 36 samples for flag priming, but only US data collections were counted for the confirmatory analysis (see Table 2). International samples also did not show a flag priming effect (weighted mean $d = .03$, 99% CI [−.04, .10]). To rule out the possibility that the priming effects were contaminated by the contents of other experimental materials, we reexamined only those participants who completed these tasks first. Again, there was no effect (Flag Priming: $t(431) = 0.33$, $p = .75$, 95% CI [−.171, .240], Cohen's $d = .03$; Currency Priming: $t(605) = -0.56$, $p = .57$, 95% CI [−.201, .112], Cohen's $d = .05$).[3]

When an effect size for the original study could be calculated, it is presented as an "X" in Figure 1. For three effects (contact, flag priming, and currency priming), the original effect is larger than for any sample in the present study, with the observed median or mean effect at or below the lower bound of the 95% confidence interval for the original effect.[4] Though the sex difference in implicit math attitudes effect was within the 95% confidence interval of the original result, the replication estimate combined with another large-scale replication (Nosek & Smyth, 2011) suggests that the original effect was an overestimate.

Variation Across Samples and Settings

Figure 1 demonstrates substantial variation for some of the observed effects. That variation could be a function of the true effect size, random error, sample differences, or setting differences. Comparing the intra-class correlation of samples across effects (ICC = .005; $F(35, 385) = 1.06$, $p = .38$, 95% CI [−.027, .065]) with the intra-class correlation of effects across samples (ICC = .75; $F(12,420) = 110.62$, $p < .001$, 95% CI [.60, .89]) suggests that very little in the variability of effect sizes can be attributed to the samples, and substantial variability is attributable to the effect under investigation. To illustrate, Figure 2 shows the same data as Figure 1 organized by sample rather than by effect. There is almost no variation in the average effect size across samples.

However, it is possible that particular samples would elicit larger magnitudes for some effects and smaller magnitudes for others. That might be missed by the aggregate analyses. Table 3 presents tests of whether the heterogeneity of effect sizes for each effect exceeds what is expected by measurement error. Cochran's Q and I^2 statistics

[3] None of the effects was moderated by which position in the study procedure it was administered.

[4] The original anchoring report did not distinguish between topics so the aggregate effect size is reported.

Table 3. Tests of effect size heterogeneity

	Heterogeneity statistics				Moderation tests					
Effect	Q	DF	p	I^2	US or international	p	η_p^2	Laboratory or online	p	η_p^2
Anchoring – babies born	59.71	35	0.01	0.402	0.16	0.69	0.00	16.14	<0.01	0.00
Anchoring – Mt. Everest	152.34	35	<.0001	0.754	94.33	<0.01	0.02	119.56	<0.01	0.02
Allowed/forbidden	180.40	35	<.0001	0.756	70.37	<0.01	0.01	0.55	0.46	0.00
Anchoring – Chicago	312.75	35	<.0001	0.913	0.62	0.43	0.00	32.95	<0.01	0.01
Anchoring – distance to NYC	88.16	35	<.0001	0.643	9.35	<0.01	0.00	15.74	<0.01	0.00
Relations between I and E math attitudes	54.84	34	<.0001	0.401	0.41*	0.52	<.001*	2.80*	0.09	<.001*
Retrospective gambler fallacy	50.83	35	0.04	0.229	0.40	0.53	0.00	0.34	0.56	0.00
Gain vs. loss framing	37.01	35	0.37	0.0001	0.09	0.76	0.00	1.11	0.29	0.00
Sex differences in implicit math attitudes	47.60	34	0.06	0.201	0.82	0.37	0.00	1.07	0.30	0.00
Low vs. high category scales	36.02	35	0.42	0.192	0.16	0.69	0.00	0.02	0.88	0.00
Quote attribution	67.69	35	<.001	0.521	8.81	<0.01	0.001	0.50	0.48	0.00
Norm of reciprocity	38.89	35	0.30	0.172	5.76	0.02	0.00	0.64	0.43	0.00
Sunk costs	35.55	35	0.44	0.092	0.58	0.45	0.00	0.25	0.62	0.00
Imagined contact	45.87	35	0.10	0.206	0.53	0.47	0.00	4.88	0.03	0.00
Flag priming	30.33	35	0.69	0	0.53	0.47	0.00	1.85	0.17	0.00
Currency priming	28.41	35	0.78	0	1.00	0.32	0.00	0.11	0.74	0.00

Notes. Tasks ordered from largest to smallest observed effect size (see Table 2). Heterogeneity tests conducted with R-package metafor. REML was used for estimation for all tests. One sample was removed from sex difference and relations between implicit and explicit math attitudes because of a systemic error in that laboratory's recording of reaction times.

*Moderator statistics are F value of the interaction of condition and the moderator from an ANOVA with condition, country, and location as independent variables with the exception of relations between impl. and expl. math attitudes for is reported the F value associated with the change in R squared after the product term between the independent variable and the moderator is added in a hierarchical linear regression model. Details of all analyses are available in the supplement.

revealed that heterogeneity of effect sizes was largely observed among the very large effects – anchoring, allowed-forbidden, and relations between implicit and explicit attitudes. Only one other effect – quote attribution – showed substantial heterogeneity. This appears to be partly attributable to this effect occurring more strongly in US samples and to a lesser degree in international samples.

To test for moderation by key characteristics of the setting, we conducted a Condition × Country (US or other) × Location (lab or online) ANOVA for each effect. Table 3 presents the essential Condition × Country and Condition × Location effects. Full model results are available in supplementary materials. A total of 10 of the 32 moderation tests were significant, and seven of those were among the largest effects – anchoring and allowed-forbidden. Even including those, none of the moderation effect sizes exceeded a η_p^2 of .022. The heterogeneity in anchoring effects may be attributable to differences in knowledge of the height of Mt Everest, distance to NYC, or population of Chicago between the samples. Overall, whether the sample was collected in the US or elsewhere, or whether data collection occurred online or in the laboratory, had little systematic effect on the observed results.

Additional possible moderators of the flag priming effect were suggested by the original authors. On the US participants only ($N \sim 4{,}670$), with five hierarchical regression models, we tested whether the items moderated the effect of the manipulation. They did not (p's = .48, .80, .62, .07, .05, all $\Delta R^2 < .001$). Details are available in the online supplement.

Discussion

A large-scale replication with 36 samples successfully replicated eleven of 13 classic and contemporary effects in psychological science, some of which are well-known to be robust, and others that have been replicated infrequently or not at all. The original studies produced underestimates of some effects (e.g., anchoring-and-adjustment and allowed versus forbidden message framing), and overestimates of other effects (e.g., imagined contact producing willingness to interact with outgroups in the future). Two effects – flag priming influencing conservatism and currency priming influencing system justification – did not replicate.

A primary goal of this investigation was to examine the heterogeneity of effect sizes by the wide variety of samples and settings, and to provide an example of a paradigm for testing such variation. Some studies were conducted online, others in the laboratory. Some studies were conducted in the United States, others elsewhere. And, a wide variety of educational institutions took part. Surprisingly, these factors did not produce highly heterogeneous effect sizes.

Intraclass correlations suggested that most of the variation in effects was due to the effect under investigation and almost none to the particular sample used. Focused tests of moderating influences elicited sporadic and small effects of the setting, while tests of heterogeneity suggested that most of the variation in effects is attributable to measurement error. Further, heterogeneity was mostly restricted to the largest effects in the sample – counter to an intuition that small effects would be the most likely to be variable across sample and setting. Further, the lack of heterogeneity is particularly interesting considering that there is substantial interest and commentary about the contingency of effects on our two moderators, lab versus online (Gosling, Vazire, Srivastava, & John, 2004; Paolacci, Chandler, & Ipeirotis, 2010), and cultural variation across nations (Henrich et al., 2010).

All told, the main conclusion from this small sample of studies is that, to predict effect size, it is much more important to know what effect is being studied than to know the sample or setting in which it is being studied. The key virtue of the present investigation is that the study procedure was highly standardized across data collection settings. This minimized the likelihood that factors other than sample and setting contributed to systematic variation in effects. At the same time, this conclusion is surely constrained by the small, nonrandom sample of studies represented here. Additionally, the replication sites included in this project cannot capture all possible cultural variation, and most societies sampled were relatively Western, Educated, Industrialized, Rich, and Democratic (WEIRD; Henrich et al., 2010). Nonetheless, the present investigation suggests that we should not necessarily assume that there are differences between samples; indeed, even when moderation was observed in this sample, the effects were still quite robust in each setting.

The present investigation provides a summary analysis of a very large, rich dataset. This dataset will be useful for additional exploratory analysis about replicability in general, and these effects in particular. The data are available for download at the Open Science Framework (https://osf.io/ydpbf/).

Conclusion

This investigation offered novel insights into variation in the replicability of psychological effects, and specific information about the replicability of 13 effects. This methodology – crowdsourcing dozens of laboratories running an identical procedure – can be adapted for a variety of investigations. It allows for increased confidence in the existence of an effect and for the investigation of an effect's dependence on the particular circumstances of data collection (Open Science Collaboration, 2014). Further, a consortium of laboratories could provide mutual support for each other by conducting similar large-scale investigations on original research questions, not just replications. Thus, collective effort could accelerate the identification and verification of extant and novel psychological effects.

Note From the Editors

Commentaries and a rejoinder on this paper are available (Crisp, Miles, & Husnu, 2014; Ferguson, Carter, & Hassin, 2014; Kahneman, 2014; Klein et al., 2014; Monin & Oppenheimer, 2014; Schwarz & Strack, 2014; doi: 10.1027/1864-9335/a000202).

Acknowledgments

We thank Eugene Caruso, Melissa Ferguson, Daniel Oppenheimer, and Norbert Schwarz for their feedback on the design of the materials. This project was supported by grants to the second and fifty-first authors from Project Implicit. Ratliff and Nosek are consultants of Project Implicit, Inc., a nonprofit organization that includes in its mission "to develop and deliver methods for investigating and applying phenomena of implicit social cognition, including especially phenomena of implicit bias based on age, race, gender or other factors." Author contributions: Designed research: R. K., B. N., K. R.; Translated materials: S. B., K. B., M. Brandt, B. B., Z. C., N. F., E. G., F. H., H. I., R. K., R. P., A. V., M. Vianello, M. Vranka; Performed research: R. A., S. B., M. Bernstein, K. B., M. Brandt, C. B., Z. C., J. C., W. C., W. D., T. D., M. E., N. F., D. F., E. G., J. A. H., J. F. H., S. J. H., J. H., H. I., M. J., J. J., H. K., R. K., L. K., J. K., C. L., R. M., W. M., A. N., J. N., G. P., R. P., K. R., A. R., K. S., J. L. S., R. S., T. S., J. S., L. V., D. T., A. V., L. V., M. Vranka, A. L. W., J. W.; Analyzed data: M. Vianello, F. H., R. K.; Wrote paper: B. N., K. R., R. K., M. Vianello, J. C. We report all data exclusions, manipulations, measures, and how we determined our sample sizes either in text or the online supplement. All materials, data, videos of the procedure, and the original preregistered design are available at the project page https://osf.io/ydpbf/.

PREREGISTERED OPEN DATA OPEN MATERIALS

References

Aronson, E. (1992). The return of the repressed: Dissonance theory makes a comeback. *Psychological Inquiry, 3*, 303–311.

Bodenhausen, G. V. (1990). Stereotypes as judgmental heuristics: Evidence of circadian variations in discrimination. *Psychological Science, 1*, 319–322.

Brandt, M. J., IJzerman, H., Dijksterhuis, A., Farach, F. J., Geller, J., Giner-Sorolla, R., . . . van 't Veer, A. (2013). The replication recipe: What makes for a convincing replication? *Journal of Experimental Social Psychology, 50*, 217–224.

Carter, T. J., Ferguson, M. J., & Hassin, R. R. (2011). A single exposure to the American flag shifts support toward Republicanism up to 8 months later. *Psychological Science, 22*, 1011–1018.

Caruso, E. M., Vohs, K. D., Baxter, B., & Waytz, A. (2013). Mere exposure to money increases endorsement of free-market systems and social inequality. *Journal of Experimental Psychology: General, 142*, 301–306.

Cesario, J. (2014). Priming, replication, and the hardest science. *Perspectives on Psychological Science, 9*, 40–48.

Collins, H. M. (1974). The TEA set: Tacit knowledge and scientific networks. *Science Studies, 4*, 165–185.

Crisp, R. J., Miles, E., & Husnu, S. (2014). Support for the replicability of imagined contact effects. Commentaries and rejoinder on Klein et al. (2014). *Social Psychology*. Advance online publication. doi: 10.1027/1864-9335/1000202

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. Commentaries and rejoinder on Klein et al. (2014). *Social Psychology*. Advance online publication. doi: 10.1027/1864-9335/1000202

Gosling, S. D., Vazire, S., Srivastava, S., & John, O. P. (2004). Should we trust web-based studies? A comparative analysis of six preconceptions about Internet questionnaires. *American Psychologist, 59*, 93.

Greenwald, A. G., Pratkanis, A. R., Leippe, M. R., & Baumgardner, M. H. (1986). Under what conditions does theory obstruct research progress? *Psychological Review, 93*, 216–229.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). Most people are not WEIRD. *Nature, 466*, 29.

Henry, J. D., MacLeod, M. S., Phillips, L. H., & Crawford, J. R. (2004). A meta-analytic review of prospective memory and aging. *Psychology and Aging, 19*, 27.

Husnu, S., & Crisp, R. J. (2010). Elaboration enhances the imagined contact effect. *Journal of Experimental Social Psychology, 46*, 943–950.

Hyman, H. H., & Sheatsley, P. B. (1950). The current status of American public opinion. In J. C. Payne (Ed.), *The teaching of contemporary affairs: 21st yearbook of the National Council of Social Studies* (pp. 11–34). New York, NY: National Council of Social Studies.

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS Medicine, 2*, e124.

Jacowitz, K. E., & Kahneman, D. (1995). Measures of anchoring in estimation tasks. *Personality and Social Psychology Bulletin, 21*, 1161–1166.

Kahneman, D. (2014). A new etiquette for replication. Commentaries and rejoinder on Klein et al. (2014). *Social Psychology*. Advance online publication. doi: 10.1027/1864-9335/1000202

Kay, A. C., & Jost, J. T. (2003). Complementary justice: Effects of ''poor but happy'' and ''poor but honest'' stereotype exemplars on system justification and implicit activation of the justice motive. *Journal of Personality and Social Psychology, 85*, 823–837.

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B. Jr., Bahník, Š., Bernstein, M. J., ... Nosek, B. A. (2014). Theory building through replication: Response to commentaries on the ''Many Labs'' replication project. Commentaries and rejoinder on Klein et al. (2014). *Social Psychology*. Advance online publication. doi: 10.1027/1864-9335/1000202

Knox, R. E., & Inkster, J. A. (1968). Postdecision dissonance at post time. *Journal of Personality and Social Psychology, 8*, 319.

Lorge, I., & Curtiss, C. C. (1936). Prestige, suggestion, and attitudes. *The Journal of Social Psychology, 7*, 386–402.

Monin, B., & Oppenheimer, D. M. (2014). The limits of direct replications and the virtues of stimulus sampling. Commentaries and rejoinder on Klein et al. (2014). *Social Psychology*. Advance online publication. doi: 10.1027/1864-9335/1000202

Moskowitz, G. B. (2004). *Social cognition: Understanding self and others*. New York: Guilford Press.

Nosek, B. A., Banaji, M. R., & Greenwald, A. G. (2002). Math = male, Me = female, therefore math ≠ Me. *Journal of Personality and Social Psychology, 83*, 44–59.

Nosek, B. A., & Smyth, F. L. (2011). Implicit social cognitions predict sex differences in math engagement and achievement. *American Educational Research Journal, 48*, 1125–1156.

Open Science Collaboration. (2012). An open, large-scale, collaborative effort to estimate the reproducibility of psychological science. *Perspectives on Psychological Science, 7*, 657–660.

Open Science Collaboration. (2014). The reproducibility project: A model of large-scale collaboration for empirical research on reproducibility. In V. Stodden, F. Leisch, & R. Peng (Eds.), *Implementing reproducible computational research (A volume in the R series)* (pp. 299–323). New York, NY: Taylor & Francis.

Oppenheimer, D. M., Meyvis, T., & Davidenko, N. (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology, 45*, 867–872.

Oppenheimer, D. M., & Monin, B. (2009). The retrospective gambler's fallacy: Unlikely events, constructing the past, and multiple universes. *Judgment and Decision Making, 4*, 326–334.

Paolacci, G., Chandler, J., & Ipeirotis, P. (2010). Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making, 5*, 411–419.

Rugg, D. (1941). Experiments in wording questions: II. *Public Opinion Quarterly, 5*, 91–92.

Schmidt, S. (2009). Shall we really do it again? The powerful concept of replication is neglected in the social sciences. *Review of General Psychology, 13*, 90–100.

Schwarz, N., Hippler, H. J., Deutsch, B., & Strack, F. (1985). Response scales: Effects of category range on reported behavior and comparative judgments. *Public Opinion Quarterly, 49*, 388–395.

Schwarz, N., & Strack, F. (2014). Does merely going through the same moves make for a ''direct'' replication? Concepts, contexts, and operationalizations. Commentaries and rejoinder on Klein et al. (2014). *Social Psychology*. Advance online publication. doi: 10.1027/1864-9335/1000202

Thaler, R. (1985). Mental accounting and consumer choice. *Marketing Science, 4*, 199–214.

Turner, R. N., Crisp, R. J., & Lambert, E. (2007). Imagining intergroup contact can improve intergroup attitudes. *Group Processes and Intergroup Relations, 10*, 427–441.

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science, 211*, 453–458.

Received March 4, 2013
Accepted November 26, 2013
Published online May 19, 2014

Richard A. Klein

Department of Psychology
University of Florida
Gainesville, FL 32611
USA
E-mail raklein@ufl.edu

Forming Impressions of Personality

A Replication and Review of Asch's (1946) Evidence for a Primacy-of-Warmth Effect in Impression Formation

Sanne Nauts,[1] Oliver Langner,[2] Inge Huijsmans,[1] Roos Vonk,[1] and Daniël H. J. Wigboldus[1]

[1]Behavioural Science Institute, Radboud University Nijmegen, The Netherlands, [2]Friedrich Schiller University Jena, Germany

Abstract. Asch's seminal research on "Forming Impressions of Personality" (1946) has widely been cited as providing evidence for a primacy-of-warmth effect, suggesting that warmth-related judgments have a stronger influence on impressions of personality than competence-related judgments (e.g., Fiske, Cuddy, & Glick, 2007; Wojciszke, 2005). Because this effect does not fit with Asch's Gestalt-view on impression formation and does not readily follow from the data presented in his original paper, the goal of the present study was to critically examine and replicate the studies of Asch's paper that are most relevant to the primacy-of-warmth effect. We found no evidence for a primacy-of-warmth effect. Instead, the role of warmth was highly context-dependent, and competence was at least as important in shaping impressions as warmth.

Keywords: replication, primacy-of-warmth, person perception

Social psychological laboratories have undergone considerable change since the publication of Asch's "Forming Impressions of Personality" in 1946, leading to the inevitable demise of punch cards and slide carousels in favor of more advanced experimental equipment. Still, the basic methodology underlying present-day person perception research is strongly grounded in Asch's paradigm-shifting paper, in which impression formation was studied in a controlled laboratory setting, yielding high internal validity and experimental precision (Fiske, Cuddy, & Glick, 2007; Gilbert, 1998). Beyond the methodological realm, Asch's studies have also laid much of the groundwork for influential theories about person perception (e.g., attribution theory; Jones & Davis, 1965; the continuum model of impression formation; Fiske, Neuberg, Beattie, & Milberg, 1987).

Written long before the dawn of bite-size science (Bertamini & Munafo, 2012) and the advice to "role-play grandma" to create a clear storyline (Bem, 1987, p. 27), "Forming Impressions of Personality" (Asch, 1946) is as interesting as it is multifaceted. Although there is not one unitary message to be taken from the work (which has been cited over 2,750 times), the message that seems to have most strongly resonated with present-day researchers concerns the primacy-of-warmth effect. Primacy-of-warmth[1] (e.g., Fiske et al., 2007; Wojciszke, 2005) entails that warmth has a primary role in impression formation, in the sense that warmth-related information has a stronger influence on impressions than competence-related information (Wojciszke, Bazinska, & Jaworski, 1998).

The present research aims to critically examine the evidence that Asch's (1946) research provides for the primacy-of-warmth effect. Moreover, we conducted a direct replication of those studies in Asch's publication that are particularly relevant to this effect. Replication attempts of Asch's work abound (e.g., Ahlering & Parker, 1989; Anderson & Barrios, 1961; Babad, Kaplowitz, & Darley, 1999; Grace & Greenshields, 1960; Hendrick & Constantini, 1970; Kelley, 1950; Luchins, 1948; Luchins & Luchins, 1986; McCarthy & Skowronski, 2011; McKelvie, 1990; Mensh & Wishner, 1947; Pettijohn, Pettijohn, & McDermott, 2009; Semin, 1989; Singh, Onglacto, Sriram, & Tay, 1997; Veness & Brierley, 1963; Wishner, 1960), but most are conceptual rather than direct replications, many are incomplete, few relate to primacy-of-warmth, and some results do not concur with Asch's original findings. Although "Forming Impressions of Personality" has been regarded as a first demonstration of the primacy-of-warmth effect (e.g., Abele & Bruckmüller, 2011; Abele & Wojciszke, 2007; Cuddy, Fiske, & Glick, 2008; Judd, James-Hawkins, Yzerbyt, & Kashima, 2005; Kervyn, Yzerbyt, & Judd, 2010; Richetin, Durante, Mari, Perugini, & Volpato, 2012; Vonk, 1994), it is unclear whether Asch's

[1] In the present research, in line with the recommendations by Fiske et al. (2007), warmth is used as an omnibus term that includes dimensions such as other-profitability (Peeters & Czapinski, 1990), morality (Wojciszke, 2005), trustworthiness (Todorov, Said, Engell, & Oosterhof, 2008), and social evaluation (Rosenberg, Nelson, & Vivekananthan, 1968).

DOI: 10.1027/1864-9335/a000179

original studies provide replicable evidence for the effect. Many studies suggest that warmth plays an important role in impression formation (for a review, see Fiske et al., 2007; Wojciszke, 2005), but we wonder if Asch has befittingly been cited as the progenitor of this effect. We believe that Asch's Gestalt theory, if anything, addresses the limitations and boundary conditions of primacy-of-warmth, and we wonder if his data provide any evidence for the effect itself. Before discussing the latter point, we first provide a short overview of Asch's main findings.

Overview of Asch (1946)

In the original publication (Asch, 1946), 10 studies were reported (total $N = 834$)[2] in which participants read different lists of traits. For example, in the classic warm-cold study (Study I), participants were either exposed to a trait-list containing *warm* or to a trait-list containing *cold*, keeping all other traits identical between groups. Participants then wrote down their impression of the target person (open-ended measure), selected which traits from a trait-pair list were most applicable to the target (trait-pair choice measure; see Appendix), and ranked the original traits according to importance for their impression (ranking measure). From this study, Asch concluded that participants treated warm and cold as relatively central in forming impressions, transforming their impressions when warm was replaced by cold. The subsequent nine studies featured variations to this paradigm, introducing other traits, manipulating the order of traits, asking participants to give synonyms for elements of the trait lists, or asking for judgments on how the provided traits are related. Table A1 of the Additional Findings provides a summary of all 10 studies.

Based on these experiments, Asch (1946) concluded that perceivers form coherent, unitary impressions of others. For such unitary impressions, perceivers attribute different meanings and weights to traits, assigning central roles to some traits (these determine the meaning/function of other traits) and peripheral roles to others (their meaning/function is determined by central traits). Traits interact dynamically in shaping each other's interpretation: Which traits become central or peripheral is fully determined by the trait context. Thus, *warm* was central in Asch's Study I when accompanied by traits like *intelligent, skillful, industrious, determined, practical*, and *cautious*, but was peripheral in Asch's Study IV when accompanied by traits like *obedient, weak, shallow, unambitious*, and *vain*. Asch suggests that changing the context does not merely lead to affective shifts (or Halo effects), but modifies the entire Gestalt of the impression and the cognitive content of the traits within this Gestalt. Or, as Asch puts it: "the gaiety of an intelligent man is no more or less than the gaiety of a stupid man: it is different in quality" (p. 287).

Interpretations of Asch's Work

Much like punch cards and slide carousels, the Gestalt-view on impression formation has slowly but surely gone out of fashion (partly because there were more simple explanations for Asch's 1946 data, e.g., Anderson, 1981; Rosenberg, Nelson, & Vivekananthan, 1968; Wishner, 1960), though some of its premises have resonated in typological models of impression formation (e.g., Anderson & Sedikides, 1991; Sedikides & Anderson, 1994). These typological models failed to gain a strong foothold in the field: Instead, dimensional models became increasingly popular. Dimensional models suggest that impressions of personality can be captured by a limited number of domains (such as warmth and competence; e.g., Fiske et al., 2007), and have given rise to an increase in research on the primacy-of-warmth effect.

Introductory textbooks presently put more emphasis on Asch's research (1946) as providing evidence for a primacy-of-warmth effect than on the Gestalt-model that was its actual focus. Many textbooks refer only to Study I, concluding that Asch's research shows that warmth is primary in impression formation[3] (e.g., Baron & Byrne, 2004; DeLamater & Meyers, 2010; Franzoi, 2009; Hogg & Vaughan, 2011; Kassin, Fein, & Markus, 2011; Pennington, 2000; Stainton-Rogers, 2011; Taylor, Peplau, & Sears, 2006; Worchel, Cooper, Goethals, & Olson, 2000; for a notable exception, see Hewstone, Stroebe, & Jonas, 2007). Although Asch acknowledges that warmth plays an important role in impression formation, in his view, any trait can be central as well as peripheral. Thus, no trait is central by design, and even traits of special importance (such as *warm* and *cold*) may become peripheral in some contexts, as the meaning and weight of any trait is context-dependent. This ever-changing, context-dependent nature of centrality that is a key element of Gestalt-models seems to be at least somewhat at odds with the much more simple and parsimonious view that is portrayed by dimensional models, in which warmth is usually seen as central (and as primary over competence).

Evidence for Primacy-of-Warmth in Asch's (1946) Data

Asch's (1946) theorizing and the results of his Study IV do not support the primacy-of-warmth effect; the reason why he has been widely cited as the progenitor of this effect is because of his first study (Study I, or the classic warm-cold study). In our view, this study does not provide unequivocal evidence for primacy-of-warmth, as is apparent from the three measures Asch used in his research

[2] A well-informed reader may notice that Asch writes in his introduction that he tested over a 1,000 participants, but the results of only 834 are reported.

[3] Although some authors additionally refer to Study VI or VII about primacy-effects.

(the open-ended, trait-pair choice, and ranking measures). We will now discuss each of these measures in turn.

In the open-ended measure, participants wrote down their general impression of the target. Asch (1946) based his conclusions to a large extent on these open-ended responses, providing many anecdotes, but never systematically analyzing the data. Consequently, the interpretation of these data was heavily contested by his contemporaries (e.g., Gollin, 1954; Luchins, 1948). Because replications of Asch's research did not include systematic analysis of open-ended responses either (e.g., Mensh & Wishner, 1947; Semin, 1989; Veness & Brierley, 1963), as yet it is unclear to what extent they provide evidence for primacy-of-warmth (or for effects that were the actual focus of Asch's paper; more information on those effects is available in our Additional Findings).

For the trait-pair choice measure, participants chose which trait (out of a pair) was most applicable to the target. The results suggest that changing a trait from positive (e.g., *warm*) to negative (e.g., *cold*) made the overall impression more negative (negative traits of the pairs were chosen more frequently). Though this effect has been replicated repeatedly (e.g., Mensh & Wishner, 1947; Veness & Brierley, 1963; Semin, 1989), it may not provide the most stringent test of the primacy-of-warmth hypothesis, as changing any positive trait into a negative one is likely to influence the overall valence of the trait-list.

For the ranking measure, participants ranked all traits from the stimulus list from most to least important to their impression. The results for this measure do not provide any evidence for a primacy-of-warmth effect: In Study I, warmth was ranked highest by 6 out of 42 participants, the exact amount that could be expected by chance (given that there are seven options). This limitation was acknowledged by Asch (1946), but seems to have been overlooked in many later references to his work. Unfortunately, the original data are reported incompletely, making it difficult to interpret which trait was primary in people's impressions (considering that it clearly was not warmth).

In sum, Asch's data (1946) do not provide clear evidence for a primacy-of-warmth effect. The open-ended responses that were important in Asch's theorizing were not systematically analyzed; the trait-pair choice measure seems unfit to test primacy-of-warmth; and the results of the ranking measure suggest that warmth was not central in determining participant's impressions. In addition, several factors make it difficult to estimate the extent of evidence for primacy-of-warmth in Asch's data: Several studies were insufficiently powered, the open-ended questions lacked a clear coding scheme, only incomplete accounts of the data were provided, and no quantitative statistical analyses were conducted.[4] In the present research, we conducted a direct replication of Asch's Studies I, III, and IV (the studies that are most relevant to the primacy-of-warmth effect; see Table A1 of the Additional Findings for an overview) to get more insight into the evidence Asch provides for a primacy-of-warmth in impression formation.

Method

Our replication attempt was highly similar to Asch's original work, but there are several methodological differences. First of all, we increased power and added statistical analyses of the ranking data and trait-pair choice data and systematic analyses of the open-ended responses, which were absent in the original publication. Second, we administered the study online through Amazon's MechanicalTurk (MTurk) instead of in a laboratory with student participants (the recent "Many Labs project" suggests that MTurk replications and laboratory replications yield highly similar results; Klein et al., 2014). Third, we randomly assigned participants to one of seven conditions to aid comparability of the studies (Asch ran the conditions in three separate studies). Fourth, the study proposal and materials were preregistered.

Participants and Design

Participants were recruited through MTurk in exchange for $1. Of 1,140 participants, 117 were removed because English was not their native language or because they failed to pass an instructional manipulation check (Oppenheimer, Meyvis, & Davidenko, 2009). The remaining 1,023 participants[5] (474 males) were on average 33 years old (range 18–75 years).

Participants were randomly assigned to one of seven trait lists (see Table 1). According to Asch (1946), *warm* and *cold* should be central in Conditions 1 and 2 when accompanied by traits like *intelligent, skillful, industrious, determined, practical*, and *cautious* (original Study I), but not in Conditions 3–5 when accompanied by traits like *obedient, weak, shallow, unambitious*, and *vain* (original Study IV). In Conditions 6 and 7 (original Study III), the same lists as in Conditions 1 and 2 were used with *warm* and *cold* replaced by *polite* and *blunt*. Asch included these lists to show that *polite* and *blunt* would be less central than *warm* and *cold*, suggesting that the centrality of a trait is determined by the interplay between that specific trait and the context.

Procedure

After providing informed consent, participants were instructed that they would see several traits on a computer screen, all of which belonged to the same person. Traits

[4] Asch's research was published in 1946, when reporting statistical analyses was not yet customary (and many analyses still had to be invented).

[5] Based on the literature by Cohen (1992) and power analysis with G*Power (Faul, Erdfelder, Lang, & Buchner, 2007), we had aimed to run 1,050 participants in total.

Table 1. Conditions included in our replication and the stimulus list participants were exposed to

Condition in Asch (1946)	Replication condition	Stimulus list
Study I, Group A	Condition 1	Intelligent, skillful, industrious, warm, determined, practical, cautious
Study I, Group B	Condition 2	Intelligent, skillful, industrious, cold, determined, practical, cautious
Study IV, Group A	Condition 3	Obedient, weak, shallow, warm, unambitious, vain
Study IV, Group B	Condition 4	Vain, shrewd, unscrupulous, warm, shallow, envious
Study IV, Group C	Condition 5	Intelligent, skillful, sincere, cold, conscientious, helpful, modest
Study III, Group A	Condition 6	Intelligent, skillful, industrious, polite, determined, practical, cautious
Study III, Group B	Condition 7	Intelligent, skillful, industrious, blunt, determined, practical, cautious

Table 2. Rankings of traits in Condition 1 (the warm-list), and average rank for each trait (N = 159)

	Trait						
Rank	Intelligent	Skillful	Industrious	Warm	Determined	Practical	Cautious
1	55.3%	1.9%	6.3%	19.5% (14%)	9.4%	5.0%	2.5%
2	23.9%	18.9%	5.7%	15.7% (35%)	17.6%	15.1%	3.1%
3	10.1%	28.9%	10.1%	17.0% (10%)	13.8%	15.1%	5.0%
4	5.0%	17.6%	17.0%	11.9% (10%)	19.5%	17.6%	11.3%
5	0.6%	18.9%	16.4%	11.3% (10%)	18.2%	20.8%	13.8%
6	3.1%	8.2%	24.5%	10.7% (7%)	14.5%	18.9%	20.1%
7	1.9%	5.7%	20.1%	13.8% (14%)	6.9%	7.5%	44.0%
AVG rank	1.89^{a}	3.80^{b}	4.86^{d}	3.67^{b}	$3.91^{b,c}$	4.21^{c}	5.67^{e}

Notes. Lower average ranks indicate that participants ranked the trait as more important in determining their impressions. Numbers in parentheses indicate the results reported in Asch's original study (1946; N = 42). Ranks not sharing the same superscript are significantly different from each other ($p < .05$).

were presented one by one for 3 s each, with 2 s between traits. Then, all traits were repeated once (cf. Asch, 1946). Next, participants were asked to type in their impression of the target person (open-ended measure). Subsequently, they were exposed to lists of trait pairs (see Appendix) and were asked to choose which trait from each pair was most in accordance with their target impression (trait-pair choice measure). Following that, all traits of the target were presented once again, and participants had to rank them in order of importance for their impression, starting with the most important trait (Rank 1) and proceeding to the least important one (Rank 6 or 7, depending on the condition; ranking measure). Finally, participants completed some demographic questions and were debriefed.

Results

Warmth in Rankings

To find out if *warm* and *cold* were more central than other traits within Conditions 1 and 2, we first investigated which traits were ranked as most influential in shaping perceivers' impressions (see Table 2). Only 19.5% of participants ranked *warm* as the most important trait in determining their impression, whereas 55.3% ranked *intelligent* as the most important trait. Wilcoxon signed rank tests confirmed that *intelligent* received lower average ranks (indicating higher importance) than *warm*, $Z(2, N = 159) = -7.27$, $p < .001$, $r = 0.41$, with mean ranks of 1.89 and 3.67, respectively. Not warmth, but intelligence, was primary in shaping participants' impressions. In fact, the rank frequencies for warmth did not significantly differ from a flat distribution, $X^2(2, N = 159) = 7.11$, $p = .31$, suggesting that warmth did not receive higher (or lower) rankings than could be expected based on chance alone. In sum, the results of the ranking data do not provide evidence for a primacy-of-warmth effect: intelligence, not warmth, was the primary determinant of participant's impressions of personality.

In Condition 2, perceivers saw the same trait-list as in Condition 1, except for *warm* (which was replaced by *cold*). As apparent from Table 3, 30.0% of participants ranked *cold* as the most important trait in determining their impression, whereas 36.2% ranked *intelligent* as the most important trait. Wilcoxon signed rank tests confirmed that *intelligent* received lower average ranks than *cold*, $Z(2, N = 130) = -4.39$, $p < .001$, $r = 0.14$, with mean ranks of 2.34 and 3.77, respectively. Unlike for *warm*, the distribution of rank frequencies for *cold* did differ from a flat distribution, $X^2(2, N = 130) = 64.22, p < .001$, Cohen's $w = 0.70$. More specifically, *cold* was selected as most important trait by 30.0% of participants and as least important trait by 29.2% of participants: Participants seemed to have a polarized view on the importance of coldness, ranking it as important and as unimportant relatively frequently.

Table 3. Rankings of traits in Condition 2 (the cold-list), and average rank for each trait ($N = 130$)

	Trait						
Rank	Intelligent	Skillful	Industrious	Cold	Determined	Practical	Cautious
1	36.2%	5.4%	3.1%	30.0% (27%)	16.9%	6.2%	2.3%
2	30.8%	15.4%	13.8%	13.1% (21%)	16.9%	6.9%	3.1%
3	10.0%	23.8%	12.3%	10.0% (2%)	23.8%	10.0%	10.0%
4	14.6%	14.6%	13.8%	7.7% (5%)	18.5%	20.0%	10.8%
5	5.4%	23.1%	21.5%	4.6% (7%)	11.5%	17.7%	16.2%
6	0.8%	11.5%	16.2%	5.4% (5%)	10.0%	27.7%	28.5%
7	2.3%	6.2%	19.2%	29.2% (33%)	2.3%	11.5%	29.2%
AVG rank	2.34^{a}	3.94^{c}	4.62^{d}	3.77^{c}	3.30^{b}	4.65^{d}	5.38^{e}

Notes. Lower average ranks indicate that participants ranked the trait as more important in determining their impressions. Numbers in parentheses indicate the results reported in Asch's original study (1946; $N = 41$). Ranks not sharing the same superscript are significantly different from each other ($p < .05$).

Concurring with Condition 1, the results for the cold-list do not provide clear evidence for a primacy-of-warmth effect. *Intelligent*, not *cold*, seemed the primary determinant of participant's impressions of personality. However, given that *cold* received relatively polarized ranks, the results are not as unequivocal as they are for Condition 1.

Tables 4–8 contain the average ranks for all remaining experimental conditions. As in Conditions 1 and 2, *intelligent* was rated as the most important trait in all conditions that included this trait (ranked highest by 53.5%–60.4% of participants), whereas *warm* and *cold* were not central in any condition that included one of these traits (ranked highest by 6.6%–7.8% of participants). Importantly, the centrality of *warm* and *cold* in Conditions 1 and 2 was even more absent in Conditions 3, 4, and 5, in accordance with Asch's hypothesis (1946) that the centrality of warmth

Table 4. Rankings of traits in Condition 3, and average rank for each trait ($N = 143$)

	Trait					
Rank	Obedient	Weak	Shallow	Warm	Unambitious	Vain
1	21.7%	25.2%	18.2%	7.0%	17.5%	10.5%
2	14.7%	19.6%	23.8%	7.0%	11.9%	23.1%
3	14.0%	17.5%	17.5%	9.8%	25.9%	15.4%
4	14.0%	21.0%	18.2%	12.6%	16.1%	18.2%
5	20.3%	11.2%	13.3%	25.2%	15.4%	14.7%
6	15.4%	5.6%	9.1%	38.5%	13.3%	18.2%
AVG rank	3.43^{b,c}	2.90^{a}	3.12^{a,b}	4.57^{d}	3.40^{b,c}	3.58^{c}

Notes. Lower average ranks indicate that participants ranked the trait as more important in determining their impressions. Ranks not sharing the same superscript are significantly different from each other ($p < .05$).

Table 5. Rankings of traits in Condition 4, and average rank for each trait ($N = 151$)

	Trait					
Rank	Vain	Shrewd	Unscrupulous	Warm	Shallow	Envious
1	44.0%	10.7%	18.7%	6.6%	16.0%	4.0%
2	16.7%	22.0%	17.3%	2.0%	27.3%	14.7%
3	18.0%	19.3%	15.3%	3.3%	23.3%	20.7%
4	15.3%	22.7%	14.0%	12.6%	16.0%	19.3%
5	3.3%	18.7%	24.0%	8.6%	11.3%	34.0%
6	2.7%	6.7%	10.7%	66.2%	6.0%	7.3%
AVG Rank	2.25^{a}	3.37^{c}	3.39^{c}	5.16^{e}	2.97^{b}	3.87^{d}

Notes. Lower average ranks indicate that participants ranked the trait as more important in determining their impressions. Ranks not sharing the same superscript are significantly different from each other ($p < .05$).

Table 6. Rankings of traits in Condition 5, and average rank for each trait (N = 129)

	Trait						
Rank	Intelligent	Skillful	Sincere	Cold	Conscientious	Helpful	Modest
1	53.5%	6.2%	17.8%	7.8%	7.0%	7.0%	0.8%
2	17.1%	21.7%	17.1%	10.1%	8.5%	17.1%	8.5%
3	7.8%	17.1%	11.6%	16.3%	15.5%	22.5%	9.3%
4	11.6%	11.6%	17.1%	6.2%	19.4%	22.5%	11.6%
5	6.2%	17.8%	20.9%	4.7%	17.8%	14.0%	18.6%
6	2.3%	19.4%	12.4%	8.5%	17.8%	10.9%	28.7%
7	1.6%	6.2%	3.1%	46.5%	14.0%	6.2%	22.5%
AVG rank	2.13^{a}	$3.96^{b,c}$	3.56^{b}	$5.02^{d,e}$	4.42^{c}	3.77^{b}	5.15^{e}

Notes. Lower average ranks indicate that participants ranked the trait as more important in determining their impressions. Ranks not sharing the same superscript are significantly different from each other (p < .05).

Table 7. Rankings of traits in Condition 6 (the polite-list), and average rank for each trait (N = 159)

	Trait						
Rank	Intelligent	Skillful	Industrious	Polite	Determined	Practical	Cautious
1	60.4%	5.0%	8.8%	10.7% (0%)	8.2%	3.1%	3.8%
2	17.6%	19.5%	10.1%	15.7% (0%)	23.9%	9.4%	3.8%
3	8.2%	27.0%	13.2%	15.7% (0%)	15.7%	13.2%	6.9%
4	6.3%	21.4%	13.8%	14.5% (10%)	19.5%	17.0%	7.5%
5	2.5%	12.6%	13.8%	13.2% (16%)	17.0%	25.8%	15.1%
6	3.1%	8.8%	17.0%	15.1% (21%)	8.8%	25.2%	22.0%
7	1.9%	5.7%	23.3%	15.1% (53%)	6.9%	6.3%	40.9%
AVG Rank	1.90^{a}	3.66^{b}	$4.58^{c,d}$	4.10^{c}	3.67^{b}	4.53^{d}	5.56^{e}

Notes. Lower average ranks indicate that participants ranked the trait as more important in determining their impressions. Numbers in parentheses indicate the results reported in Asch's original study (1946; N = 19). Ranks not sharing the same superscript are significantly different from each other (p < .05).

Table 8. Rankings of traits in Condition 7 (the blunt-list), and average rank for each trait (N = 152)

	Trait						
Rank	Intelligent	Skillful	Industrious	Blunt	Determined	Practical	Cautious
1	57.2%	6.6%	7.9%	5.9% (0%)	15.8%	5.9%	0.7%
2	19.7%	17.8%	17.1%	5.9% (15%)	23.7%	10.5%	5.3%
3	5.9%	24.3%	13.8%	12.5% (12%)	22.4%	15.1%	5.9%
4	4.6%	20.4%	16.4%	6.6% (19%)	20.4%	22.4%	9.2%
5	7.2%	13.8%	12.5%	9.9% (23%)	10.5%	26.3%	19.7%
6	2.6%	10.5%	20.4%	15.8% (4%)	5.3%	12.5%	32.9%
7	2.6%	6.6%	11.8%	43.4% (27%)	2.0%	7.2%	26.3%
AVG rank	2.03^{a}	3.75^{c}	4.17^{d}	5.30^{e}	3.10^{b}	4.19^{d}	5.46^{e}

Notes. Lower average ranks indicate that participants ranked the trait as more important in determining their impressions. Numbers in parentheses indicate the results reported in Asch's original study (1946; N = 26). Ranks not sharing the same superscript are significantly different from each other (p < .05).

is context-dependent. In fact, in these conditions, warmth and coldness received the lowest rank out of the entire trait lists, suggesting that they were the least important traits in determining participants' impressions.

Warmth in Open-Ended Descriptions

We further investigated the evidence for primacy-of-warmth in Conditions 1 and 2 by applying content analysis to the open-ended responses. If the warm-cold dimension was at the heart of participants' impressions, *warm* and *cold* should be mentioned more often in their descriptions of the target person than any other trait from the presented lists. We thus simply counted the occurrence of all presented traits in participants' descriptions of the target person (plus close synonyms and common incorrect spellings, e.g., *inteligent* instead of *intelligent*). In Condition 1, *warm* and *intelligent* were mentioned about equally often, $F < 1$, with means of 0.23 and 0.22, respectively. Further, both were mentioned more frequently than any other trait (means between 0.01 and 0.11, all Fs > 6.32, all ps < .05, all η_p^2's = .04–.06). Contrary to the predictions based on a primacy-of-warmth approach, participants were as likely to mention intelligence in their description of the target person as they were to mention warmth.

In Condition 2, *cold* versus *intelligent* were mentioned equally often, $F < 1$, with means of 0.27 and 0.24, respectively. All other traits were mentioned less frequently than both intelligence and coldness (the difference between *cold* and *determined* was only marginally significant; means between 0.01 and 0.15, all Fs > 3.18, all ps < .08, all η_p^2's = .04–.05). Contrary to primacy-of-warmth, participants mentioned intelligence in their descriptions of the target person as much as coldness. These results are consistent with those for the ranking measure, in that neither provides evidence for a primacy-of-warmth effect in impression formation. Both measures suggest that warmth is not the primary determinant of perceivers' impressions, and that intelligence (a competence-related trait) seems at least equally important.

A possible disadvantage of the above analysis is that some warmth-related inferences may not have been part of the stimulus list: Instead of responding that the target person is warm, participants may have inferred the target to be trustworthy, helpful, or considerate. To pick up on these indirect warmth inferences, we generated an index of how warmth related the traits mentioned in the descriptions were. All traits mentioned by participants were rated by a separate group of participants ($N = 33$) on how warm and competent a person with that specific trait is (on a 7-point scale). To determine which words in participants' descriptions were traits, we used Anderson's list of personality traits (Anderson, 1968); only words included in this list were considered in the present analysis[6]. We generated a warmth index for 188 traits in this way: First, we calculated scores for warmth- and competence-relatedness by reverting the ratings to absolute values of the scores centered around the midpoint of the scale (e.g., the ratings one and seven would both be reverted to three, as both scores have a distance of three points to the midpoint of the scale). Next, we calculated the difference between competence-relatedness scores and warmth-relatedness scores, forming a warmth-index. Positive warmth-indices appear for traits that are more strongly related to warmth than to competence. Contrary to predictions based on primacy-of-warmth, participants used traits more strongly related to competence in Condition 1, $t(136) = -3.81$, $p < .001$, with an average warmth-index of −0.33, Cohen's $d = -0.32$. In Condition 2, the average warmth-index was not significantly different from zero, $t(103) = -0.68$, $p = .50$, $M = -0.08$, suggesting that the traits participants used were overall equally related to competence and warmth.

In sum, the descriptions participants provided about the target person contained many traits that were not part of the originally presented trait lists, suggesting that participants went beyond the information given and made inferences about the target person's other traits. However, even when taking the inferred traits into account (instead of limiting our search to the words *warm* and *cold*), we did not find evidence for primacy-of-warmth. Instead, the used traits were at least as strongly related to competence as they were related to warmth, suggesting that warmth was not at the heart of participants' descriptions of the target person.

Finally, to check whether our textual analysis may have missed subtle references to warmth, we asked an independent coder to rate for 350 (out of 1,023) randomly selected descriptions to what extent warmth or coldness was conveyed (more information is available in the Additional Findings). The descriptions of 54% of participants in Condition 1 and 36% in Condition 2 did not include any reference to warmth, showing that a substantial amount of participants did not refer to the warm-cold dimension, but solely focused on competence. As apparent from Table A3 in the Additional Findings, the function, meaning, and weight of warmth (if it was mentioned) differed strongly across conditions: For example, in some conditions, warmth was interpreted as meaning the person was truly nice and kind-hearted; in others, it was interpreted as a way for cold-hearted people to manipulate others. More information on the interpretation of warmth in different conditions is available in the Additional Findings.

In sum, the open-ended descriptions do not provide evidence for a primacy-of-warmth effect. Participants were not more likely to mention warmth in their descriptions of the target person than to mention intelligence; the traits they discussed in their descriptions were at least as strongly related to competence as they were to warmth; and a large part of participants did not make any references to warmth whatsoever.

Changes in Valence

One reason for Asch (1946) to conclude that warmth was central in impression formation was that the valence of impressions in his studies seemed to change dramatically when replacing *warm* by *cold* (as in Asch's original Study I), but not when replacing *polite* by *blunt* (as in Asch's original Study IV). To test this effect, which was not

[6] Some participants did not use any trait words in their description of the target person that are part of the Anderson (1968) trait-list. These participants were excluded from this analysis.

quantified in Asch's original paper, we used textual analysis for assessing the valence of participants' descriptions of the target person in the open-ended responses.

After removing capitals and punctuation, we used a sentiment dictionary (Wilson, Wiebe, & Hoffmann, 2005) to establish the average valence of all descriptions. The dictionary we used contains the valence (positive, negative, or neutral) of 8,220 words in the English language. An average valence index for each description was determined by first counting the number of positive and negative words, and then subtracting the number of negative words from the number of positive words. In the resulting index, higher scores reflect more positive descriptions. As expected, descriptions were more positive for *warm*, M (Condition 1) = 3.24, than for *cold*, M (Condition 2) = 1.77, $F(1, 288) = 31.54$, $p < .001$. $\eta_p^2 = .10$. Changing *polite* to *blunt*, however, did not affect the valence of the impression ($F < 1$). In line with Asch's theorizing, changing *warm* to *cold* had a more pronounced influence on perceiver's impressions than changing *polite* to *blunt*. Thus, although the ranking measure and use of warmth-related terms in open-ended descriptions do not provide evidence for a strong version of the primacy-of-warmth effect, the warm-cold dimension nevertheless had a stronger influence on the overall valence of impressions than the polite-blunt dimension did.

Additional Analyses

The Additional Findings contain additional analyses that have no direct relevance to the primacy-of-warmth effect, but are related to Asch's hypotheses (1946) about the process underlying the above mentioned change in valence (pitting a change-in-meaning-effect, e.g., Hamilton & Zanna, 1974; Zanna & Hamilton, 1977, against a simple Halo-effect). They also contain analyses suggesting that almost all participants formed unified impressions in which they went beyond the information given, creating elaborate narratives about things that were not included in the original trait lists they had been exposed to (such as other traits, occupations, and gender). These exploratory analyses include modern-day data-analytical approaches to quantify some of the ideas that Asch had about his data, but was unable to test.

Discussion

In the present replication attempt, we aimed to critically examine the extent to which Asch's seminal "Forming impressions of personality" (1946) provides evidence for a primacy-of-warmth effect. Ample research suggests that warmth is often primary over competence in people's impressions of others (e.g., Fiske et al., 2007; Wojciszke, 2005), and Asch's classic warm-cold study often is one of the first and foremost references for this effect. In our replication of Asch's studies, we failed to find any evidence for primacy-of-warmth. Even in those conditions in which primacy-of-warmth should have been most pronounced (the classic warm-cold studies), participants indicated that *intelligent* was at least as influential a trait in forming their impressions. Moreover, participants' descriptions of the target person centered on competence at least as much as on warmth, and a substantial amount of participants did not refer to warmth in their descriptions at all.

Although it may seem as if the present replication attempt proves Asch (1946) wrong, note that Asch never claimed that warmth should be primary over competence. Centrality, in his view, was a property multiple traits could possess simultaneously, a property determined by "the whole system of relations between traits" (p. 284). In fact, Asch was upfront about the fact that warmth, though important, was not primary in his studies: "That the rankings are not higher is due to the fact that the lists contained *other* central traits." (p. 7, emphasis added). The present research coincides with Asch's idea that the centrality of warmth is highly context-dependent. The warm-cold dimension played an important (though not primary) role in determining participant's impressions when accompanied by traits such as *intelligent, skillful, industrious, determined, practical,* and *cautious* (Condition 1), but it became entirely peripheral in the context of other traits (Conditions 3 through 5). In line with Asch's predictions, the weight and meaning of warmth was not fixed, being relatively important in some contexts but not others.

It could be argued that Asch's studies (1946) were not optimally designed to capture a primacy-of-warmth effect. For example, his stimulus lists contained unequal amounts of warmth- and competence-related traits and the ranking measure presupposes that perceivers can reliably indicate which traits influenced their impressions (which may not be the case; Nisbett & Wilson, 1977). Many methodological advances have been made in the 68 years since the publication of Asch's seminal paper, and there now seems to be converging evidence for the central role warmth plays in shaping impressions of personality (e.g., from face perception research, Todorov, Said, Engell, & Oosterhof, 2008; research on morality, Wojciszke, 2005; and research on the perception of persons and groups, Fiske et al., 2007). In light of these recent findings, it may seem unimportant that Asch's data do not provide evidence for primacy-of-warmth, because, after all, the effect seems present in more modern studies. Still, knowing about the lack of primacy-of-warmth in Asch's studies is important. With over 2,750 references, Asch's work has been "the stuff of textbooks" (Fiske et al., 2007, p. 78), forming part of the foundation on which this later research has been built. By focusing on an incomplete and incorrect interpretation of Asch's work, researchers forfeit the chance to learn from the subtleties and complexities of his ideas and the intricacies of his thinking, and run the risk of overestimating the evidence there is for the primacy-of-warmth effect.

Asch's data (1946) suggest that, in the context of certain traits, warmth may not always be primary over competence. What are these conditions? Is warmth generally primary over competence in forming impressions, or is this effect limited to very specific circumstances? The present research suggests that Asch's data do not provide evidence for a primacy-of-warmth effect; if anything, competence seems more primary in his studies. Asch may not be the progenitor of primacy-of-warmth, but he did father the Gestalt-view on impression formation; A view that has lost its position at the forefront of science. For all its disadvantages, we believe this Gestalt-view (or other typological accounts of impression formation) may raise and answer questions that do not readily follow from dimensional models of impression formation. Asch's work, in our view, deserves a position at the forefront of science not because of its peripheral message about warmth, but because of its central message about the way in which people form impressions of personality, which constitutes the Gestalt of Asch's work.

Acknowledgments

We would like to thank the editor and two anonymous reviewers for their useful comments on earlier drafts of this article. This research was partially supported by a grant of the Center for Open Science. We report all data exclusions, manipulations, and measures, and how we determined our sample sizes. The authors declare no conflict-of-interest with the content of this article. Contribution statement: Designed research: S. N., I. H., D. W. Performed research: S. N., I. H. Analyzed data: S. N., O. L., I. H. Wrote paper: S. N., O. L., I. H., D. W., R. V. Materials, data, the preregistered proposal, and additional findings are available on the project page on the Open Science Framework http://osf.io/r6idy/

PREREGISTERED OPEN DATA OPEN MATERIALS

References

Abele, A. E., & Bruckmüller, S. (2011). The bigger one of the "Big Two"? Preferential processing of communal information. *Journal of Experimental Social Psychology, 47*, 935–948.

Abele, A. E., & Wojciszke, B. (2007). Agency and communion from the perspective of self versus others. *Journal of Personality and Social Psychology, 93*, 751–763.

Ahlering, R. F., & Parker, L. D. (1989). Need for cognition as moderator of the primacy effect. *Journal of Research in Personality, 23*, 313–317.

Anderson, C. A., & Sedikides, C. (1991). Thinking about people: Contributions of a typological alternative to associationistic and dimensional models of person perception. *Journal of Personality and Social Psychology, 60*, 203–217.

Anderson, N. H. (1968). Likableness ratings of 555 personality-trait words. *Journal of Personality and Social Psychology, 9*, 272–279.

Anderson, N. H. (1981). *Foundations of information integration theory*. New York, NY: Academic Press.

Anderson, N. H., & Barrios, A. A. (1961). Primacy effects in personality impression formation. *Journal of Abnormal and Social Psychology, 63*, 346–360.

Asch, S. E. (1946). Forming impressions of personality. *Journal of Abnormal and Social Psychology, 41*, 258–290.

Babad, E., Kaplowitz, H., & Darley, J. (1999). A "classic" revisited: Students' immediate and delayed evaluations of a warm/cold instructor. *Social Psychology of Education, 3*, 81–102.

Baron, R. A., & Byrne, D. E. (2004). *Social psychology*. Boston, MA: Allyn & Bacon.

Bem, D. J. (1987). Writing the empirical journal article. In M. P. Zanna & J. M. Darley (Eds.), *The complete academic: A practical guide for the beginning social scientist* (pp. 171–201). New York, NY: Random House.

Bertamini, M., & Munafo, M. R. (2012). Bite-size science and its undesired side-effects. *Perspectives on Psychological Science, 7*, 67–71.

Cohen, J. (1992). A power primer. *Quantitative Methods in Psychology, 112*, 155–159.

Cuddy, A. J., Fiske, S. T., & Glick, P. (2008). Warmth and competence as universal dimensions of social perception: The stereotype content model and the BIAS map. *Advances in Experimental Social Psychology, 40*, 61–149.

DeLamater, J. D., & Myers, D. J. (2010). *Social psychology*. Belmont, CA: Wadsworth.

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*, 175–191.

Fiske, S. T., Cuddy, A. J. C., & Glick, P. (2007). Universal dimensions of social cognition: Warmth and competence. *Trends in Cognitive Sciences, 11*, 77–83.

Fiske, S. T., Neuberg, S. L., Beattie, A. E., & Milberg, S. J. (1987). Category-based and attribute-based reactions to others: Some informational conditions of stereotyping and individuating processes. *Journal of Experimental Social Psychology, 23*, 399–427.

Franzoi, S. L. (2009). *Social psychology* (5th ed.). New York, NY: McGraw-Hill.

Gilbert, D. T. (1998). Ordinary personology. In D. T. Gilbert, S. T. Fiske, & L. Gardner (Eds.), *The Handbook of Social Psychology* (4th ed., Vol. 2, pp. 89–150). New York, NY: McGraw-Hill.

Gollin, E. S. (1954). Forming impressions of personality. *Journal of Personality, 23*, 65–76.

Grace, H. A., & Geenshields, C. M. (1960). Effect of closure on formation of impressions. *Psychological Reports, 6*, 94.

Hamilton, D. L., & Zanna, M. P. (1974). Context effects in impression formation: Changes in connotative meaning. *Journal of Personality and Social Psychology, 29*, 649–654.

Hendrick, C., & Constantini, A. V. (1970). Effects of varying trait inconsistency and response requirements on the primacy effect in impression formation. *Journal of Personality and Social Psychology, 18*, 158–164.

Hewstone, M., Stroebe, W., & Jonas, K. (2007). *Introduction to social psychology: A European perspective* (4th ed.). Hoboken, NJ: Wiley.

Hogg, M., & Vaughan, G. (2011). *Social psychology* (6th ed.). Upper Saddle River, NJ: Prentice Hall.

Jones, E. E., & Davis, K. E. (1965). From acts to dispositions: The attribution process in person perception. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 2, pp. 219–266). New York, NY: Academic Press.

Judd, C. M., James-Hawkins, L., Yzerbyt, V., & Kashima, Y. (2005). Fundamental dimensions of social judgment: Understanding the relations between judgments of competence and warmth. *Journal of Personality and Social Psychology, 89*, 899–913.

Kassin, S., Fein, S., & Markus, H. R. (2011). *Social psychology* (8th ed.). Stamford, CT: Cengage Learning.

Kelley, H. H. (1950). The warm-cold variable in impressions of persons. *Journal of Personality, 18*, 431–439.

Kervyn, N., Yzerbyt, V., & Judd, C. M. (2010). Compensation between warmth and competence: Antecedents and consequences of a negative relation between the two fundamental dimensions of social perception. *European Review of Social Psychology, 21*, 155–187.

Klein, R. A., Ratliff, K., Nosek, B. A., Vianello, M., Pilati, R., Devos, T., . . . Nier, J. A. (2014). Investigating variation in replicability: The ''many labs'' replication project. *Social Psychology, 45*, 142–152.

Luchins, A. S. (1948). Forming impressions of personality: A critique. *Journal of Abnormal and Social Psychology, 43*, 318–325.

Luchins, A. S., & Luchins, E. H. (1986). Primacy and recency effects with descriptions of moral and immoral behavior. *The Journal of General Psychology, 113*, 159–177.

McCarthy, R. J., & Skowronski, J. J. (2011). You're getting warmer: Level of construal affects the impact of central traits on impression formation. *Journal of Experimental Social Psychology, 47*, 1304–1307.

McKelvie, S. J. (1990). The Asch primacy effect: Robust but not infallible. *Journal of Social Behavior & Personality, 5*, 135–150.

Mensh, I. N., & Wishner, J. (1947). Asch on ''Forming impressions of personality'': Further evidence. *Journal of Personality, 16*, 188–191.

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review, 84*, 231–259.

Oppenheimer, D. M., Meyvis, T., & Davidenko, N. (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology, 45*, 867–872.

Peeters, G., & Czapinski, J. (1990). Positive-negative asymmetry in evaluations: The distinction between affective and informational negativity effects. *European Review of Social Psychology, 1*, 33–60.

Pennington, D. C. (2000). *Social cognition* (1st ed.). London, UK: Routledge.

Pettijohn, T. F. II., Pettijon, T. F., & McDermott, L. A. (2009). Active learning exercises for teaching classic research on impression formation in social psychology courses. *The Open Education Journal, 2*, 78–81.

Richetin, J., Durante, F., Mari, S., Perugini, M., & Volpato, C. (2012). Primacy of warmth versus competence: A motivated bias? *The Journal of Social Psychology, 152*, 417–435.

Rosenberg, S., Nelson, C., & Vivekananthan, P. S. (1968). A multidimensional approach to the structure of personality impressions. *Journal of Personality and Social Psychology, 9*, 283–294.

Sedikides, C., & Anderson, C. A. (1994). Causal perceptions of intertrait relations: The glue that holds person types together. *Personality and Social Psychology Bulletin, 20*, 294–302.

Semin, G. R. (1989). The contribution of linguistic factors to attribute inferences and semantic similarity judgments. *European Journal of Social Psychology, 19*, 85–100.

Singh, R., Onglacto, M. L. U., Sriram, N., & Tay, A. B. G. (1997). The warm-cold variable in impression formation: Evidence for the positive-negative asymmetry. *British Journal of Social Psychology, 36*, 457–477.

Stainton-Rogers, W. (2011). *Social psychology* (2nd ed.). Philadelphia, PA: Open University Press.

Taylor, S. E., Peplau, L. A., & Sears, D. O. (2006). *Social psychology* (12th ed.). Upper Saddle River, NJ: Prentice-Hall.

Todorov, A., Said, C. P., Engell, A. D., & Oosterhof, N. N. (2008). Understanding evaluation of faces on social dimensions. *Trends in Cognitive Sciences, 12*, 455–460.

Veness, T., & Brierley, D. W. (1963). Forming impressions of personality: Two experiments. *British Journal of Social and Clinical Psychology, 2*, 11–19.

Vonk, R. (1994). Trait inferences, impression formation, and person memory: Strategies in processing inconsistent information about persons. *European Review of Social Psychology, 5*, 111–149.

Wilson, T., Wiebe, J., & Hoffmann, P. (2005, October). Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing* (pp. 347–354). Morristown, NJ: Association for Computational Linguistics.

Wishner, J. (1960). Reanalysis of ''impressions of personality''. *Psychological Review, 67*, 96–112.

Wojciszke, B. (2005). Morality and competence in person- and self-perception. *European Review of Social Psychology, 16*, 155–188.

Wojciszke, B., Bazinska, R., & Jaworski, M. (1998). On the dominance of moral categories in impression formation. *Personality and Social Psychology Bulletin, 24*, 1245–1257.

Worchel, S., Cooper, J., Goethals, G., & Olson, J. (2000). *Social psychology*. Belmont, CA: Wadsworth.

Zanna, M. P., & Hamilton, D. L. (1977). Further evidence for meaning change in impression formation. *Journal of Experimental Social Psychology, 13*, 224–238.

Received February 14, 2013
Accepted November 27, 2013
Published online May 19, 2014

Sanne Nauts

Department of Social Psychology
Montessorilaan 3
6525 HR Nijmegen
The Netherlands
Tel. + 31 24 3615682
Fax +31 24 3612677
E-mail s.nauts@psych.ru.nl

Appendix

Checklist (Trait-Pair Choice Measure) as Used in Asch (1946)

"Choose the characteristic that is most in accordance with the view you have formed."

1. generous – ungenerous
2. shrewd – wise
3. unhappy – happy
4. irritable – good natured
5. humorous – humorless
6. sociable – unsociable
7. popular – unpopular
8. unreliable – reliable
9. important – insignificant
10. ruthless – humane
11. good looking – unattractive
12. persistent – unstable
13. frivolous – serious
14. restrained – talkative
15. self–centered – altruistic
16. imaginative – hard headed
17. strong – weak
18. dishonest – honest

Revisiting Schachter's Research on Rejection, Deviance, and Communication (1951)

Eric D. Wesselmann,[1] Kipling D. Williams,[2] John B. Pryor,[1] Fredrick A. Eichler,[1] Devin M. Gill,[1] and John D. Hogue[1]

[1]Illinois State University, Normal, IL, USA, [2]Purdue University, West Lafayette, IN, USA

Abstract. We conducted a replication of the original Schachter (1951) deviation-rejection study. Schachter's groundbreaking demonstration of the deviation-rejection link has captivated social psychologists for decades. The findings and paradigm were so compelling that the deviation-rejection link is often taken for granted and sometimes may be misrepresented (Berkowitz, 1971; Wahrman & Pugh, 1972). Because there have only been two direct replications, one of which by the original author, we believed it was important to revisit the original study. We replicated Schachter's main finding, albeit with a smaller effect size. One intriguing possibility is that we found somewhat weaker reactions to deviates because society may be becoming more tolerant of individuals who hold deviate opinions. We hope that our replication study will inspire other researchers to revisit the deviation-rejection link.

Keywords: deviance, rejection, Stanley Schachter, group locomotion, replication

Schachter's (1951) demonstration that groups reject opinion deviates from future interactions has captivated social psychologists for decades. In Schachter's research, discussion groups deliberated about how to deal with a juvenile delinquent named Johnny Rocco. Whereas most participants advocated leniency, a confederate argued articulately and logically for harsh discipline, standing his ground against all counterarguments. Two other confederates (the *mode* and the *slider*) either stuck with the consensus throughout the discussion or initially followed the *deviate* and later conformed to the group. Schachter found that over time participants attempted to achieve unanimity by increasing the amount of communication toward the deviate confederate. After it became clear to participants that he would not change his opinion, they stopped communicating with him entirely. At the end of the study, participants tended to suggest less prestigious roles for the deviate in subsequent discussion groups and typically did not choose him for future group meetings. Most citations to the original Schachter study only focused on participants' rejection of the deviate from future groups (Berkowitz, 1971).

Replication Value

Although not widely cited, there were several conceptual replications and extensions of the original findings (Arnold & Greenberg, 1980; Berkowitz & Howard, 1959; Earle, 1986; Israel, 1956; Kruglanski & Webster, 1991; Lauderdale, 1976; Lauderdale, Smith-Cunnien, Parker, & Inverarity, 1984; Levine & Ranelli, 1978; Levine, Saxe, & Harris, 1976; Sampson & Brandon, 1964), but there have only been two *direct replications* (Emerson, 1959; Schachter et al., 1954) using the original procedures and stimulus material. In the Schachter et al. (1954) follow-up, researchers conducted group discussion sessions in seven western-European countries. In this study, the researchers constructed groups of adolescent boys under the guise of an aviation club and asked the boys to discuss which type of model airplane would be best for their group (the deviate confederate chose the least attractive model). Across all countries sampled, the groups rejected the deviate compared to the other group members. The rejection tendency varied somewhat by country but the researchers interpreted these small differences cautiously (and noted that the paradigm instructions may not have translated well to some of the countries). Emerson (1959) found similar results to Schachter's first study using the original Johnny Rocco case with a new sample – high school boys. Emerson did note that his study found nuances different from Schachter's original study; the most important being the replicated deviation-rejection effect was statistically weaker than the original effect. Because there are only two direct replications, we conducted a replication of Schachter's original study and made minor alterations. The primary alteration is that we only manipulated the

DOI: 10.1027/1864-9335/a000180

behavior of the three confederates (i.e., the deviate, the slider, and the mode) and removed his other manipulations, which are not often discussed.

Method

We report all data exclusion, manipulations, and measures, and how we determined our sample size.

Participants

Schachter's original groups were composed of 5–7 male participants and three male confederates. We advertised the 75-min. study focusing on the dynamics of group discussion and decision-making and recruited participants via emails to a research opportunity listserve. Volunteers from this listserve represent a broad cross-section of university students from different majors. They were offered $20.00 Wal-Mart gift cards for participation. We organized group sessions so that there were 10 group members scheduled for each session (seven participants in addition to three confederates). We conducted all experimental sessions regardless of how many participants showed up.

Schachter treated each group as a unit of analysis for his primary rejection DVs. A meta-analysis by Tata et al. (1996) reported an overall deviation-rejection effect across studies as d = 1.49 (they reported Schachter's effect size as d = 1.84). Using a power analysis, we set the anticipated effect size at d = 1.49, power level of .95, and an alpha of .05. The analysis suggests a minimal total sample size of 26 groups for a two-tailed hypothesis; because the primary comparisons are within-subjects (deviate vs. mode) we divided this number in half (Cohen, 1988). Thus, we set our minimal target number at 13 group sessions.

Eighty-two men participated and comprised a total of 17 groups, which ranged between 3 and 7 participants (mode = 5). During debriefing, two participants requested that their data be discarded. The final sample consisted of 80 men, of which 73.7% were Caucasian/White, 8.7% African-American/Black, 3.7% Asian/Indian, 1.2% Hispanic, 6.2% Multiracial, and 6.2% unreported.[1] Participants were on average 21.21 years of age (SD = 2.83). No participants indicated having prior knowledge of the Schachter (1951) study.

Procedure

The experimenter instructed participants that the experiment would take approximately 75-min. As each group member (both participants and the confederates) entered the discussion room they completed informed consent documents. The experimenter had name tags prepared for each group member and placed them in assigned seating in a semi-circle. This was to ensure that the confederates would all be sitting in the same location so that they could hear each participant's opinions before they had to indicate their vote. The experimenter then gave a brief introduction, asked group members to introduce themselves, and then explained that he was trying to obtain more funds to give them all the opportunity to come back for future discussions but that he would describe that more at the end of the study. The experimenter gave this information to create the illusion that they may have future interactions with each other (Schachter's original participants thought they were signing up a group that would have reoccurring meetings).

The experimenter then gave participants a case study of a juvenile delinquent named *Johnny Rocco*. He explained that research has found this material was useful for facilitating discussion in new groups such as theirs. The experimenter gave the group 5 min to read the case study. Then, participants orally indicated their opinions of how Johnny should be treated. They did so using the Love/Punishment scale, a 7-point rating scale ranging from *giving Johnny nothing but love, kindness, and friendship* (1) to *giving Johnny nothing but a severe disciplinary environment by punishing him* (7). Schachter's sample endorsed lenient attitudes toward the Johnny Rocco case (Schachter reported most participants chose positions 2–4; see online material A). In our sample, participants' final choices ranged from 1 to 6 (M = 3.47, SD = .64). No one chose 7 at any point during the group discussions. The confederates all answered last so that they could calibrate their opinions based on the rest of the group.

The primary discussion was 45 min. The experimenter took a census approximately 20 min into the discussion and again at the end of the discussion. The experimenter then handed out questionnaires that included the dependent variables, and mentioned that he was petitioning the psychology department to give him more funding possibly to run future group sessions so they may be given the opportunity to meet again. If that occurred, the researchers would use their questionnaire answers to guide decisions about any future meetings. The experimenter gave these instructions to increase the psychological realism that the participants may interact with each other in the future (and thus their ratings of the other group members may have a meaningful impact on future interactions). After the participants completed these measures, the experimenter then debriefed them about the purpose of the study and gave them their gift cards.

The Confederates

We trained three white male confederates; the race and gender of the confederates were held consistent to avoid any confounding variables related to minority racial or gender status with the deviate group role. We trained the

[1] We neglected to collect demographic information during the first three groups. We contacted those participants via email afterwards to obtain this information and for any participants who did not respond we coded as unreported.

confederates to perform all three roles: *mode*, *deviate*, and *slider*. We counterbalanced confederates' role assignments using a Latin square design.

In keeping with Schachter's original procedures, each confederate had to speak once every 5 min. If during any 5-min interval no one addressed him, he initiated a communication. When possible, all communications (whether initiated or in answer to one of the other participants) was a rephrasing of the position he was maintaining at the time.

The *mode* took the side of the participants from the beginning, choosing an option consistently on the lenient side of the Love/Punishment scale (assuming that was the modal opinion at the beginning). If the group's consensus shifted during the discussion, the mode shifted as well. If ratings scores were equally divided, the mode would make a decision based on his best judgment.

The *deviate* chose #7 on the Love/Punishment scale, which was the most extreme position on the Punishment side, at the beginning and maintained it throughout the discussion. We trained the deviate to resist attempts to minimize differences between him and members at the other position. When it was impossible for the deviate simply to rephrase his position, we allowed him variations on two standard arguments: (1) despite the fact that Johnny was shown love and affection by one adult, he went back to stealing, and (2) it could not be said that discipline would not work because it had not consistently been applied to Johnny.

The *slider* chose #7 at the beginning of the session but gradually was persuaded so that at the end of the discussion he was at the modal position with the rest of the group. Schachter (1950, p. 18) stated that this member allowed "himself to be influenced step-by-step to the modal position" but did not elucidate how this member would vote at the midpoint vote. We trained the slider to choose an option between #7 and the group's current modal position. In the event that there was not an even halfway point between the deviate rating and majority rating, the slider chose the rating based on his best judgment and at the end of the discussion his rating would always match that of the modal rating.

Neither Schachter's (1951) article nor his original dissertation (1950) contained a confederate script but instead only the basic instructions given to each confederate. We developed confederate scripts based upon these instructions, representative arguments given in Schachter's reenactment in the documentary *The Social Animal* (Mayer & Norris, 1970) and other relevant topics (e.g., news or media events, book, or film examples). We instructed confederates to use these scripts as guidelines, due to the dynamic nature of group discussions; they were allowed to adapt the scripted arguments if necessary to account for each group's idiosyncrasies.

Dependent Variables

Communication During Discussion

We surreptitiously recorded the experimental sessions in order to code the number of communications to each confederate. The moderator also took notes on the discussions as a backup in case of equipment failure (see online material B).

Committee Nomination Measure

Participants assigned each group member to one of three committees for any future group meetings. They were instructed not to rank themselves and to only assign each group member to *one* committee.[2] These committees were (from least to most desirable) the Correspondence committee, the Steering committee, and the Executive committee (see online material C).

Sociometric Test

Participants ranked their preferences of the other group members for future group interactions. They were instructed to write the rank number next to each group member's name (#1-preferred most, #2, preferred next, etc.) and write *no* next to any group member that they absolutely did not want to work with. Schachter reasoned that higher rankings meant greater rejection.

Results and Discussion

Confirmatory Analyses

Communication Pattern[3]

Schachter's analysis identified the *peak communication* time intervals for each confederate and conducted a binary test to assess if there was more or less communication from the peak communication interval to the final time interval. His analysis neglected the mean differences over the course of the group discussion and inflated the probability of

[2] Unfortunately, 19 participants ignored the directions and nominated one or more of the confederates into multiple committees. We excluded their data from this analysis as it violated the χ^2 assumption of independence; specifically, each participant receives one vote per confederate to one committee.

[3] We did not ask participants to write their names on their questionnaire packets to increase the perceived anonymity of their answers. Thus, we cannot match participants' questionnaire answers with their communications during the group discussion. This would have been the only way to assess how participants' individual ratings of the confederates corresponded to their communications to them during the discussion.

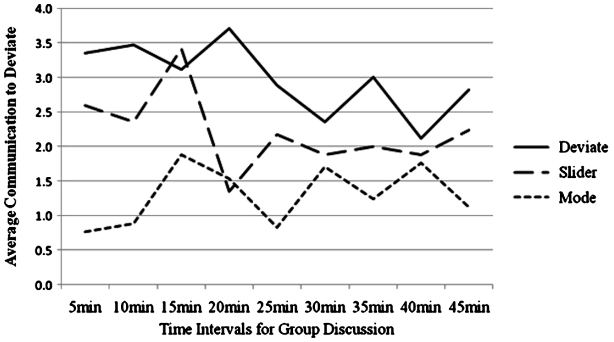

Figure 1. $N = 17$ groups. All communication is averaged on the group level and represented over the course of the group discussions. After 20 min, a second "vote" was taken amongst all members to indicate their current position on the love/punishment scale.

Type I error (i.e., conducting a separate test for each role). To alleviate these issues, we conducted a 6 (first three 5-min intervals and last three 5-min intervals) × 3 (Deviate, Slider, Mode) within-subjects factorial ANOVA.[4] We treated each group as the unit of analysis and averaged the number of communications to each confederate in each 5-min interval and both factors as within-participants variables.

We did not find a statistically significant main effect for change over time in overall communication to the confederates, $F(5, 80) = 1.23$, $p = .30$, $\eta_p^2 = .07$, but we did find a significant main effect for the groups' differential communication between the confederates, $F(2, 32) = 20.83$, $p < .01$, $\eta_p^2 = .57$, with the most communications toward the deviate (Table 1). We did not observe a significant interaction between communication to the different confederates and the point of the conversation (first vs. second half), $F(10, 160) = .99$, $p = .45$, $\eta_p^2 = .06$. It is important to note that Schachter did not test an interaction between confederates and communications over time but instead calculated binary tests for each individual confederate. The pattern of communication to each confederate in our data does still replicate Schachter's findings; communications to the deviate were the most during the first half of the discussion and then reduced over time; communications to the mode stayed consistent throughout the discussion and communications to the slider were higher at the beginning and then became parallel with the mode after the 20-min vote when he conformed to the group (Figure 1).

Committee Nomination Measure

Schachter originally conducted multiple *t*-tests for each confederate's likelihood of being assigned to each role. This method increases the likelihood of Type I error. We chose instead to conduct a chi-square test for independence between the confederate role and committee assignment to control for error and account for the three-confederate design. We found no association between confederate roles and their assigned committees, $\chi^2(4) = .79$, $p = .94$ (Table 2). These results do not replicate Schachter's original findings.

Sociometric Test

Schachter originally created equivalence scores for each rank assigned to confederates based on each group's size. We focused on participants' ranking of each confederate relative to each other (1 = most preferred to 3 = least preferred) because we had wider variability in group sizes than Schachter did.[5] We conducted a Friedman's related samples test (the nonparametric equivalent of a repeated-measures ANOVA) to examine participants' vote preferences for each confederate. There was an overall significant difference among confederate rankings, $\chi^2(2) = 14.74$, $p < .01$. Descriptively, participants ranked the slider ($M_{\text{Rank}} = 1.74$) as most preferred, followed by the mode ($M_{\text{Rank}} = 1.91$), and finally the deviate ($M_{\text{Rank}} = 2.34$). We conducted a post hoc Wilcoxon Signed-Rank test (with Bonferonni corrections) to determine significant differences between participants' confederate rankings; participants ranked both the slider and mode as more preferred than the deviate ($T = 806.50$, $r = -.28$ and $T = 1{,}044.50$, $r = -.19$, respectively), but there was no difference between the mode and the slider's rankings ($T = 1{,}280.50$, $r = -.10$). In addition to rankings, some participants actively voted "no" for some of the confederates. Twelve participants voted "no" for the deviate, two participants for the slider, and no participants for the mode. These results replicate Schachter's original findings.

Tata et al. (1996) reported a meta-analytic effect size for the general deviation-rejection link as $d = 1.49$ and Schachter's effect size as $d = 1.84$ (in both cases, comparing rankings of the deviate to the mode). If we convert our effect size for the deviate-mode comparison from *r* to *d*, the effect is considerably smaller than either Schachter's

[4] Due to small sample size, the full 9 (5-min time interval) × 3 (confederate) within-participants analysis did not converge. As we are conceptually emulating Schachter's analyses that assessed the difference in communication from the beginning to end of the discussion, we trimmed the middle three 5-min intervals.

[5] The instructions specified that participants rank *each* group member and not to rank themselves. Unfortunately, some participants ranked themselves, which influences the interpretation of each confederate's numeric ranking; further, some participants chose not to rank all of the confederates. We excluded three participants' data from these analyses because they did not rank all three confederates. Other participants did not assign numerical values to each member but instead appeared to rank them hierarchically on the list; we inferred this because they did not simply list each group member in the assigned seating order. We excluded one participants' data because he did list all members of the group in the order they sat during the discussion. We also conducted the analyses with only participants who ranked the confederates numerically ($N = 50$); the overall pattern replicated our findings with the larger sample, $\chi^2(2) = 7.96$, $p < .05$. Post hoc analyses revealed that participants ranked the slider significantly more preferable ($M_{\text{Rank}} = 1.74$) than the deviate ($M_{\text{Rank}} = 2.30$), $T = 363.50$, $p < .01$, $r = -.27$. There was, however, no difference in the ranking between the mode ($M_{\text{Rank}} = 1.96$) and deviate ($T = 495.00$, $p > .10$, $r = -.14$) or the mode and the slider ($T = 520.00$, $p > .20$, $r = -.12$).

Table 1. Means and standard deviations from beginning to end of conversation

	Beginning of discussion (min)			End of discussion (min)		
Confederate	5	10	15	35	40	45
Deviate	3.35 (1.80)	3.47 (3.16)	3.12 (2.29)	3.00 (2.72)	2.12 (2.34)	2.82 (2.19)
Slider	2.59 (1.66)	2.35 (1.84)	3.41 (3.22)	2.00 (2.72)	1.88 (2.20)	2.24 (1.95)
Mode	0.77 (0.83)	0.88 (1.32)	1.88 (3.22)	1.24 (1.09)	1.77 (1.99)	1.12 (1.32)

Notes. N = 17 groups. *SD* are represented in parentheses.

Table 2. Committee by confederate role placement frequencies

	Committee			
Role	Executive	Steering	Correspondence	Total
Mode	25	19	17	61
Slider	26	22	13	61
Deviate	26	20	15	61
Total	77	61	45	183

Note. For this variable, the vote is the unit of analysis; each of the 61 participants had three votes.

original effect or the meta-analytic effect (|*d*| = .39, Rosenthal, 1991). To conclude, our current study replicated the primary finding in Schachter's original deviation-rejection (the vote outcome), albeit with a smaller effect size. Of his other two dependent variables, we failed to replicate the committee assignment variable, but our data trend supported his original findings for communication patterns. One divergence from Schachter's original procedures in our study is that Schachter's participants believed they were guaranteed future sessions, while our participants only believed future sessions a possibility. This difference could have made the deviate's threat to group harmony less salient in our study. However, similar to Schachter's participants the majority of our participants also expressed a general perception of group cohesion and a desire to continue meeting (see online material D). One intriguing possibility is that our study found somewhat weaker reactions to deviates because society is becoming more tolerant of individuals who hold deviate opinions; only future research can answer this question. We hope that our replication study will inspire other researchers to revisit the deviation-rejection link.

Acknowledgments

This project was funded by the Center for Open Science. The authors declare no conflict-of-interest with the content of this article. We express our gratitude to Matthew Hesson-McInnis for his invaluable expertise as a statistical consultant, as well the graduate and undergraduate research assistants who participated in our confederate training sessions. Designed research: E. D. W., K. D. W., J. B. P., F. A. E., D. G., J. D. H.; Performed research: E. D. W., F. A. E., D. G., J. D. H.; Analyzed data: E. D. W., F. A. E., D. G., J. D. H.; Wrote paper: E. D. W., K. D. W., J. B. P., F. A. E., D. G., J. D. H. All materials, data, and the preregistered design are available at http://osf.io/s92i4/.

PREREGISTERED OPEN DATA OPEN MATERIALS

References

Arnold, D. W., & Greenberg, C. I. (1980). Deviate rejection within differentially manned groups. *Social Psychology Quarterly, 43*, 419–424.

Berkowitz, L. (1971). Reporting an experiment: A case study in leveling, sharpening, and assimilation. *Journal of Experimental Social Psychology, 7*, 237–243.

Berkowitz, L., & Howard, R. C. (1959). Reactions to opinion deviates as affected by affiliation need (n) and group member interdependence. *Sociometry, 22*, 81–91.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Earle, W. B. (1986). The social context of social comparison: Reality versus reassurance? *Personality and Social Psychology Bulletin, 12*, 159–168.

Emerson, R. M. (1959). Deviation and rejection: An experimental replication. *American Sociological Review, 19*, 688–693.

Israel, J. (1956). *Self-evaluation and rejection in groups: Three experimental studies and a conceptual outline*. Uppsala, Sweden: Almqvist & Wiksell.

Kruglanski, A. W., & Webster, D. M. (1991). Group members' reactions to opinion deviates and conformists at varying degrees of proximity to decision deadline and of environmental noise. *Journal of Personality and Social Psychology, 61*, 212–225.

Lauderdale, P. (1976). Deviance and moral boundaries. *American Sociological Review, 41*, 660–676.

Lauderdale, P., Smith-Cunnien, P., Parker, J., & Inverarity, J. (1984). External threat and the definition of deviance. *Journal of Personality and Social Psychology, 46*, 1058–1068.

Levine, J. M., & Ranelli, C. J. (1978). Majority reaction to shifting and stable attitudinal deviates. *European Journal of Social Psychology, 8*, 55–70.

Levine, J. M., Saxe, L., & Harris, H. J. (1976). Reaction to attitudinal deviance: Impact of deviate's direction and distance of movement. *Sociometery, 39*, 97–107.

Mayer, H. (Producer/Director) & Norris, M. (Writer). (1970). *The social animal* [Motion picture]. Bloomington, IN: Indiana University Audio-Visual Center.

Rosenthal, R. (1991). *Meta-analytic procedures for social research*. Newbury Park, CA: Sage.

Sampson, E. E., & Brandon, A. C. (1964). The effects of role and opinion deviation on small group behavior. *Sociometry, 27*, 261–281.

Schachter, S. (1950). *Deviation, rejection and communication*, (Unpublished doctoral dissertation). University of Michigan, Ann Arbor, MI.

Schachter, S. (1951). Deviation, rejection and communication. *Journal of Abnormal and Social Psychology, 46*, 190–207.

Schachter, S., Nuttin, J., de Monchaux, C., Maucorps, P. H., Osmer, D., Duijker, H., . . . Israel, J. (1954). Cross-cultural experiments on threat and rejection. *Human Relations, 7*, 403–439.

Tata, J., Anthony, T., Lin, H., Newman, B., Tang, S., Millson, M., & Sivakumar, K. (1996). Proportionate group size and rejection of the deviate: A meta-analytic integration. *Journal of Social Behavior and Personality, 11*, 739–752.

Wahrman, R., & Pugh, M. D. (1972). Competence and conformity: Another look at Hollander's Study. *Sociometry, 35*, 376–386.

Received February 28, 2013
Accepted November 26, 2013
Published online May 19, 2014

Eric D. Wesselmann

Department of Psychology
Illinois State University
Campus Box 4620
Normal, IL 61790-4620
USA
E-mail edwesse@ilstu.edu

Replication

Revisiting the Romeo and Juliet Effect (Driscoll, Davis, & Lipetz, 1972)

Reexamining the Links Between Social Network Opinions and Romantic Relationship Outcomes

H. Colleen Sinclair,[1] Kristina B. Hood,[1] and Brittany L. Wright[2]

[1]Mississippi State University, MS, USA, [2]University of Texas at Austin, TX, USA

Abstract. We conducted a replication and extension of Driscoll, Davis, and Lipetz's (1972) classic longitudinal survey of the Romeo and Juliet effect, wherein they found that increases in parental interference were linked to increases in love and commitment. Using the original measures, 396 participants were followed over a 3–4 month period wherein they reported love, commitment, trust, and criticism for their partners as well as levels of perceived interference from friends and family. Participants also completed contemporary, validated measures of the same constructs similar to those often implemented in studies of social network opinion. Repeating the analyses employed by Driscoll and colleagues, we could not find evidence for the Romeo and Juliet effect. Rather, consistent with the social network effect (Felmlee, 2001), participants reporting higher levels of interference or lower levels of approval reported poorer relationship quality regardless of outcome measured. This effect was likewise evident in a meta-analysis.

Keywords: social networks, social support, love, romantic relationships, parent child relations

The "Romeo and Juliet effect" was coined by Driscoll, Davis, and Lipetz (1972) when they discovered that couples who reported an increase in parental interference in their romantic relationship also evidenced an increase in love over the same 6-month period. Since the original study, replications of the effect have been elusive. Few studies find support for anything approximating the effect (Felmlee, 2001; Parks, Stan, & Eggert, 1983; Sprecher, 2011). Instead, most find the "social network effect" (Felmlee, 2001) whereby disapproval from one's social network – whether family or friends – leads to *declines* in romantic relationship quality (see Allan, 2006; Parks, 2007).

Nonetheless, much like the story, the lore of the Romeo and Juliet effect persists, and the finding continues to be cited in popular culture, blogs, textbooks, and articles (e.g., DeWall, Maner, Deckman, & Rouby, 2011; Fisher, 2004; Miller, 2011). Reis (2011), in his review of the history of relationships research, touted the effect as "ever-popular" (p. 219). A limitation of the counterevidence for the Romeo and Juliet effect is the fact that none of the follow-up studies used the original scales.

One essential difference between the original study and subsequent follow ups has been the operationalization of social network opinions. Driscoll and colleagues (1972) focused on whether couple members had communicated to one another that they felt the other's parents were *interfering in their relationship*. In contrast, with few exceptions (e.g., Johnson & Milardo, 1984), the majority of subsequent studies examined network approval on a single continuum, such that if a party scored low on approval indices then they were considered disapproving of the relationship. Although interference is a behavioral manifestation of disapproval – a particularly active, direct form of disapproval – it may not be fair to compare studies measuring perceived approval, or the lack thereof, as equivalent to studies assessing interference.

Further, in the original study, both members of the couple completed interference scales about each respective set of parents only. Subsequent studies have assessed perceived network opinions from various sources, including perceived global assessments of network opinions (combining all parents, friends, society, etc.), asking about one source (i.e., parents, friends, or siblings), or asking separate questions for multiple sources. Assessing multiple network sources may have important implications for the Romeo and Juliet effect. Felmlee (2001) found that parental disapproval led to a decreased likelihood of break up but only when friends approved. In fact, some research suggests that friend opinions may carry more weight than that of parents (Etcheverry & Agnew, 2004), at least in Western cultures (MacDonald & Jessica, 2006). Likewise, other research has suggested that approval from other sources within the

DOI: 10.1027/1864-9335/a000181

network can serve as a buffer to the presence of disapproving or interfering sources (Wright & Sinclair, 2012). Thus, the Romeo and Juliet effect, if it exists, might be less about the disapproval of their family and more about the approval of their allies. Even Romeo and Juliet had the friar and the nanny.

The current study addresses the gaps between the Driscoll and colleagues (1972) study and subsequent studies. First, we assess both parent and friend opinion. We also included the scales from Driscoll et al., which have not been used since, and administered more recent, validated measures for the same constructs. Including both original and contemporary measures allows us to address whether failures to replicate the effect were due to measurement differences.

The Romeo and Juliet hypothesis asserts that interference is linked to greater love and commitment for one's partner. Therefore, using the original scales, if the effect exists we would expect to find this same relationship in a contemporary sample. It could be the case that this effect is in fact limited to the effects of *increases* in interference from *parents*. In which case it is not exactly competing with the social network hypothesis, but rather occurs in certain circumstances. Whereas when it comes to the contemporary assessment of network approval, we anticipate replication of the social network effect such that the perceived approval from friends and family (and increases in that approval) is linked to greater love and commitment.

Method

Participants

We recruited participants from Amazon's Mechanical Turk (mTurk) crowdsourcing database. As many mTurk workers did not meet the basic criteria of being in a romantic relationship, we ended up with 1,602 participants attempting the screening survey. Due to time limitations, we were only able to keep the survey open for a month, at the end of that month we had 976 who met the study criteria of (1) were in a relationship lasting more than 6 weeks, (2) were not dating another person, and (3) their friends, partner's friends, parents, and partner's parents were currently aware of their relationship, and (4) they completed the survey responsibly. Additional eliminations for invalid email addresses, nonconsent for recontact, or closed mTurk accounts resulted in 718 eligible participants completing wave one in July/August who were then contacted in November/December for wave two. Of the 458 participants who returned, 396 provided usable data. Participants were paid $1 for the wave 1 and $2 for the wave 2 (and were entered into a gift card drawing).

Married participants made up 48.5% of the sample, the remainder were in various stages of dating. Of participants returning for wave 2, only 19 were broken up. Approximately half (50.8%) of the relationships had a duration of 4 years or less. Ages ranged from 18 to 70, with an average of 31.58 (*SD* = 9.78). The sample was predominantly female (70.7%) and Caucasian (76.3%, 5.1% Asian or Pacific Islander, 0.5% Hispanic or Latino, 4.5% African-American, 0.5% American Indian, 12.9% Multiracial, or multiethnic). There were some demographic (e.g., age) differences between those who returned and those who did not, as well as some differences on variables of interest. However, all differences between wave 1 and 2 samples had very small effect sizes. We report all data exclusions, sample differences, measures, and how we determined our sample sizes in supplemental materials. All materials, data, and the preregistered design are available at: https://osf.io/6wxgf/.

Materials

Driscoll and Colleagues Measures

Participants first completed all of the original scales included in the Driscoll and colleagues study. Item order was randomized within scale. Unless otherwise noted, participants responded on a Likert scale of 1 (= *not at all*) to 6 (= *extremely*).

Social Network Interference

Six items assessed interference for each social network source (own friends, own parents, partner's parents, partner's friends). For example, participants were asked "How often has your romantic partner communicated to you that *your parents* are a bad influence?" They were asked the same questions about their communication with their partner "How often have you communicated to your romantic partner that *his/her friends* interfere?" Participants responded on a 5-point Likert scale where 1 = *not at all* and 5 = *all the time*. As in the original study, scores for both sets of parents were combined to an overall index of parental interference (α = .90 for both waves). Asking the questions about friends was an extension of the previous study, but again both sets were combined and the 12 items had a reliability of α = .91 in both waves.

Love

Driscoll and colleagues used four items to assess love (e.g., "how much do you love your partner"). Reliability was α = .84 in wave 1 and .89 in wave 2.

Commitment

A single item was originally used to assess commitment, "How committed are you to your marriage (or to marrying your current partner)?"

Additional Measures

Driscoll and colleagues also included five items to assess *trust* of the partner (e.g., "how dependable is your partner") and six items to assess "criticalness" (e.g., "how critical are you of your partner"). Reliability was .89 in wave 1 and .92 in wave 2 for trust. Reliability was .75 for criticism at both waves.

Contemporary Measures

After finishing all of the original scales, participants completed the contemporary survey. Item order on each index was randomized. All relationship quality indices were responded to on a 9-point Likert scale with 1 = *not at all true* and 9 = *definitely.*

Social Network Opinion Scale

Eight items were compiled from an array of studies examining social network opinions. Four items assessed approval and four reverse-scored items assessed disapproval were administered for each of the four sources (i.e., friends, parents, partner's parents, partner's friends) and used the same 5-point Likert response format as the original interference scale. Parents and partner's parents items were combined (α = .91 in wave 1 and .93 in wave 2) as were friends and partner's friends items (α = .93 in wave 1 and .95 in wave 2).

Love

The 15-item Hatfield and Sprecher (1986) Passionate Love Scale was administered. Reliability was α = .89 in wave 1 and .92 in wave 2.

Commitment

We used the 10-item (five reversed) Lund (1985) Commitment scale. Reliability was α = .91 in wave 1 and .92 in wave 2.

Additional Measures

The 2-item Perceived Criticism Measure (Hooley & Teasdale, 1989) was already part of Driscoll and colleagues' criticism measure, so we kept Driscoll et al.'s scale. However, for trust, the 17-item (four reversed) Rempel et al. (1985) scale was used. Reliability was α = .94 in wave 1 and .95 in wave 2.

Scores were averaged across all measures such that higher scores indicate higher levels of the respective factor. Also, when difference scores were computed, they were computed such that higher scores indicated an increase in that component.

Known Differences Between Current and Original Study

Unlike the original study, the present study is not a part of an on-going longitudinal marital intervention initiative with couples.[1] Thus, the primary difference between the two studies is the sample. The online administration format and timeframe also differed in our study. Due to time constraints, we cut the follow-up window in half from what originally was used based on the recommendations of the project reviewers.

Results

Driscoll and colleagues conducted correlational analyses between the variables at Time 1 and Time 2 plus correlations between difference scores in parental interference and differences in each the relationship quality indices. Following Driscoll et al., correlations were computed for the total sample as well as for dating and married samples separately (Table 1). Table 2 includes the correlations between the difference scores. Note, Time 2 parental and friend interference were correlated at .51 overall, .54 for daters, and .53 for those married, all p's < .001.

Contrary to the Romeo and Juliet effect, higher interference was consistently linked with lower relationship quality (e.g., lower love, trust, commitment; higher criticism). Increases in parental interference were not related to increases in love and commitment over time. However, as was found by Driscoll and colleagues, increases in interference were linked to reductions in trust and increases in criticism among daters.

We repeated these analyses for the contemporary measures. The results are presented in Tables 3 and 4. Note, Time 2 parental and friend approval were correlated at .69 overall, .73 for daters, and .64 for those married, p < .001. Across analyses, measures, samples, and sources, higher levels of social network approval were linked to higher relationship quality. Further increases in approval, particularly friend approval, were linked to increases in relationship quality.

Measures of parental interference negatively correlated with measures of parental approval at −.46 to −.57 across

[1] Although we have individual instead of dyadic data, it is important to note that the original authors found no differences by couple member regarding the existence of the effect. In fact, the couple scores were simply combined into a single index of parental interference based on this rationale (and evidence of intracouple homogeneity on indices). Accordingly, as the dyadic nature of the data was not integral to the effect being found initially we do not think the lack of dyadic data threatens the comparison.

Table 1. Correlations between original social network interference measures and relationship quality indices across Time 1 and Time 2

		Time 1		Time 2	
		Parental interference *M* = 1.62, *SD* = 0.70	Friend interference *M* = 1.47, *SD* = 0.57	Parental interference *M* = 1.49, *SD* = 0.65	Friend interference *M* = 1.33, *SD* = 0.58
Parental interference (T1)		–		.70**	.32**
	Dating			.68**	.31**
	Married			.73**	.41**
Friend interference (T1)		.43**	–	.30**	.57**
	Dating	.40**		.35**	.61**
	Married	.47**		.24**	.54**
Love (T1, *M* = 5.09, *SD* = 0.83)		*−.08*	−.19**	−.13*	−.22**
	Dating	*−.04*	−.27**	*−.07*	−.21*
	Married	−.15*	*−.08*	−.19*	−.17*
Commitment (T1, *M* = 5.16, *SD* = 1.18)		*−.02*	−.16**	*−.06*	*−.19***
	Dating	*.05*	−.16*	*.04*	−.10
	Married	−.22**	*−.13*	−.28**	*−.25***
Trust (T1, *M* = 4.97, *SD* = 0.92)		−.21**	−.27**	−.21**	−.19*
	Dating	−.15*	−.33**	*−.10*	−.18*
	Married	−.27**	−.20*	−.32**	−.24**
Criticism (T1, *M* = 2.25, *SD* = 0.79)		.35**	.38**	.29**	.25**
	Dating	.30**	.38**	.26**	.28**
	Married	.40**	.40**	.31**	.27**
Love (T2, *M* = 4.99, *SD* = 0.98)		*−.06*	−.12*	−.13*	−.20**
	Dating	*.04*	−.14*	*−.05*	−.19*
	Married	−.21*	*−.07*	−.21*	−.15*
Commitment (T2, *M* = 5.03, *SD* = 1.35)		*−.02*	−.19**	−.10*	−.26**
	Dating	*.05*	−.17*	*−.07*	−.19*
	Married	−.22*	−.20*	−.21*	−.24**
Trust (T2, *M* = 4.83, *SD* = 1.07)		−.15*	−.19**	−.24**	−.23**
	Dating	*−.05*	−.21**	−.18*	−.26**
	Married	−.26**	−.15*	−.31**	−.20*
Criticism (T2, *M* = 2.22, *SD* = 0.83)		.25**	.31**	.33**	.35**
	Dating	.15*	.33**	.32**	.46**
	Married	.36**	.27**	.33**	.18*

Notes. *$p < .05$, **$p < .001$, values in italics $p > .05$.

Table 2. Correlations between difference scores in interference and relationship quality using original measures

		Increases in parental interference	Increases in friend interference
Increase in love		*−.05*	*−.07*
	Dating	*−.13*	*−.11*
	Married	*.07*	*.00*
Increase in commitment		*−.09*	*−.06*
	Dating	−.19*	*−.11*
	Married	*.13*	*.10*
Increase in trust		−.18**	−.16*
	Dating	−.28**	−.23**
	Married	*−.01*	*−.01*
Increase in criticism		.22**	.21**
	Dating	.28**	.25**
	Married	*.12*	*.10*

Notes. *$p < .05$, **$p < .001$, values in italics $p > .05$.

the different administrations. Correlations between measures of friend interference and friend approval ranged from −.52 to −.78.

Meta-Analysis

To examine consistency of effects, a meta-analysis of 22 studies, including the present data, was conducted to assess how peer and family networks correlate with romantic relationship outcomes (see Appendix). To be included, studies needed to employ assessments of social network opinions and commitment or love. We searched PsycINFO and Scopus using relevant terms and used citation searches to find articles that referenced the Driscoll et al. study or other highly cited social network studies (e.g., Felmlee, 2001; Sprecher & Felmlee, 1992) in July and August of 2013. Reference lists from published review papers and meta-analyses were searched for eligible studies. We also reviewed archived abstracts from past SPSP conference programs and submitted a call for papers on various listservs for unpublished data and manuscripts under review or in press.

The 22 eligible studies included 17 published articles, one dissertation, data from three unpublished datasets, and data described in the current article. For all studies, we calculated the Fisher's Z statistic which was then converted to the Hedges' g effect sizes for measures of relationship commitment and love. For studies with more than one outcome measure for a given construct, we calculated the effect size for each measure first, then calculated an average Hedges' g effect for each study outcome. All meta-analyses were conducted using Comprehensive Meta-Analysis (v.2) software developed by Biostat (Borenstein et al., 2005).

We examined the extent to which network type: friend, family, or combined social network were associated with love (Table 5, Figure 1) and commitment (Table 6, Figure 2) by calculating a weighted mean effect size for each network type. Random effects models and the Q statistic are reported for each network type. Eleven studies provided data on the overall effect of network approval on love. Network approval was moderately and positively associated with love 0.49 ($p < .001$, 95% confidence interval [CI] = 0.26–0.72). Correspondingly, friend approval 0.40 ($p < .05$, 95% CI = 0.09–0.71), family approval 0.32 ($p < .01$, 95% CI = 0.10–0.55), and combined network approval 1.02 ($p < .01$, 95% CI = 0.39–1.65) were positively associated with love. Those who reported approval from their social networks reported higher ratings of love for their romantic partner.

Similar results were found with commitment. Network approval was moderately and positively associated with commitment 0.62 ($p < .001$, 95% CI = 0.50–0.74). Likewise, friend approval 0.70 ($p < .001$, 95% CI = 0.54–0.86), family approval 0.56 ($p < .001$, 95% CI = 0.39–0.84), and combined network approval 0.63 ($p < .001$, 95% CI = 0.43–0.84) were all positively associated with commitment. Those who reported approval from their social networks reported higher levels of commitment.

Discussion

Using the scales employed by Driscoll and colleagues as well as contemporary measures we found no evidence for the Romeo and Juliet effect. Instead, with both the original and contemporary measures we found consistent support for the *social network effect* such that the greater the approval (and lower the interference or disapproval) the better the relationship fared. Likewise perceived increases in social network support corresponded to increases in a number of positive aspects of the relationship (e.g., love, commitment, trust) and decreases in criticism. This finding was observed in our replications of the original study and in a meta-analysis of the accumulated literature.

Limitations

The present study was not a perfect replication of the original study. As noted above, and documented in the preregistered proposal, there were differences in sample and administration. Future studies may wish to specifically recruit couples, particularly those experiencing conflict, from the community for an in-person survey over a 6–8 month timeframe. Additional samples may also be considered. Although evidence seems to favor that the majority of romantic relationships are harmed by interference, it may be possible to find couples who thrive despite network disapproval. Studies could investigate, then, how

Table 3. Correlations between social network opinion measures and contemporary relationship quality indices across Time 1 and Time 2

		Time 1		Time 2	
		Parental approval M = 4.16, SD = 0.71	Friend approval M = 4.19, SD = 0.66	Parental approval M = 4.14, SD = 0.78	Friend approval M = 4.14, SD = 0.75
Parental approval (T1)		–		.82**	.51**
	Dating			.77**	.41**
	Married			.85**	.57**
Friend approval (T1)		.59**	–	.48**	.70**
	Dating	.57**		.50**	.71**
	Married	.58**		.48**	.79**
Love (PLS: T1, M = 7.36, SD = 1.18)		.25**	.33**	.22**	.25**
	Dating	.29**	.38**	.21**	.26**
	Married	.23**	.29**	.24**	.25**
Commitment (Lund: T1, M = 7.70, SD = 1.31)		.47**	.57**	.42**	.47**
	Dating	.42**	.59**	.34**	.39**
	Married	.44**	.47**	.46**	.50**
Trust (Rempel: T1, M = 7.25, SD = 1.40)		.46*	.52**	.43**	.45**
	Dating	.44**	.55**	.39**	.42**
	Married	.50**	.51**	.48**	.52**
Criticism (T1)		−.44**	−.47**	−.35**	−.34**
	Dating	−.44**	−.48**	−.28**	−.30**
	Married	−.48**	−.50**	−.45**	−.45**
Love (PLS: T2, M = 7.22, SD = 1.42)		.23**	.25**	.29**	.41**
	Dating	.17*	.24**	.32**	.46**
	Married	.29**	.26**	.27**	.35**
Commitment (Lund: T2, M = 7.59, SD = 1.52)		.39**	.41**	.50**	.62**
	Dating	.28**	.38**	.48**	.61**
	Married	.44**	.37**	.47**	.55**
Trust (Rempel: T2, M = 7.14, SD = 1.57)		.40**	.39**	.50**	.60**
	Dating	.33**	.38**	.53**	.64**
	Married	.46**	.38**	.47**	.55**
Criticism (T2)		−.34**	−.34**	−.41**	−.47**
	Dating	−.27**	−.34**	−.41**	−.53**
	Married	−.40**	−.31**	−.40**	−.38**

Notes. *p < .05, **p < .001. Note, the criticism measure is the same as that used in the original study. All other measures are contemporary validated scales (PLS – Hatfield & Sprecher Passionate Love Scale, Lund Commitment Scale, and Rempel Trust Scale).

Table 4. Correlations between difference scores in approval and relationship quality using contemporary measures

		Increases in parental approval	Increases in friend approval
Increase in love		.24**	.38**
	Dating	.38*	.41**
	Married	*.02*	.30**
Increase in commitment		.29*	.43**
	Dating	.47**	.51**
	Married	*.11*	.23**
Increase in trust		.34**	.50**
	Dating	.49**	.54**
	Married	.15*	.40**
Increase in criticism		−.30*	−.35**
	Dating	−.46**	−.38**
	Married	*−.07*	−.26**

Notes. *p < .05, **p < .001, values in italics p > .05.

Table 5. Random effects models of network type as a predictor of romantic love

		Weighted mean (Hedges *g*) (95% confidence interval)	Homogeneity of effect sizes		
Network type	*k*	Random effects	*Q*	*P*	I^2
Complete	5	1.02 (0.39–1.65)	3.07	0.55	0.00
Family	8	0.32 (0.10–0.55)	12.22	0.09	42.73
Friend	7	0.40 (0.09–0.71)	8.27	0.22	27.48
Overall	11	0.49 (0.26–0.72)	21.70	0.02	53.92

Table 6. Random effects models of network type as a predictor of relationship commitment

		Weighted mean (Hedges *g*) (95% confidence interval)	Homogeneity of effect sizes		
Network type	*k*	Random effects	*Q*	*P*	I^2
Complete	7	0.63 (0.43–0.84)	7.64	0.270	21.46
Family	7	0.56 (0.39–0.73)	20.41	0.002	70.58
Friend	7	0.70 (0.54–0.86)	7.26	0.300	17.36
Overall	16	0.62 (0.50–0.74)	32.58	0.005	53.96

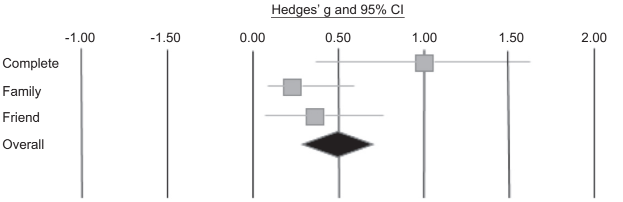

Figure 1. Forest plot of effect sizes for romantic love. Each row represents a network type, with the square the reported effect size (Hedges' *g*) and the bar representing a 95% CI of the effect size. The diamond represents the overall estimated effect size and the distribution of plausible effects sizes for a 95% CI (width of diamond).

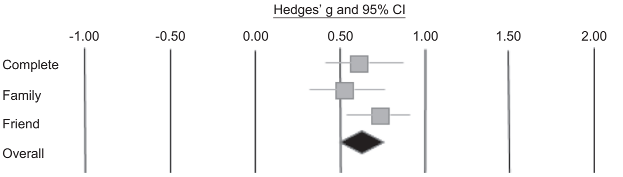

Figure 2. Forest plot of effect sizes for commitment. Each row represents a network type, with the square representing the reported effect size (Hedges' *g*) and the bar representing a 95% CI of the effect size. The diamond represents the overall estimated effect size and the distribution of plausible effects sizes for a 95% CI (width of diamond).

these couples differ from those who do not endure. Also, future research employing online samples may want to investigate additional tactics for increasing return response rates, especially if intent on following participants for a longer duration.

Implications

Even taking the limitations into account, the current study makes a number of contributions. First, by employing the original measures we were better able to draw comparisons to the Driscoll et al. study. Second, we were able to show that although the more infamous findings of the original study did not recur – interference did not correspond to increased love or commitment – we did replicate the two lesser known findings that interference is linked with poorer outcomes on trust and criticism. Third, we were able to demonstrate that there is overlap between measures of interference and network approval, which supports making comparisons between the original study and the subsequent studies, albeit with caution. The measures were not identical and further work should parse the dimensions of network opinion (e.g., approving vs. disapproving, active vs. passive). Fourth, by using diverse relationship quality indices, examining multiple network sources, and employing meta-analysis, we were able show the remarkable consistency of the social network effect – even on the original scales employed by Driscoll and colleagues.

Conclusion

When romanticizing the story of Romeo and Juliet we tend to overlook the fact that, in the end, even Romeo and Juliet confirmed the social network effect. Although they tried to stay together *despite* network disapproval, their relationship ultimately ended. Though few have a relationship end so dramatically, the accumulated evidence makes clear that relationship love and commitment is threatened, not strengthened, by the lack of support of others.

Note From the Editors

A commentary and a rejoinder on this paper are available (Dricsoll, 2014; Wright, Sinclair, & Hood, 2014; doi: 10.1027/1864-9335/a000203).

Acknowledgments

The authors would like to thank Sining Wu and Alison Patev for their assistance on the project, as well as the Advanced Social Psychology lab. We greatly appreciated the reviews received from Richard Driscoll, Susan Sprecher, Diane Felmlee, and anonymous reviewers on various stages of the project. Designed research: H. C. S., K. B. H., B. L. W.; Analyzed data: H. C. S., K. B. H., B. L. W.; Wrote paper: H. C. S., K. B. H., B. L. W. The authors declare no conflict-of-interest with the content of this article. This project received funding from the Center for Open Science (see http://osf.io/6wxgf/), Psi Chi, and the Mississippi State University Office for Research and Economic Development.

PREREGISTERED OPEN DATA OPEN MATERIALS

References

Allan, G. (2006). Social networks and personal communities. In A. L. Vangelisti & D. Perlman (Eds.), *The Cambridge handbook of personal relationships* (pp. 657–671). New York, NY: Cambridge University Press.

Borenstein, M., Hedges, L., Higgins, J., & Rothstein, H. (2005). *Comprehensive Meta-Analysis (Version 2.2.040) [Computer software]*. Engelwood, NJ: Biostat.

DeWall, C. N., Maner, J. K., Deckman, T., & Rouby, D. A. (2011). Forbidden fruit: Inattention to attractive alternatives provokes implicit relationship reactance. *Journal of Personality and Social Psychology, 100*, 621–629.

Dricsoll, R. (2014). Romeo and Juliet through a narrow window. Commentary and rejoinder on Sinclair, Hood, and Wright (2014). *Social Psychology*. Advance online publication. doi: 10.1027/1864-9335/a000203

Driscoll, R., Davis, K. E., & Lipetz, M. E. (1972). Parental interference and romantic love: The Romeo & Juliet effect. *Journal of Personality and Social Psychology, 24*, 1–10.

Etcheverry, P. E., & Agnew, C. R. (2004). Subjective norms and the prediction of romantic relationship state and fate. *Personal Relationships, 11*, 409–428.

Felmlee, D. (2001). No couple is an island: A social stability network perspective on dyadic stability. *Social Forces, 79*, 1259–1287.

Fisher, H. (2004). *Why we love: The nature and chemistry of romantic love*. New York, NY: Henry Holt.

Hatfield, E., & Sprecher, S. (1986). Measuring passionate love in intimate relations. *Journal of Adolescence, 9*, 383–410.

Hooley, J. M., & Teasdale, J. D. (1989). Predictors of relapse in unipolar depressives Expressed emotion, marital distress, and perceived criticism. *Journal of Abnormal Psychology, 98*, 229–235.

Johnson, M. P., & Milardo, R. M. (1984). Network interference in pair relationships: A social psychological recasting of Slater's theory of social regression. *Journal of Marriage and the Family, 46*, 893–899.

Lund, M. (1985). The development of investment and commitment scales for predicting continuity of personal relationships. *Journal of Social and Personal Relationships, 2*, 3–23.

MacDonald, G., & Jessica, M. (2006). Family approval as a constraint in dependency regulation: Evidence from Australia and Indonesia. *Personal Relationships, 13*, 183–194.

Miller, R. (2011). *Intimate relationships*. New York, NY: McGraw Hill.

Parks, M. R. (2007). *Personal relationships and personal networks*. Mahwah, NJ: Erlbaum.

Parks, M. R., Stan, C. M., & Eggert, L. L. (1983). Romantic involvement and social network involvement. *Social Psychology Quarterly, 46*, 116–131.

Reis, H. T. (2011). A history of relationship research in social psychology. In A. W. Kruglanski & W. Stroebe (Eds.), *Handbook of the history of social psychology* (pp. 363–382). New York, NY: Psychology Press.

Rempel, J. K., Holmes, J. G., & Zanna, M. P. (1985). Trust in close relationships. *Journal of Personality and Social Psychology, 49*, 95–112.

Sprecher, S. (2011). The influence of social networks on romantic relationships: Through the lens of the social network. *Personal Relationships, 18*, 630–644.

Sprecher, S., & Felmlee, D. (1992). The influence of parents and friends on the quality and stability of romantic relationships: A three-wave longitudinal investigation. *Journal of Marriage and the Family, 54*, 888–900.

Wright, B. L., & Sinclair, H. C. (2012). Pulling the strings: Effects of friend and parent opinions on dating choices. *Personal Relationships, 19*, 743–748.

Wright, B. J., Sinclair, H. C., & Hood, K. B. (2014). In search of Romeo and Juliet. Commentary and rejoinder on Sinclair, Hood, and Wright (2014). *Social Psychology*. Advance online publication. doi: 10.1027/1864-9335/a000203

Received February 28, 2013
Accepted January 24, 2014
Published online May 19, 2014

H. Colleen Sinclair

Department of Psychology
Mississippi State University
Social Science Research Center
255 Lee Blvd.
Mississippi State, MS 39762
USA
Tel. +1 662 325-4177
E-mail csinclair@ssrc.msstate.edu

Appendix

Papers Included in the Meta-Analysis

Blair, K. L., & Holmberg, D. (2008). Perceived social network support and well-being in same-sex versus mixed-sex romantic relationships. *Journal of Social and Personal Relationships, 25*, 769–791.

Blaney, A., & Sinclair, H. C. (2013). *Who are Romeo and Juliet? Identifying individual differences among those who exhibit the Romeo & Juliet effect*. Poster presented at the 2013 Society for Personality and Social Psychology conference in New Orleans, LA. Unpublished dataset.

Bryan, L., Fitzpatrick, J., Crawford, D., & Fischer, J. (2001). The role of network support and interference in women's perception of romantic, friend, and parental relationships. *Sex Roles, 45*, 481–499.

Bryant, C. M. (1996). *Subcultural variations in the social network support of Latinos, African Americans, and Anglos: What is the association between the development of heterosexual relationships and the support of friends and family members?* (Doctoral dissertation, University of Texas at Austin).

Bryant, C. M., & Conger, R. D. (1999). Marital success and domains of social support in long-term relationships: Does the influence of network members ever end? *Journal of Marriage and the Family, 61*, 437–450.

Driscoll, R., Davis, K. E., & Lipetz, M. E. (1972). Parental interference and romantic love: The Romeo & Juliet effect. *Journal of Personality and Social Psychology, 24*, 1–10.

Etcheverry, P. E., & Agnew, C. R. (2004). Subjective norms and the prediction of romantic relationship state and fate. *Personal Relationships, 11*, 409–428.

Etcheverry, P. E., Le, B., & Charania, M. R. (2008). Perceived versus reported social referent approval and romantic relationship commitment and persistence. *Personal Relationships, 15*, 281–295.

Knobloch, L. K., & Donovan-Kicken, E. (2006). Perceived involvement of network members in courtships: A test of the relational turbulence model. *Personal Relationships, 13*, 281–302.

Le, B., Buniva, E. J., Kozakewich, E., & Etcheverry, P. E. (2008). *Social network relationship approval and commitment to long-distance relationships*. Unpublished manuscript.

Lehmiller, J. J., & Agnew, C. R. (2006). Marginalized relationships: The impact of social disapproval on romantic relationship commitment. *Personality and Social Psychology Bulletin, 32*, 40–51.

Lehmiller, J. J., & Agnew, C. R. (2007). Perceived marginalization and the prediction of romantic relationship stability. *Journal of Marriage and Family, 69*, 1036–1049.

Lewis, R. A. (1973). Social reaction and the formation of dyads: An interactionist approach to mate selection. *Sociometry, 26*, 409–418.

Mitchell, H., Sinclair, H. C., & Barkley, T. (2005). *The Romeo & Juliet effect revisited: Re-examining the influence of familial versus peer (dis)approval on romantic relationship quality*. Poster presented at the 2005 Society for Personality and Social Psychology conference, New Orleans, LA. Unpublished dataset.

Parks, M. R., Stan, C. M., & Eggert, L. L. (1983). Romantic involvement and social network involvement. *Social Psychology Quarterly, 46*, 116–131.

Parks, M. R., & Eggert, L. L. (1991). The role of social context in the dynamics of personal relationships. In W. H. Jones & D. Perlman (Eds.), *Advances in personal relationships: A research annual, Vol. 2. Advances in personal relationships* (pp. 1–34). Oxford, UK: Jessica Kingsley.

Sinclair, H. C. (2007). *Investigating effects of social opinion on romantic relationships*, Paper presented at the 2007 Southeastern Psychological Association conference in New Orleans, LA. Unpublished dataset.

Sprecher, S. (1988). Investment model, equity, and social support determinants of relationship commitment. *Social Psychology Quarterly, 51*, 318–328.

Sprecher, S., & Felmlee, D. (1992). The influence of parents and friends on the quality and stability of romantic relationships: A three-wave longitudinal investigation. *Journal of Marriage and the Family, 54,* 888–900.

Zak, A., Coulter, C., Giglio, S., Hall, J., Sanford, S., & Pellowski, N. (2002). Do his friends and family like me? Predictors of infidelity in intimate relationships. *North American Journal of Psychology, 4*, 287–290.

Zhang, S., & Kline, S. L. (2009). Can I make my own decision? A cross-cultural study of perceived social network influence in mate selection. *Journal of Cross-Cultural Psychology, 40*, 3–23.

Breakthrough or One-Hit Wonder?

Three Attempts to Replicate Single-Exposure Musical Conditioning Effects on Choice Behavior (Gorn, 1982)

Ivar Vermeulen,[1] Anika Batenburg,[1] Camiel J. Beukeboom,[1] and Tim Smits[2]

[1]Faculty of Social Sciences, Communication Science, VU University Amsterdam, The Netherlands,
[2]Faculty of Social Sciences, Institute for Media Studies, Katholieke Universiteit Leuven, Belgium

Abstract. Three studies replicated a classroom experiment on single-exposure musical conditioning of consumer choice (Gorn, 1982), testing whether simultaneous exposure to liked (vs. disliked) music and a pen image induced preferences for the shown (vs. a different) pen. Experiments 1 and 2 employed the original music, Experiment 3 used contemporary music. Experiments 2 and 3 employed hypothesis-blind experimenters. All studies incorporated post-experimental inquiries exploring demand artifacts. Experiments 1 and 2 (original music; N = 158, N = 190) showed no evidence for musical conditioning, and were qualified (conclusive) replication failures. Experiment 3 (contemporary music; N = 91) reproduced original effects, but with significantly smaller effect size. Moreover, it had limited power and showed extreme scores in one experimental group. Aggregated, the three studies produced a null effect. Exploration of demand artifacts suggests they are unlikely to have produced the original results.

Keywords: music in advertising, musical conditioning, demand characteristics, direct replication, Gorn

This paper focuses on replicating the first experiment in Gerald Gorn's article "The effects of music in advertising on choice behavior: A classical conditioning approach," published in the *Journal of Marketing*, 1982. The original experiment's findings are taken as evidence that music can unobtrusively, and through single exposure, condition consumer choice behavior. The study has an almost iconic status and is impressively proliferated through the literature. Google Scholar reports 662 citations, Web of Knowledge 243 (October 30, 2013). Moreover, the study appears in nearly every student textbook on persuasion and consumer psychology (e.g., Cialdini, 2001; Fennis & Stroebe, 2010; Peck & Childers, 2008; Saad, 2007). Several replications of the original study failed, but, as will be described shortly, none exactly followed the original procedures. A direct replication is still lacking.

The original experiment (Gorn, 1982, Experiment 1) involved a 2 (Pen color: light blue vs. beige) × 2 (Music: liked vs. disliked) between-subjects design, conducted among 244 undergraduate students in a management course. Participants were asked, during class time, to evaluate a piece of music that an advertising agency considered for a pen commercial. Depending on the experimental condition, they were then exposed to a picture of a light blue or beige pen on a big screen while either "liked" or "disliked" music (an excerpt from the movie Grease vs. classical Indian music) played for 1 minute. To thank participants, they were offered a light blue or a beige pen (one of which was previously "advertised" on screen). Upon leaving the classroom they could choose a pen from one of two boxes, with question sheet drop-off boxes next to them.

Results showed that 79% of participants in the "liked" music conditions chose the pen in the color displayed on screen. Only 30% of participants in the "disliked" music conditions chose the displayed pen. Furthermore, when asked afterwards for their reasons to choose a particular pen color, only 2.5% of participants mentioned the music. These findings suggest that simple, fairly unobtrusive cues like music can influence consumer choice behavior following single exposure.

Replicating Gorn's experiment is important, not only because of its impact on the persuasion literature, but also because its rather unconventional procedure has repeatedly been criticized (e.g., Allen & Madden, 1985; Kellaris & Cox, 1989). Replicating the study constitutes a challenge, as it involves stimuli susceptible to cultural trends and changes (e.g., musical preferences).

Existing Evidence

The original study (Gorn, 1982) produced a rather strong (Cohen, 1988) effect size: φ = .49. Some found this noteworthy because it does not seem to employ a particularly powerful conditioning procedure (e.g., Bierley, McSweeney, & Vannieuwkerk, 1985). Participants were exposed to stimuli only once instead of repeatedly;

DOI: 10.1027/1864-9335/a000182

although conditioning effects have been shown after single trials (Stuart, Shimp, & Engle, 1987, Experiment 1), such effects usually require very strong stimuli, like nauseating drugs or intense shocks. Moreover, in Pavlovian conditioning, strongest results are usually reported when conditioned stimuli (pens) are presented before unconditioned stimuli (music) rather than simultaneously.

Some scholars suggested that the strong effects found in the original study originated in demand artifacts (Allen & Madden, 1985; Kellaris & Cox, 1989). Participants' awareness of the study's purpose may have elicited behavior congruent to inferred expectations. Gorn inferred participants did not know the study's purpose, as only five out of 205 mentioned music as a reason for pen choice. However, Allen and Madden (1985) commented that the original post-experimental inquiry lacked detail and rigor.

To elucidate these issues, several scholars conducted replication studies. The literature contains two studies demonstrating results congruent to the original (Bierley et al., 1985; Groenland & Schoormans, 1994), and three that failed to show significant findings (Allen & Madden, 1985; Kellaris & Cox, 1989, Experiments 1 and 3). However, the experimental procedures of all replications (both successful and unsuccessful) included fundamental modifications that could have caused different findings. For example, experiments were conducted in cubicles instead of a classroom, researchers used alternative liked and disliked musical, or even nonmusical, stimuli (e.g., humor segments), participants were offered pens from one rather than two boxes, exposed to one color pen only, or asked to answer questions about the advertised pens' characteristics before pen choice. Therefore, none of these studies can be considered direct replications. For a more extensive review of prior replication attempts, see the preregistered proposal (http://osf.io/z6e8j).

Aforementioned replication attempts failed to show conclusive evidence that attributes the original findings to demand characteristics. Shimp, Hyatt, and Snyder (1991, 1993) conclude, after analyzing the failed replications by Kellaris and Cox (1989), that the original findings (Gorn, 1982) more likely originate in successful conditioning than in demand artifacts. In sum, several replications of the original study were conducted but none were direct, and the aggregated knowledge remains inconclusive.

Study Outline and Power

The current replication attempt follows a three-step approach. Experiment 1 replicates the original procedures using (close to) original materials and (following the original study) a fully informed experimenter team. Experiment 2 employs the same procedures and materials, but uses non-informed experimenters; it also employs more extensive post-experimental questionnaires to explore demand characteristics. Experiment 3 mirrors the basic procedures of Experiment 2, but uses contemporary rather than the 1980s musical stimuli.

Planned sample sizes (250, 240, and 200, respectively) were based on estimated attendance of the classes in which the experiments would take place, and on a priori power analysis using *G*Power* (Erdfelder, Faul, & Buchner, 1996). This analysis showed that, assuming $\alpha = .05$ and inclusion of 80% of participants in the main analyses (cf. the original study), $N = 69$ would suffice to detect the original effect size ($\varphi = .49$; logOR = 2.17, 95% CI = 1.52–2.82)[1] with .95 power. Recently, Simonsohn (2013) suggested samples for replication studies should multiply the original sample by 2.5 (thus, in this case, $N = 610$ per study) to enable reliable detection of the effect size that would have given the original study .33 power (in this case, $\varphi_{33\%} = .11$; logOR = .44). However, such sample sizes (classes of 610 students) are unattainable in the current experimental set-up. For replications of large studies, Simonsohn suggests testing results against a practically or theoretically "small" point null effect size. We found no theoretical footholds to determine such a point null. Based on a simple return-on-investment advertising scenario, we set a practical point null effect size at half the original effect size ($\varphi = .25$; logOR = 1.04). Note however, that this point null is fairly arbitrary, and was set post hoc to enable categorization of replication outcomes (cf. Simonsohn, 2013).

Due to falling student numbers and low class attendance (for Experiment 3), we did not reach planned sample sizes, but effective samples of $N = 158$, $N = 190$, and $N = 91$ instead. Note that these samples still provide > .95 power to detect the original effect. Actual obtained power and sensitivity analyses will be presented with each sample.

Experiment 1

Experiment 1 directly replicated Gorn's original study including (1) the original musical stimuli most likely used, (2) pens in two pretested colors, (3) a balanced 2 (Music: liked vs. disliked) × 2 (Advertised Pen: color 1 vs. color 2) design, (4) original instructions and procedures, and (5) presence of an experimenter team aware of the hypothesis tested.

Participants

Participants were 160 second year BA Communication students, recruited through attendance of a persuasion class taught at a Belgian university. At onset, participants were unaware of the study's purpose, as (1) no informed consent was asked, (2) the experiment was not announced, (3) no prior references to the original study had been made earlier in the participants' curriculum. Students did not receive credits or money for participation.

[1] We will report log odds ratios (logOR) for all χ^2 tests; logOR's are approximately normally distributed, and therefore easy to interpret (logOR = 0 indicates no effect, and logOR is in the center of its CI; Bland & Altman, 2000).

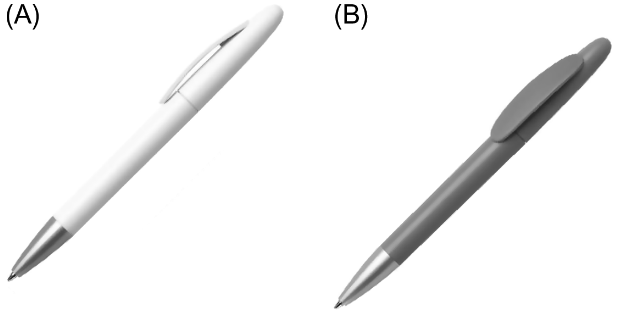

Figure 1. White pen (A), light blue pen (B).

Materials

Pens

In an online pretest, 48 Dutch MA Communication students ($M_{\text{age}} = 22.77$, $SD = 3.83$; 19 male, 28 female, 1 unknown) judged 13 pens differing only in color. Of the pens scoring similar to the collective mean (which excluded beige), the difference between white and light blue pens was the smallest in all possible pairs ($M = 4.31$ vs. 4.25; scale 1–7; $F(1, 47) = .04$, 95% $\text{CI}_{\text{diff}} = -.57$ to .70, $p = .84$, $\eta^2 = .00$) – hence these were selected as stimuli. See electronic supplementary materials for details. Pens were offered in two unmarked carton boxes holding 150 pens. For display on the slides, professional pictures from the manufacturer's website were used (see Figure 1).

Music

The original paper left the exact musical stimuli unspecified. Aided by the original author, we selected five songs from the musical Grease (liked music) and five classical Indian songs by Parveen Sultana (disliked music), which we pretested in 47 students ($M_{\text{age}} = 22.00$, $SD = 1.98$; 21 male, 26 female; 29 Dutch, 18 Belgian; see supplementary materials). Based on this pretest, and avoiding possible lyrical confounds, we selected the Grease song "Summer Nights" ($M = 5.37$; scale 1–7) as liked music and the Parveen Sultana song "Aaj Kaun Gali Gayo Shyam" ($M = 2.56$) as disliked music. Both differed significantly in evaluation, $F(1, 46) = 126.48$, 95% $\text{CI}_{\text{diff}} = 2.31\text{–}3.31$, $p < .001$, $\eta^2 = .73$.

Procedure

The informed experimenter team consisted of one of the authors (teaching the class) and four non-naive assistants.

Participants received an announcement a week prior to the experiment, explaining that the class would be split in two due to scheduling problems; division was approximately 50/50 based on alphabetical order. Halfway each class, students with odd student ID numbers were asked to follow an assistant to a waiting room, while even numbered students stayed. After completing the study, even numbered students left for another waiting room while odd numbered students returned.

While two assistants distributed the music evaluation forms, the experimenter explained – following the original script – that an advertising agency was trying to select music to use in a commercial for a pen produced by one of its clients. Participants would hear some music that was being considered while they would see an image of the pen that the agency was planning to advertise on a PowerPoint slide.

While displaying the pen for one minute, a music excerpt was played over the class' sound system. Afterwards, participants evaluated the music on the form. Subsequently, they were told that they would receive either a white or a blue pen for their help, donated by the manufacturer. The experimenter held up each pen briefly and commented that if they wanted a white one, they should go to the box positioned left of the class room's exit, whereas if they wanted a blue one, they should go to the box positioned on the right[2], and drop off their question form next to the boxes. Participants were invited to line up for the exit, thus exposing them equally to both boxes. After collecting their pens, participants were given a brief questionnaire with four open questions (see below) and were asked to indicate pen choice, age, gender, and the last three digits of their student ID. Afterwards, the class continued. Debriefing took place in the subsequent class.

Measures

Music evaluation was measured using three 5-point Likert scale items (see supplementary materials; $\alpha = .93$).

Pen choice was measured by (1) assessing at which table music evaluation forms were handed in, (2) an unobtrusive code on the post-choice question forms handed out at the same tables, and (3) participants' self-reported pen color choice on the post-choice form. Each form asked for age, gender, and last three student ID digits, enabling us to link responses. Participants with conflicting pen choice measurements were excluded.

Reasons for pen choice were measured using three open answer boxes. See supplementary materials for details.

Hypothesis awareness – participants' awareness of the hypothesized relationship between music played and pen choice – was assessed in one open question asking participants about the goal of the session they just attended. Answers were coded between 0 and 5 for increasing hypothesis awareness (see supplementary materials for details).

[2] Here, we diverted slightly from the procedure described in the original paper, which stated that the boxes were positioned on the left and right side of the classroom. In response to our concerns that this positioning could lead to participants bumping into each other or going with the flow, the original author stated that he had actually used equidistant placement of the boxes on opposite sides of the exit door. Hence, this is what we used.

Table 1. Frequencies (and percentages) of choice for advertised and non-advertised pen in the liked and disliked music conditions, in the original study and the current three replications

	Liked music		Disliked music	
	Advertised pen	Non-advertised pen	Advertised pen	Non-advertised pen
Original Experiment (Gorn, 1982, Experiment 1)	74 (79%)	20 (21%)	30 (30%)	71 (70%)
		N = 195; $\chi^2(1)$ = 47.01, p < .001, φ = .49		
Experiment 1 (exact replication)	34 (48%)	37 (52%)	39 (54%)	33 (46%)
		N = 143; $\chi^2(1)$ = .56, p = .45, φ = −.06		
Experiment 2 (exact replication with additions)	38 (43%)	50 (57%)	38 (53%)	34 (47%)
		N = 160; $\chi^2(1)$ = 1.46, p = .23, φ = −.10		
Experiment 3 (replication with updated music)	21 (57%)	16 (43%)	8 (23%)	27 (77%)
		N = 72; $\chi^2(1)$ = 8.59, p = .003, φ = .35		

Notes. Depicted chi-square test results are confirmatory analysis equal to original study, excluding participants with deviant music evaluation.

Results

Two participants were removed from the sample because we could not reliably assess pen choice (form code and self-reported choice were incongruent). Analyses were conducted on the remaining 158 participants (M_{age} = 20.28, SD = 1.48; 37 male, 121 female). Music evaluations for "Summer Nights" were more positive (M = 4.18, SD = 0.53) than for the Indian music (M = 2.27, SD = 0.67; F(1, 157) = 384.96, 95% CI_{diff} = 1.72–2.10, p < .001, η^2 = .71), indicating that the music manipulation was successful.

Confirmatory Analyses

First, we replicated the original study's analysis by excluding participants who either (somewhat) disliked the liked music (evaluation below 3) or liked the disliked music (evaluation above 3), leaving 143 participants (M_{age} = 20.26, SD = 1.43; 35 male, 108 female; $N_{\text{liked\_music}}$ = 71 (35 white pen, 36 blue pen), $N_{\text{disliked\_music}}$ = 72 (33 white pen, 39 blue pen). Actual power of this sample to detect the original effect of φ = .49 is 1.00, power to detect the point null effect of φ = .25 is .85. Sensitivity analysis shows that the sample provides .95 power to detect a φ = .30 effect size, and .8 power to detect φ = .23.

A chi-square test of "advertised" versus "non-advertised" pen choice against "liked" versus "disliked" music showed no effect of music on pen choice, $\chi^2(1)$ = .56, p = .45, φ = −.06, logOR = −.25, 95% CI = −.91 to .41). For cell frequencies, see Table 1. Because the 95% CI included 0, the main hypothesis was rejected. The obtained logOR is significantly lower than the original 2.17 (with 95% CI = 1.52–2.82). The obtained CI did not include the logOR of 1.04 associated with the point null effect of φ = .25. Based on the former criterion (e.g., Asendorpf et al., 2013), the current replication failed; based on the latter it should be regarded a conclusive failure (cf. Simonsohn, 2013).

Testing the main hypothesis with all 158 participants included (M_{age} = 20.28, SD = 1.48; 37 male, 121 female; $N_{\text{liked\_music}}$ = 73 [35 white pen, 38 blue pen], $N_{\text{disliked\_music}}$ = 85 [41 white, 44 blue]) showed no effect of music on pen choice, $\chi^2(1)$ = .21, p = .65, φ = −.04, logOR = −.15, 95% CI = −.77 to .48; for liked music the ratio between advertised versus non-advertised pen choice was 36 versus 37; for disliked music 45 versus 40. As the 95% CI includes 0, the hypothesis was rejected.

Exploratory Analyses

Only 3.8% of participants mentioned music as a reason for choice, and 66.5% mentioned color preference. These results emulate those obtained in the original study. In describing the goal of the study, 46.2% of participants mentioned "influencing pen choice," indicating that many inferred pen choice was the main outcome variable. Hypothesis awareness was marginally higher for the liked music condition (M = 2.08, SD = 1.68) then for the disliked music condition (M = 1.56, SD = 1.62; F(1, 157) = 3.87, 95% CI_{diff} = −.02 to 1.04, p = .05, η^2 = .02), suggesting that demand artifacts could be more prominent in the former. However, logistic regression showed no effect of hypothesis awareness on choosing the "hypothesized" pen (OR = .93, 95% CI = .77–1.12, p = .44), indicating that hypothesis awareness did not transfer systematically into compliant or contravening pen choice. Thus, the current failure to replicate cannot be attributed to systematic biases in choice behavior of differentially hypothesis-aware participants. More exploratory results are reported in the supplementary materials.

Discussion

The current experiment showed no effect of music on pen choice. Because the 95% CI_{OR} included 0, and did not include the originally obtained OR, nor an OR associated with the point null effect, the current results amount to a (conclusive) replication failure. Although some prior studies suggested that hypothesis awareness and resulting demand artifacts may have amplified or reduced the original study's effects, our results shows no systematic relation between hypothesis awareness and choice behavior.

Experiment 2

Compared to Experiment 1, two differences applied: (1) to avoid demand artifacts resulting from the presence of an involved researcher the experimenter team was naive; (2) to further explore possible demand artifacts, a more extensive post-experimental inquiry was conducted.

Participants

Participants were 195 second year BA Communication students, recruited through attendance of a persuasion class taught at a Dutch university. Participants were unaware of the purpose of the study, and did not receive credits or money for participation – they received credits for completing a complementary (unplanned) post-experimental questionnaire 2 weeks later.

Materials, Procedure, and Measures

Materials were the same as in Experiment 1. A male professional actor, posing as a researcher from another department, conducted the experiment with four assistants. All received a thorough briefing on the study's procedures, but were left naive regarding its purposes and hypotheses. Procedures were equal to Experiment 1's except for the inclusion of a more extensive post-experimental inquiry.

The planned post-experimental questionnaire contained questions (adapted from Allen & Madden, 1985) pertaining to demand artifacts (see supplementary materials). Two weeks later, 116 of the 195 participants filled out a secondary (not planned in preregistered proposal) online post-experimental questionnaire focusing on possible variations in experimental procedures, credibility of cover story, additional reasons for choice, and demand artifacts (see supplementary materials). Participants were debriefed in the subsequent class. Other measures were the same as in Experiment 1; music evaluation's α was .93.

Results

Five participants were excluded for having attended the course (featuring the original study) previously, leaving 190 participants. "Summer Nights" was evaluated more positively (M = 3.72, SD = .75) than the Indian music (M = 2.11, SD = .76; $F(1, 189) = 210.47$, 95% $\mathrm{CI}_{\mathrm{diff}}$ = 1.39–1.83, $p < .001$, $\eta^2 = .53$), indicating a successful music manipulation.

Confirmatory Analyses

After excluding participants with an evaluation below 3 for the Grease song, and above 3 for the Indian music, 160 participants remained ($M_{\mathrm{age}} = 21.17$, SD = 2.03; 42 male, 118 female; $N_{\mathrm{liked\_music}} = 88$ [51 white pen, 37 blue pen], $N_{\mathrm{disliked\_music}} = 72$ [28 white, 44 blue]; actual power to detect $\varphi = .49$: 1.00; power to detect the point null $\varphi = .25$: .89; .95 sensitivity: $\varphi = .28$; .8 sensitivity: $\varphi = .22$). A chi-square test showed no effect of music on pen choice, $\chi^2(1) = 1.46$, $p = .23$, $\varphi = -.10$, logOR = −.39, 95% CI = −1.01 to .24). See Table 1 for cell frequencies. The 95% CI included 0, the obtained logOR is not included in the original CI, and the obtained CI did not include the point null logOR of 1.04. Therefore, the current replication should be regarded a conclusive failure (Simonsohn, 2013).

Testing the main hypothesis on all 190 participants ($M_{\mathrm{age}} = 21.21$, SD = 1.97; 55 male, 134 female, 1 n/a; $N_{\mathrm{liked\_music}} = 110$ [61 white pen, 49 blue pen], $N_{\mathrm{disliked\_music}} = 80$ [48 white, 32 blue]), similarly showed no effect of music on pen choice, $\chi^2(1) = 1.17$, $p = .28$, $\varphi = -.08$, logOR = −.32, 95% CI = −.90 to .26); liked music 49 (advertised pen) versus 61 (non-advertised pen); disliked music 42 versus 38. As the 95% CI includes 0, the hypothesis was rejected.

Exploratory Analyses

Again, few participants (3.7%) mentioned music as a reason for pen choice, and many (55.8%) mentioned color preference; 34.2% mentioned "influencing pen choice" as study goal. Contrasting to Experiment 1, hypothesis awareness did not differ between liked and disliked music conditions, $F(1, 184) = .70$, 95% $\mathrm{CI}_{\mathrm{diff}}$ = −.22 to .41, $p = .42$, $\eta^2 = .00$. No effect of hypothesis awareness on choosing the "hypothesized" pen was found (OR = .94, 95% CI = .75–1.18, p = .59), indicating no systematic compliant or contravening behavior in hypothesis guessers.

In the unplanned secondary post-experimental inquiry, 74.1% of participants reported choosing the pen for own reasons; 7.4% indicated to have complied with the perceived study goal, whereas 18.8% contravened. Logistic regression shows that hypothesis awareness (determined from first post-experimental questionnaire) predicts these latter two behaviors combined (OR = 1.68, 95% CI = 1.21–2.34, p = .002); within the participants reacting on perceived study goals, hypothesis awareness elicited contravention rather than compliance (OR = 2.76, 95% CI = 1.17–6.55, p = .02). These results indicate that hypothesis awareness induces goal-contravening behaviors rather than goal-compliant behaviors. See supplementary materials for further exploratory analyses.

Discussion

The second experiment also did not show an effect of music on pen choice. The 95% CI included 0, and did not include the originally obtained effect size, nor the point null effect size. Thus, like Experiment 1, the current replication should be regarded a conclusive failure. The extensive post-experimental questionnaires showed that hypothesis awareness yields contravention to, rather than compliance with, perceived research goals. This makes sense: only by choosing the opposing pen participants can demonstrate they "outsmarted" the experimenters. Our results corroborate analyses by Shimp et al. (1991, 1993), who showed it was unlikely that demand artifacts caused the original findings (Gorn, 1982).

Experiment 3

Possibly, both direct replications described above failed because of the reuse of the original, over 30 years old, musical stimuli. Exposure to outdated music might elicit cognitive reflection on the experimental situation, and induce a state of involvement that could impede associative learning (Gorn, 1982, Experiment 2). Reflection might also enhance hypothesis awareness, in turn eliciting reactive responses (as seen in Experiment 2). Alternative to both prior experiments, Experiment 3 uses contemporary music.

Participants

Participants were 93 first year BA Communication students, recruited through attendance of an introduction on communication taught at a Dutch university. They were unaware of the purpose of the study, and did not receive credits or money for participation.

Materials, Procedure, and Measures

Because previous research attributed the original findings to the musical selections' differences in familiarity, lyrics, cultural origin, genre, tempo, and instrumentation (Kellaris & Cox, 1989), our aim was to select music similar on all these characteristics, and differing only in elicited affect. The pretest (N = 47) described above (see supplementary materials) also tested six contemporary pop songs against poor but professionally produced renditions by cover artists. Based on this pretest, we selected two renditions of the Rihanna song "We found love" as liked and disliked music. Both versions featured female singers and the same tempo, song sequence, and lyrics. Mean evaluations (M = 5.60 vs. 2.48) differed significantly, $F(1, 46) = 196.38$, 95% $\text{CI}_{\text{diff}} = 2.67\text{–}3.57$, $p < .001$, $\eta^2 = .81$.

The procedure of Experiment 3 emulated Experiment 2, omitting the elaborate post-experimental inquiry. Measures emulated Experiment 1; music evaluation's α was .87.

Results

Two participants were excluded because they were also enrolled in the second year BA class where Experiment 2 took place. Analyses were conducted on the remaining 91 participants. Music evaluation for the liked music (Rihanna) was more positive (M = 3.73, SD = .77) than for the disliked music (cover artist; M = 2.19, SD = .77; $F(1, 90) = 90.89$, 95% $\text{CI}_{\text{diff}} = 1.23\text{–}1.86$, $p < .001$, $\eta^2 = .51$), indicating the music manipulation was successful.

Confirmatory Analyses

After excluding participants with an evaluation below 3 for the liked music, and above 3 for the disliked music, 72 participants remained (M_{age} = 19.26, SD = 2.13; 17 male, 55 female; $N_{\text{liked\_music}}$ = 37 [15 white pen, 22 blue pen], $N_{\text{disliked\_music}}$ = 35 [17 white, 22 blue]; actual power to detect $\varphi = .49$: .99; power to detect the point null $\varphi = .25$: .56; .95 sensitivity: $\varphi = .42$; .8 sensitivity: $\varphi = .33$). This time, the chi-square test analyzing advertised versus non-advertised pen choice against liked versus disliked music showed a significant effect, $\chi^2(1) = 8.59$, $p = .003$, $\varphi = .35$, logOR = 1.49, 95% CI = .47–2.51. Cell frequencies were in the hypothesized direction (see Table 1). Because the 95% CI did not include 0, the main hypothesis was accepted. However, the obtained logOR of 1.49 is significantly lower than the original 2.17 (with 95% CI = 1.52–2.82). Employing this criterion (e.g., Asendorpf et al., 2013) the replication failed, even though the main hypothesis was accepted. Note that Simonsohn (2013) argued against considering replications failed when obtained effect sizes differ from the original. Instead he suggests considering replications that establish a significant effect in the hypothesized direction, and not significantly smaller than the point null effect, as successful (Simonsohn, 2013). Given that we acquired relatively small (.56) power to detect the point null effect, we concordantly qualify the current findings as a somewhat unreliable replication success.

Testing the main hypothesis on all 91 participants (M_{age} = 19.25, SD = 1.97; 20 male, 71 female; $N_{\text{liked\_music}}$ = 45 [16 white pen, 29 blue pen], $N_{\text{disliked\_music}}$ = 46 [24 white, 22 blue]) showed similar results: liked music promoted advertised pen choice, $\chi^2(1) = 4.03$, $p = .045$, $\varphi = .21$, logOR = .87, 95% CI = .01–1.73; for liked music, advertised versus non-advertised pen choice was 23 versus 22; for disliked music 14 versus 32. As the 95% CI did not include 0, the hypothesis was accepted.

Exploratory Analyses

Exploratory analyses showed that the effects observed in the current study stem largely from one experimental group (disliked music/blue pen), where 20 out of 22 participants chose the white pen. Worried about experimental anomaly, we tested whether this group differed from other groups regarding reasons provided for pen choice (e.g., "one box was better accessible"; "I followed a friend") or hypothesis awareness. We found no significant differences (see supplementary materials). Also, the research assistants on site reported no anomalies in their post-experimental assessment report.

Of the total sample, 5.5% mentioned music influence and 58.2% color preference as a reason for choice; 58.2% mentioned the central item "influencing pen choice" as study goal. Hypothesis awareness was similar for liked and disliked music, $F(1, 86) = .67$, 95% $\mathrm{CI}_{\mathrm{diff}} = -.74$ to .39, $p = .54$, $\eta^2 = .00$. Notably, hypothesis awareness negatively affected choosing the "hypothesized" pen (OR = .71, 95% CI = .50–1.00, $p = .05$). More aware participants tended to contravene study goals, indicating that the observed effects of music on pen choice cannot be attributed to compliant behavior of participants "in the know."

Discussion

The final experiment showed the hypothesized effect of music on pen choice. The obtained effect size was significantly smaller than the original, but exceeded the null point effect. Although these results qualify a "successful replication" cf. Simonsohn (2013), achieved power was fairly low, and observed effects originated mostly in one experimental group. Therefore, conclusions from Experiment 3 should be drawn cautiously. Observed effects cannot be attributed to demand artifacts – hypotheses-aware participants chose the "hypothesized" pen significantly less often.

Aggregated Results

Aggregating our data, and excluding participants with "deviant" musical taste, we found no effect of music on pen choice, $N = 375$; $\chi^2(1) = .00$, $p = .99$, $\varphi = .00$, logOR = .00, 95% CI = −.41 to .40. The 95% CI included 0, and did not include the original 2.17, nor the OR of 1.04 associated with a point null effect, nor the OR of 0.44 associated with Simonsohn's (2013) $\varphi_{33\%}$ criterion (for which our aggregated sample provides .57 power). In unison, the three experiments failed to replicate the original results.

Adding the reconstructed data (195 cases) from the original study to ours (thereby aggregating all known direct replications; $N = 570$) the original effect still holds up, $\chi^2(1) = 15.54$, $p < .001$, $\varphi = .17$, logOR = .67, 95% CI = .33–1.00, due to the strong effect obtained in the original study. However, the aggregated effect size is small (Cohen, 1988).

Conclusion

Five conclusions can be drawn: (1) two well-powered replications failed to reproduce the original effect (Gorn, 1982). If the reader accepts the proposed $\varphi = .25$ point null effect, both can be considered conclusive replication failures cf. Simonsohn (2013); (2) a smaller replication using updated and matching musical selections – sufficiently powered to reliably detect the original effect, but featuring one experimental group with extreme scores – reproduced the original findings, but with a significantly smaller effect size. All in all, we labeled it a somewhat unreliable successful replication; (3) in aggregate, our studies conclusively failed to replicate the original effect; (4) aggregated data from all four known direct replications (including the original study) still show an effect of music on pen choice, though with considerably smaller effect size; (5) hypothesis awareness tends to elicit contravening rather than compliant responses in participants, rendering it unlikely that the original results were due to demand artifacts, as previously implied (Allen & Madden, 1985; Kellaris & Cox, 1989).

If, as suggested by Experiments 1 and 2, musical conditioning effects on pen choice do not exist, more replications would be needed to fully neutralize the original effect. Notably, 2,207 additional cases would be needed to push the aggregated effect below significance level (assuming future replications would consistently produce null effects).

Alternatively, if, as implied by Experiment 3, the proposed effect does sometimes emerge, moderators or confounds may be at play. Note that the original musical stimuli differ on more characteristics than elicited affect alone: for example, familiarity, lyrics, cultural origin, genre, tempo, and instrumentation. It is possible that these specific differences either amplified effects in the original 1982 sample (Kellaris & Cox, 1989) or dampened effects in the present samples. Note that Experiment 3's stimuli were not only contemporary, but also matched on all above, possibly confounding, differences, which discounts them as alternative explanations of observed effects. In addition, the current results discount participants' hypothesis awareness as an alternative explanation of the observed effects.

Limitations

The current research has several limitations, some of which may inform future replicators. First, it is uncertain whether we fully reproduced the original musical materials in Experiments 1 and 2. We selected the song "Summer Nights" as liked music because it posed no lyrical confounds, in contrast to the other candidate Grease songs. Yet, possibly, the original study used the song "You're the one that I want" to advertise pens. If so, this would provide a compelling alternative explanation of the strong effects observed.

Second, it was impossible to fully reproduce the original experimental context, not only because times and locations were different, but also because in general classroom experiments are very susceptible to noise. Small procedural variations or disruptions may influence entire experimental conditions, and group dynamics may influence individuals' behaviors. One might even question whether the current experimental set-up is suited to reliably determine subtle conditioning effects. To advance knowledge on the potential of single-exposure musical conditioning of consumer choices, future (conceptual) replicators might better employ individualized experiments.

Third, we did not achieve planned power for our experiments. This is problematic for Experiment 3, which had relatively low power and therefore was susceptible to chance findings. Possibly, the extreme scores observed in one experimental group were such a chance finding. Future replicators might prefer experimental set-ups in which sample sizes can be fully controlled. Had Experiment 3's sample size been as planned, we could have attributed it more weight.

Finally, our research's theoretical contribution is limited. By emulating the original, rather unconventional, experimental set-up, we did not advance much in answering whether consumer choice can be conditioned through single exposure to music. We did establish, however, that the original well-cited findings (Gorn, 1982) were not – to paraphrase our title – a one-hit wonder, as testified by Experiment 3. To determine whether the findings constitute a theoretical ''breakthrough,'' however, much care should be taken to eliminate possible confounds, preferably in non-classroom conceptual replications.

Acknowledgments

We thank Gerry Gorn for his support and for providing us with many details about the original experiment that helped us replicate it as faithfully as possible. We also thank Thijs Brenk and Sander Wensink, as well as 539 anonymous students at VU Amsterdam and KU Leuven, who spent considerable time and effort in participating in pretests and experimental sessions. We report all data exclusions, manipulations, and measures, and how we determined our sample sizes. This research was funded with a replicator grant of $1980 from the Center for Open Science. The authors declare no conflict-of-interest with the content of this article. Designed research: I.V., A.B., C.B.; Performed research: A.B., C.B., T.S., I.V.; Coded data: C.B., T.S., I.V.; Analyzed data: I.V., C.B., A.B., T.S.; Wrote paper: I.V., A.B., C.B., T.S. All materials, data, descriptions of the procedure, and the preregistered design are available at http://osf.io/ietwv/, registered at http://osf.io/cbn4x/. The design was registered, prior to data collection, at http://osf.io/z6e8j.

PREREGISTERED OPEN DATA OPEN MATERIALS

References

Allen, C. T., & Madden, T. J. (1985). A closer look at classical conditioning. *Journal of Consumer Research, 12*, 301–315.

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.

Bierley, C., McSweeney, F. K., & Vannieuwkerk, R. (1985). Classical conditioning of preferences for stimuli. *Journal of Consumer Research, 12*, 316–323.

Bland, J. M., & Altman, D. G. (2000). The odds ratio. *British Medical Journal, 320*, 1468.

Cialdini, R. B. (2001). *Influence: Science and Practice* (4th ed.). Boston, MD: Allyn & Bacon.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Mahwah, NJ: Erlbaum.

Erdfelder, E., Faul, F., & Buchner, A. (1996). GPOWER: A general power analysis program. *Behavior Research Methods, Instruments, & Computers, 28*, 1–11.

Fennis, B. M., & Stroebe, W. (2010). *The psychology of advertising*. New York: Psychology Press.

Gorn, G. J. (1982). The effects of music in advertising on choice behavior: A classical conditioning approach. *Journal of Marketing, 46*, 94. doi: 10.2307/1251163

Groenland, A. G. E., & Schoormans, J. P. L. (1994). Comparing mood-induction and affective conditioning as mechanisms influencing product evaluation and product choice. *Psychology and Marketing, 11*, 183–197.

Kellaris, J. J., & Cox, A. D. (1989). The effects of background music in advertising: A reassessment. *Journal of Consumer Research, 16*, 113. doi: 10.1086/209199

Peck, J., & Childers, T. L. (2008). Effects of Sensory Factors on Consumer Behavior. In C. P. Haugtvedt, P. M. Herr, & F. R. Kardes (Eds.), *Handbook of consumer psychology* (pp. 193–219). New York, NY: Erlbaum.

Saad, G. (2007). *The evolutionary bases of consumption*. Mahwah, NJ: Erlbaum.

Shimp, T. A., Hyatt, E. M., & Snyder, D. J. (1991). A critical appraisal of demand artifacts in consumer research. *Journal of Consumer Research, 18*, 273–283.

Shimp, T. A., Hyatt, E. M., & Snyder, D. J. (1993). A critique of Darley and Lim's ''alternative perspective''. *Journal of Consumer Research, 20*, 496–501.

Simonsohn, U. (2013). *Evaluating replication results.* Retrieved from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2259879

Stuart, E. W., Shimp, T. A., & Engle, R. W. (1987). Classical conditioning of consumer attitudes: Four experiments in an advertising context. *Journal of Consumer Research, 14*, 334–351.

Received February 28, 2013
Accepted December 21, 2013
Published online May 19, 2014

Ivar Vermeulen

Faculty of Social Sciences
Communication Science
VU University Amsterdam
De Boelelaan 1081
1081 HV Amsterdam
The Netherlands
Tel. +31 20 598-9190
E-mail i.e.vermeulen@vu.nl

What Does It Take to Activate Stereotypes? Simple Primes Don't Seem Enough

A Replication of Stereotype Activation (Banaji & Hardin, 1996; Blair & Banaji, 1996)

Florian Müller and Klaus Rothermund

Friedrich-Schiller-Universität Jena, Germany

Abstract. According to social cognition textbooks, stereotypes are activated automatically if appropriate categorical cues are processed. Although many studies have tested effects of activated stereotypes on behavior, few have tested the process of stereotype activation. Blair and Banaji (1996) demonstrated that subjects were faster to categorize first names as male or female if those were preceded by gender congruent attribute primes. The same, albeit smaller, effects emerged in a semantic priming design ruling out response priming by Banaji and Hardin (1996). We sought to replicate these important effects. Mirroring Blair and Banaji (1996) we found strong priming effects as long as response priming was possible. However, unlike Banaji and Hardin (1996), we did not find any evidence for automatic stereotype activation, when response priming was ruled out. Our findings suggest that automatic stereotype activation is not a reliable and global phenomenon but is restricted to more specific conditions.

Keywords: stereotype activation, response priming, semantic priming, replication

The idea that stereotypes are activated automatically upon encountering a member of a social category is taken for granted by many researchers and has also found its way into standard Social Psychology textbooks (Fiske, 1998; Schneider, 2004). This notion is by no means new, as Gilbert and Hixon (1991) point out:

> Many theorists have assumed that the activation of a stereotype is an automatic and inevitable consequence of encountering the object of that stereotype. For instance, Allport (1954, p. 21) argued that "Every event has certain marks that serve as a cue to bring the category of prejudgment into action.... A person with dark brown skin will activate whatever concept of Negro is dominant in our mind." (p. 509)

More specifically, activation of stereotypes and their subsequent impact on judgment and behavior can be conceptualized as a three-step process (Fiske, 1998; Moskowitz, Li, & Kirk, 2004; Schneider, 2004). First, a person is categorized as a member of a social group. Second, traits associated with this category are activated. And, third, judgment of and behavior toward the target person are influenced by these activated traits. While not immune to moderating factors (Blair, 2002) this sequence is thought to proceed in an automatic fashion largely beyond control (Bargh, Chen, & Burrows, 1996; Devine, 1989).

Indeed a considerable number of studies have shown effects of stereotype activation on attitudes and evaluations (Banaji & Greenwald, 1995; Banaji, Hardin, & Rothman, 1993; Devine, 1989; Dovidio, Evans, & Tyler, 1986; Lepore & Brown, 1997), thus addressing the final phase of this three-step model.

However, only few studies investigated the second step, the very process of stereotype activation itself. Gilbert and Hixon (1991) found an increase in participants' stereotype congruent word completions after perceiving a member of a stereotyped category. However, the considerable delay between category activation and word completion does not rule out controlled processes (Wentura & Rothermund, in press). Perdue and Gurtman (1990) found that participants were faster evaluating negative trait words if those were preceded by an old prime (vs. a young prime), whereas they were faster evaluating positive trait words if those were preceded by a young prime (vs. old prime). While this might

DOI: 10.1027/1864-9335/a000183

be taken as evidence for an automatic activation of stereotypic traits, the results can also be explained by response priming based on the valence dimension inherent in both primes and targets (Wentura & Degner, 2010).

Finally, in frequently cited studies, Blair and Banaji (1996) and Banaji and Hardin (1996) investigated activation of stereotypes. Blair and Banaji (1996) found that first name targets were categorized faster as being male or female if those were preceded by gender congruent attribute primes (Exps. 1 & 2). In a similar vein, Banaji and Hardin (1996) showed that participants were faster to categorize pronouns as male or female if they were preceded by gender congruent primes (Exp. 1). This effect held even if response priming was ruled out (Exp. 2).

While these influential studies are among the few that actually investigated the very process of stereotype activation, a few issues deserve mentioning. First, priming effects in Blair and Banaji (1996) were strongest for "non-trait" words; that is, attributes related to gender by designating typically male or female activities, roles, objects, or professions (e.g., ballet, master, flowers, mechanic). Priming effects were not significant for stereotypic trait attributes (e.g., courageous, logical, sentimental, warm).

Second, the priming effects in Blair and Banaji (1996) are prone to an alternative explanation in terms of response priming (Wentura & Degner, 2010). Thus, congruency effects in a gender categorization task can be explained by the fact that *both* primes and targets can be categorized as typically male or female, leading to response facilitation or interference in case of congruent or incongruent prime/target pairs, without having to assume an automatic spreading of activation from primes to targets. This issue was addressed in Banaji and Hardin (1996, Exp. 2). Participants were asked to categorize targets as pronouns or non-pronouns, which rules out response priming as an explanation. Only small effects were found, with priming effects present only for attribute primes related to gender by definition (e.g., mother, father), but not for attribute primes related to gender by normative base-rates (e.g., secretary, doctor). Considering the findings by Blair and Banaji (1996) it appears reasonable to assume that trait attributes are even less likely to produce these effects.

Given the implications and the importance of the discussed studies, it appears paramount to test (a) whether category activation via attribute priming can be replicated (Blair & Banaji, 1996), (b) whether those effects encompass attribute primes related to gender not only by definition, and (c) whether priming effects still hold if response priming is ruled out (Banaji & Hardin, 1996).

To accomplish these goals we conducted a study consisting of two experiments. The aim of the first experiment was a direct replication of the gender congruency effects reported by Blair and Banaji (1996). The aim of the second experiment was a replication of the gender congruency effects if response priming is ruled out (Banaji & Hardin, 1996). Here the gender categorization task was replaced by a name vs. town categorization task. This semantic priming design rules out response priming, a replication of Banaji and Hardin (1996, Exp. 2). This condition provides the crucial test of the assumed automatic gender stereotype activation effect in the absence of response priming. Both experiments were combined in a within-subjects design, with order of experiments counterbalanced across participants.

Overview

The goal of the current study was to replicate the finding that activation of stereotypically male or female attributes facilitates processing of targets denoting category membership (Banaji & Hardin, 1996; Blair & Banaji, 1996). To disentangle stereotype activation from response priming effects, each participant completed two tasks. A gender classification task (male vs. female names) corresponding to the Banaji and Hardin (1996, Exps. 1 & 2) study, and a semantic classification task (name vs. town) that was orthogonal to gender, corresponding to the semantic priming design used by Banaji and Hardin (1996, Exp. 2). The variation in the target task constitutes the within-subject factor *Task Type* (gender categorization vs. semantic catgorization). The complete study was implemented using the Psychopy software package[1] (Peirce, 2007, 2009) and run on standard PC hardware (Microsoft Windows XP) connected to a 17″ CRT monitor displaying XGA resolution at 85 Hz.

Method

Sample

In order to detect *Prime* × *Target* interaction effects from Banaji and Hardin (1996) and Blair and Banaji (1996) with a power $(1 - \beta)$ of .95 a sample of $N = 135$ was needed for each experiment.[2] To ensure sufficient sample size in case that order effects emerge (rendering only the experiment run first for each participant available for analyses) and to guard against dropout a total of 300 participants were recruited. Six participants who did not finish the experiment were excluded from further analyses, resulting in a final sample size of $N = 294$, 51% female, age: $M(SD) = 23.84$ (5.9), range: 18–55.

[1] Version 1.75.01, http://www.psychopy.org

[2] Calculation of the necessary sample size was based on the smallest effect that was reported in the original studies, which was found for the Prime Gender × Target Gender interaction in Banaji and Hardin (1996, Exp. 2) with $F(1, 56) = 4.63$. The sample size that is necessary to detect an effect of this size with $\alpha = \beta = .05$ was computed with G*Power 3.1.5.1 (Faul, Erdfelder, Lang, & Buchner, 2007).

Procedure

Upon arrival at the laboratory participants were seated at a computer in individual, soundproof cubicles. They were told to follow the instructions provided to them on the computer screen and to contact the experimenter if they had any questions. Participants learned that they were going to work on two different tasks, each requiring a binary categorization of words via button press. A detailed description was provided at the beginning of each experimental task, followed by a short practice block (eight trials) to familiarize participants with the upcoming task. To ensure fast and correct responses the practice block was repeated if participants' mean reaction time exceeded 1,000 ms or if their accuracy was below 80%. To guard against order effects, sequence of tasks was counterbalanced, constituting the between-subject factor *Experiment Order* (gender categorization first vs. semantic categorization first). Additionally, key assignment in both tasks was counterbalanced. Both tasks differed only with respect to the employed target stimuli and the required response.

Gender Categorization Task (GCT)

After a fixation cross (500 ms) participants were presented with a stereotypically male or female attribute prime (200 ms). After an Inter-Stimulus-Interval (ISI) of 100 ms (blank screen; SOA = 300 ms), a male or female first name was presented as target stimulus until a response was registered. Participants had to indicate as quickly and accurately as possible whether the name was female or male by pressing the assigned button on the keyboard, with button assignment counterbalanced across participants.

Semantic Categorization Task (SCT)

The procedure was identical to the GCT, with the following exception: After the ISI, either a first name (25% male, 25% female) or the name of a well-known city (50%) was presented as a target stimulus until a response was registered. Participants had to indicate as quickly and accurately as possible whether the target was the name of a person (regardless of gender) or the name of a town by pressing the assigned button on the keyboard. Again, button assignment was counterbalanced across participants.

Materials

Primes

A total of 62 prime words were used, one half of them being stereotypically male, the other half being stereotypically female. Of those 54 comprised male or female stereotypes counterbalanced on valence (positive, neutral, negative) and word type (noun, verb, adjective), with three exemplars representing each combination. The remaining eight primes were related to gender by definition and were adapted from Banaji and Hardin (1996). Their inclusion offers an additional test whether our procedure in general is suited to detect priming effects (see Table 1).

Table 1. Overview of gender-related words used as primes

Gender	Valence	Wordtype	Exemplars (translated)	Exemplars (original)
male	positive	noun	computer, exercise, muscles	Computer, Sport, Muskeln
		verb	to repair, to build, to protect	reparieren, bauen, beschützen
		adjective	brave, fearless, strong	tapfer, mutig, stark
	neutral	noun	skat, regular's table, car	Skat, Stammtisch, Auto
		verb	to fight, to shave, to do math	kämpfen, rasieren, rechnen
		adjective	dominant, objective, big	dominant, sachlich, groß
	negative	noun	beer belly, bald head, sweat	Bierbauch, Glatze, Schweiss
		verb	to curse, to hit, to drink	fluchen, boxen, saufen
		adjective	aggressive, rude, brutal	aggressiv, grob, brutal
	by definition		man, father, brother, king	Mann, Vater, Bruder, König
female	positive	noun	ballet, flowers, family	Ballett, Blumen, Familie
		verb	to dance, to groom, to bake	tanzen, pflegen, backen
		adjective	empathetic, caring, affectionate	einfühlsam, fürsorglich, zärtlich
	neutral	noun	household, chit-chat, diet	Haushalt, Tratsch, Diät
		verb	to put on make-up, to cook, to sew	schminken, kochen, nähen
		adjective	sensitive, emotional, domestic	sensibel, emotional, häuslich
	negative	noun	fashion, kitchen, purse	Mode, Küche, Handtasche
		verb	to do dishes, to clean, to iron	spülen, putzen, bügeln
		adjective	gossipy, hysterical, touchy	geschwätzig, hysterisch, zickig
	by definition		woman, mother, sister, queen	Frau, Mutter, Schwester, Königin

Table 2. Overview of first names and city names uses as targets

Type	Word
first name, male	Achim, Albert, Bernhard, Dieter, Daniel, Frank, Fritz, Georg, Hans, Helmut, Horst, Jens, Jürgen, Klaus, Manfred, Markus, Norbert, Peter, Friedrich, Thomas, Ulrich, Volker, Wolfgang, Hannes, Ingo, Kurt, Hubert, Stefan, Florian, Thorsten, Alexander
first name, female	Annette, Angelika, Anja, Birgit, Claudia, Erna, Eva, Helga, Heike, Julia, Daniela, Karin, Katja, Kerstin, Monika, Petra, Renate, Sabine, Sandra, Silke, Susanne, Tanja, Ulrike, Kerstin, Hanna, Linda, Nina, Katharina, Carolin, Anna, Franziska
city name	Aachen, Augsburg, Berlin, Bonn, Dortmund, Dresden, Düsseldorf, Frankfurt, Freiburg, Hamburg, Jena, Hannover, Kassel, Kiel, Konstanz, Leipzig, Mainz, Marburg, München, Münster, Nürnberg, Rostock, Wiesbaden, Bottrop, Koblenz, Gladbach, Essen, Stuttgart, Bremen, Magdeburg, Chemnitz, Göttingen, Gera, Mannheim, Krefeld, Flensburg, Minden, Lübeck, Potsdam, Darmstadt, Halle, Karlsruhe, Weimar, Erfurt, Rostock, Bielefeld, Duisburg, Paderborn, Cottbus, Bochum, Braunschweig, Ahlen,Würzburg, Wolfsburg, Rudolstadt, Fulda, Suhl, Ulm, Heidelberg, Wuppertal, Regensburg, Passau

Targets

In the GCT, a total of 62 first names were used, 50% male and 50% female. Care was taken to only select names that are easily and unambiguously recognized as male or female. In the SCT, an additional 62 city names were employed (see Table 2).

Trials

In the GCT, each prime was randomly paired with a male and a female target name, yielding a total of 62 primes × 2 names = 124 trials. In the SCT, each prime was additionally paired with two city names, yielding a total of 248 trials.

Known Differences From Original Studies

In the current experiments we opted to exclude control or nonword primes, for two reasons: First, those primes are not essential for the focal research question, in fact those trials are not part of any analyses of interest. Second, being able to use all completed trials for the analysis increases the power to detect even small effects.

Also we employed a new set of prime and target stimuli. This was mainly because the current experiments were conducted in Germany and the original studies are now 17 years old. Thus it was questionable whether our subjects would endorse the original stereotypic stimuli to the same extent as the original subjects. As in Blair and Banaji (1996), attributes were balanced with regard to valence, with an additional set of attribute stimuli of neutral valence. Also we systematically varied the word type of the attributes to estimate whether effects vary depend on the type of attributes (adjectives = personality traits, nouns and verbs refer to behavior and appearance).

Finally, we opted to employ the aforementioned name vs. town classification task instead of the original pronoun versus non-pronoun classification (Banaji & Hardin, 1996). This is because the latter allows only a limited set of different target stimuli (*he, she, it, me vs. is, do, as, all* in Banaji & Hardin, 1996) resulting in a huge number of target repetitions during the task. Employing the name versus town classification allowed us to eliminate target repetitions completely.

Results

While all trials from the GCT entered into the analyses, only trials featuring a first name target in the SCT could be analyzed with regard to stereotype congruency. Thus, each experiment yielded a total of 124 trials for the analyses. RTs from trials with incorrect responses (4.8%) or exceeding the third quartile of the respective intraindividual distribution by more than 1.5 interquartile ranges (4.3%; outlier values according to Tukey, 1977) were removed from the analyses.

Including the control factors *Experiment Order* and *Participant Gender* did not reveal any higher order interactions with the priming factors (all *p*'s > .20). We thus report findings in which we aggregate across conditions for these balancing factors. Average RTs within conditions were subjected to a 2 (*Prime Gender*: male vs. female) × 2 (*Target Gender*: male vs. female) × 2 (*Experiment Type*: GCT vs. SCT) repeated measures ANOVA, yielding a significant *Prime Gender* × *Target Gender* interaction, $F(1, 293) = 39.68$, $p = 1.09 \times 10^{-9}$, $\eta_p{}^2 = 0.12$, that was further qualified by the three-way interaction of *Prime Gender* × *Target Gender* × *Experiment Type*, $F(1, 293) = 25.74$, $p = 6.95 \times 10^{-7}$, $\eta_p{}^2 = 0.08$. Inspection of Figure 1 reveals that responses to targets are facilitated by gender congruent primes only in the GCT but not in the SCT. Conducting separate ANOVAs for each experimental task confirmed a significant *Prime Gender* × *Target Gender* interaction in the GCT, $F(1, 293) = 75.54$, $p = 2.59 \times 10^{-16}$, $\eta_p{}^2 = 0.20$, indicating that responses to male and female names were faster after stereotypically congruent primes ($M = 551$ ms) than after stereotypically incongruent primes ($M = 564$ ms). No such congruency effect was obtained for the SCT, $F(1, 293) = 1.03$, $p = .31$, $\eta_p{}^2 = 0.003$, indicating

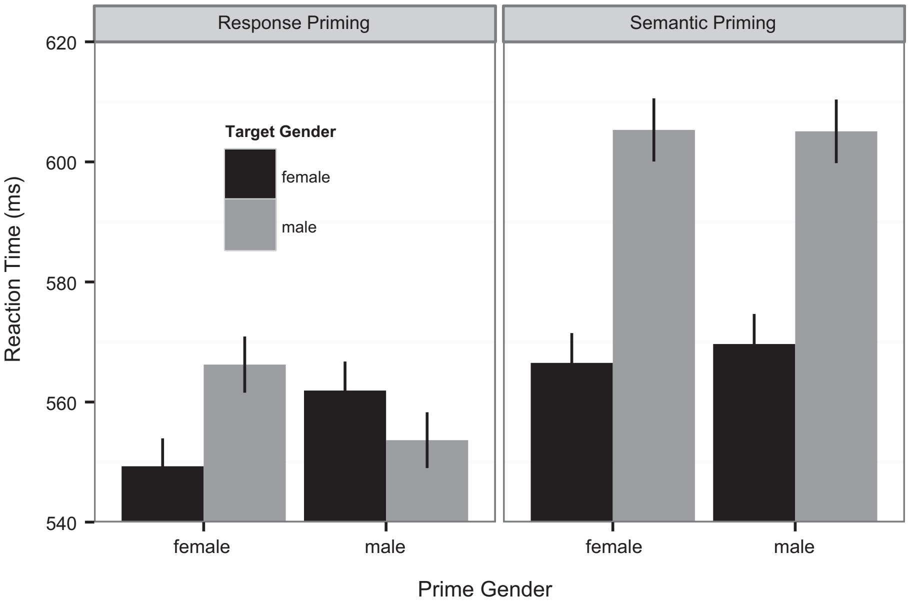

Figure 1. Reaction Times depending on *Prime Gender* and *Target Gender*. Priming effects were only found in the response priming condition (GCT).

that prime stereotypicality did not affect responding to male and female names in a semantic categorization task unrelated to gender.

To ensure the robustness of our results, analyses were repeated using log-transformed values (after removal of outliers as mentioned above) and by using values that were standardized on participants individual standard deviation of response latencies (i.e., using the *D* measure, see Greenwald, Nosek, & Banaji, 2003, p. 201). The same pattern of findings emerged for these analyses as well (for details consult supplementary materials).

Effects of Prime Valence

Including the factor *Valence* (negative, neutral, positive) revealed no interactions of Valence with stereotype priming, all *p*'s > .13 (Greenhouse Geisser corrected).

Effects of Word Type

Including the factor *Word Type* (noun, verb, adjective) revealed a significant four-way interaction of all factors, $F(2, 586) = 3.65$, $p = 0.03$, $\eta_p^2 = 0.01$. Separate ANOVA's for each experimental task confirmed a significant *Prime Gender* × *Target Gender* × *Word Type* interaction in the GCT, $F(2, 586) = 5.21$, $p = .006$, $\eta_p^2 = 0.02$, but not in the SCT, $F(2, 586) = 0.36$, $p = .69$ (Greenhouse Geisser corrected). ANOVAs conducted for each value of *Word Type* in the GCT yielded a more pronounced *Prime Gender* × *Target Gender* interaction for nouns, $F(1, 293) = 48.25$, $p = 2.43 \times 10^{-11}$, $\eta_p^2 = .14$, versus verbs, $F(1, 293) = 9.49$, $p = .002$, $\eta_p^2 = .03$, and adjectives, $F(1, 293) = 15.64$, $p = 9.61 \times 10^{-5}$, $\eta_p^2 = .05$.

Effects of Stereotype Relation

Including the factor *Stereotype Relation* (stereotypically related primes vs. gender-defining primes) revealed a significant four-way interaction of all factors, $F(1, 293) = 15.2$, $p = 1.19 \times 10^{-4}$, $\eta_p^2 = 0.05$. Separate ANOVAs for each experimental task confirmed a significant *Prime Gender* × *Target Gender* × *Word Type* interaction in the GCT, $F(1, 293) = 24.97$, $p = 1 \times 10^{-6}$, $\eta_p^2 = 0.08$, but not in the SCT, $F(1, 293) = 0.59$, $p = .44$. Finally, ANOVAs conducted for each value of *Stereotype Relation* within the GCT revealed a more pronounced *Prime Gender* × *Target Gender* interaction for primes related to gender by definition, $F(1, 293) = 67.0$, $p = 8.42 \times 10^{-15}$, $\eta_p^2 = 0.19$, compared to primes that were part of a gender stereotype, $F(1, 147) = 50.41$, $p = 9.43 \times 10^{-12}$, $\eta_p^2 = 0.15$.

Discussion

The current study sought to replicate the finding that activation of stereotypically male or female attributes facilitates processing of targets denoting category membership (Banaji & Hardin, 1996; Blair & Banaji, 1996). The response priming paradigm in a gender classification task successfully replicated the effects reported by Blair and Banaji (1996). However, these effects are readily explained by a response priming mechanism and thus do not provide strong support for claims of an automatic stereotype activation. If automatic stereotype activation does take place effects should also be obtained with a semantic priming paradigm. In contrast to the findings reported by Banaji and Hardin (1996), our semantic priming paradigm

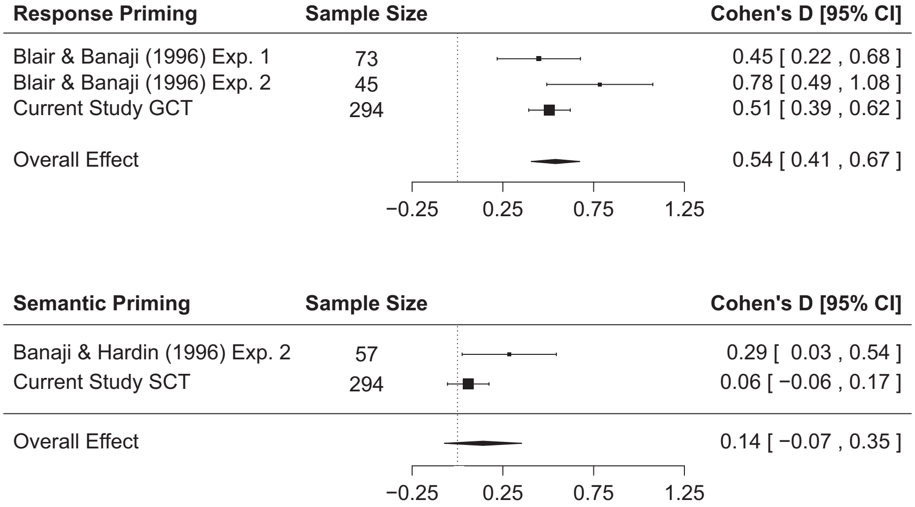

Figure 2. Priming Effects compared across studies depending on type of priming.

Table 3. Mean (*SD*) of reaction time (ms) for each condition separately for each experiment (*N* = 294)

		Prime Gender	
Experiment type	Target gender	Male	Female
GCT	male	554 (80)	566 (80)
	female	562 (83)	549 (80)
SCT	male	605 (91)	605 (90)
	female	570 (86)	567 (85)

Notes. GCT = Gender Categorization Task; SCT = Semantic Categorization Task.

(Semantic categorization task: names vs. cities) revealed no trace of stereotype congruency effects (see Figure 2 for a comparison of the observed effects).

Because our experiment had sufficient power ($1 - \beta > .95$) to detect effects of the magnitude reported by Banaji and Hardin (1996), this failure to replicate the original results cannot be attributed easily to insufficient sample size. These results might be less surprising though if one considers that Banaji and Hardin (1996) also found only small priming effects and these effects were limited to those words that relate to gender by definition (e.g., mother, king, sister). This is reminiscent of similar findings from the domain of affective priming: Affective congruency effects typically do not emerge in semantic priming designs if effects of response priming are controlled (e.g., De Houwer, Hermans, Rothermund, & Wentura, 2002; Eder, Leuthold, Rothermund, & Schweinberger, 2012; Klauer & Musch, 2002; Klinger, Burton, & Pitts, 2000; Voss, Rothermund, Gast, & Wentura, 2013; Werner & Rothermund, 2013).

However, the failure of finding global semantic activation effects for stereotypic primes does not rule out the possibility that stereotypes can be activated automatically under more specific conditions. Recent experiments using semantic priming paradigms support the notion that the activation of specific subsets of stereotypic attributes does take place if category and context information are combined in a compound prime (Casper, Rothermund, & Wentura, 2010, 2011). This context-dependent activation of stereotypes also was confirmed in studies on self-stereotyping (Casper & Rothermund, 2012) and on the activation of stereotype-related behaviors (Müller & Rothermund, 2012).

Acknowledgments

We report all data exclusions, manipulations, and measures, and how we determined our sample sizes. The research reported in this article was supported by Grant No. 2013-011 from the Center for Open Science. The authors declare no conflict-of-interest with the content of this article. Designed research: F. M., K. R.; Performed research: F. M.; Analyzed data: F. M., K. R.; Wrote paper: F. M., K. R. All materials, data, video of the procedure, and the preregistered design are available at https://osf.io/tv8ai/.

PREREGISTERED OPEN DATA OPEN MATERIALS

References

Allport, G. W. (1954). *The nature of prejudice*. Oxford, UK: Addison-Wesley.

Banaji, M. R., & Greenwald, A. G. (1995). Implicit gender stereotyping in judgments of fame. *Journal of Personality and Social Psychology, 68*, 181–198.

Banaji, M. R., & Hardin, C. D. (1996). Automatic stereotyping. *Psychological Science, 7*, 136–141.

Banaji, M. R., Hardin, C., & Rothman, A. J. (1993). Implicit stereotyping in person judgment. *Journal of Personality and Social Psychology, 65*, 272–281.

Bargh, J. A., Chen, M., & Burrows, L. (1996). Automaticity of social behavior: Direct effects of trait construct and stereotype activation on action. *Journal of Personality and Social Psychology, 71*, 230–244.

Blair, I. V. (2002). The malleability of automatic stereotypes and prejudice. *Personality and Social Psychology Review, 6*, 242–261.

Blair, I. V., & Banaji, M. R. (1996). Automatic and controlled processes in stereotype priming. *Journal of Personality and Social Psychology, 70*, 1142–1163.

Casper, C., & Rothermund, K. (2012). Gender self-stereotyping is context-dependent for men but not for women. *Basic and Applied Social Psychology, 34*, 434–442.

Casper, C., Rothermund, K., & Wentura, D. (2010). Automatic stereotype activation is context dependent. *Social Psychology, 41*, 131–136. doi: 10.1027/1864-9335/a000019

Casper, C., Rothermund, K., & Wentura, D. (2011). The activation of specific facets of age stereotypes depends on individuating information. *Social Cognition, 29*, 393–414.

De Houwer, J., Hermans, D., Rothermund, K., & Wentura, D. (2002). Affective priming of semantic categorization responses. *Cognition and Emotion, 16*, 643–666.

Devine, P. G. (1989). Stereotypes and prejudice: Their automatic and controlled components. *Journal of Personality and Social Psychology, 56*, 5–18.

Dovidio, J. F., Evans, N., & Tyler, R. B. (1986). Racial stereotypes: The contents of their cognitive representations. *Journal of Experimental Social Psychology, 22*, 22–37.

Eder, A., Leuthold, H., Rothermund, K., & Schweinberger, S. R. (2012). Automatic response activation in sequential affective priming: An ERP study. *Social Cognitive and Affective Neuroscience, 7*, 436–445.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*, 175–191.

Fiske, S. T. (1998). Stereotyping, prejudice, and discrimination. In D. T. Gilbert, S. T. Fiske, & G. Lindzey (Eds.), *The handbook of social psychology* (4th, Vol. 1 & 2, pp. 357–411). New York, NY: McGraw-Hill.

Gilbert, D. T., & Hixon, J. (1991). The trouble of thinking: Activation and application of stereotypic beliefs. *Journal of Personality and Social Psychology, 60*, 509–517.

Greenwald, A. G., Nosek, B. A., & Banaji, M. R. (2003). Understanding and using the Implicit Association Test: I. An improved scoring algorithm. *Journal of Personality and Social Psychology, 85*, 197–216.

Klauer, K. C., & Musch, J. (2002). Goal-dependent and goal-independent effects of irrelevant evaluations. *Personality and Social Psychology Bulletin, 28*, 802–814.

Klinger, M. R., Burton, P. C., & Pitts, G. S. (2000). Mechanisms of unconscious priming: I. Response competition, not spreading activation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*, 441–455.

Lepore, L., & Brown, R. (1997). Category and stereotype activation: Is prejudice inevitable? *Journal of Personality and Social Psychology, 72*, 275–287.

Müller, F., & Rothermund, K. (2012). Talking loudly but lazing at work: Behavioral effects of stereotypes are context dependent. *European Journal of Social Psychology, 42*, 557–563.

Moskowitz, G. B., Li, P., & Kirk, E. R. (2004). The Implicit Volition Model: On the preconscious regulation of temporarily adopted goals. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 36, pp. 317–413). San Diego, CA: Elsevier Academic Press.

Peirce, J. W. (2007). PsychoPy – Psychophysics software in Python. *Journal of Neuroscience Methods, 162*, 8–13.

Peirce, J. W. (2009). Generating stimuli for neuroscience using PsychoPy. *Frontiers in Neuroinformatics, 2*, 1–8.

Perdue, C. W., & Gurtman, M. B. (1990). Evidence for the automaticity of ageism. *Journal of Experimental Social Psychology, 26*, 199–216.

Schneider, D. J. (2004). *The psychology of stereotyping*. New York, NY: Guilford Press.

Tukey, J. W. (1977). *Exploratory data analysis*. Reading, UK: Addison-Wesley.

Voss, A., Rothermund, K., Gast, A., & Wentura, D. (2013). Cognitive processes in associative and categorical priming: A diffusion model analysis. *Journal of Experimental Psychology: General, 142*, 536–559.

Wentura, D., & Degner, J. (2010). A practical guide to sequential priming and related tasks. In B. Gawronski & B. K. Payne (Eds.), *Handbook of implicit social cognition: Measurement, theory, and applications* (pp. 95–116). New York, NY: Guilford.

Wentura, D., & Rothermund, R. (in press). Priming is not priming is not priming. *Social Cognition*.

Werner, B., & Rothermund, K. (2013). Attention please: No affective priming effects in a valent/neutral-categorization task. *Cognition and Emotion, 27*, 119–132.

Received March 4, 2013
Accepted January 24, 2013
Published online May 19, 2014

Florian Müller

Friedrich-Schiller-Universität Jena
Institut für Psychologie
Am Steiger 3/Haus 1
07743 Jena
Germany
Tel. +49 3641 945123
E-mail florian.mueller@uni-jena.de

A Replication Attempt of Stereotype Susceptibility (Shih, Pittinsky, & Ambady, 1999)

Identity Salience and Shifts in Quantitative Performance

Carolyn E. Gibson, Joy Losee, and Christine Vitiello

Georgia Southern University, Statesboro, GA, USA

Abstract. Awareness of stereotypes about a person's in-group can affect a person's behavior and performance when they complete a stereotype-relevant task, a phenomenon called stereotype susceptibility. Shih, Pittinsky, and Ambady (1999) primed Asian American women with either their Asian identity (stereotyped with high math ability) or female identity (stereotyped with low math ability) or no priming before administering a math test. Of the three groups, Asian-primed participants performed best on the math test, female-primed participants performed worst. The article is a citation classic, but the original studies and conceptual replications have low sample sizes and wide confidence intervals. We conducted a replication of Shih et al. (1999) with a large sample and found a significant effect with the same pattern of means after removing participants that did not know the race or gender stereotypes, but not when those participants were retained. Math identification did not moderate the observed effects.

Keywords: stereotype susceptibility, priming, group identity, Asian American women, replication, stereotype threat

Stereotype susceptibility is a phenomenon in which awareness of stereotypes about a person's in-group and other out-groups affects a person's behavior and performance on tasks related to the stereotype (Shih, Pittinsky, & Ambady, 1999). Negative stereotypes about a person's in-group can hinder performance (stereotype threat; Steele & Aronson, 1995) and positive stereotypes about a person's in-group can facilitate performance (stereotype boost; Shih, Ambady, Richeson, Fujita, & Gray, 2002). Shih et al. (1999) found that Asian American women performed better on a mathematics test when their ethnic identity was made salient compared to a control group. In contrast, Asian American women performed worse on a mathematics test when their gender identity was made salient compared to a control group.

Studies have shown that stereotype susceptibility can affect other behavior besides performance on a test, such as learning, self-control, aggressive behavior, decision-making (Inzlicht & Kang, 2010), and gender diversity in STEM fields (Shapiro & Williams, 2012; Tsui, 2007). Stereotype susceptibility could have broad implications, but some studies have shown mixed results in real-world applications (Stricker & Ward, 2004; Wei, 2012). Stoet and Geary (2012) argued that there is a mismatch between the empirical evidence for the effect and the certainty reported in academic citations of that evidence. For example, studies of stereotype susceptibility tend to have small sample sizes and wide confidence intervals (Stricker & Ward, 2004; see supplements for summary table). In novel areas of research, low power ironically can increase *both* Type I and Type II errors (Button et al., 2013) even though power is traditionally associated with Type II error risk alone. Shih and colleagues' (1999) seminal paper has had substantial impact (760 citations; Google Scholar, January 12, 2014), but used small samples. Thus, given its importance and potential applicability, we sought to replicate the original Shih et al. result to improve confidence in the effect.

Constraints on Stereotype Susceptibility Effects

Stereotype Awareness

Explicit awareness of a stereotype is an important factor for stereotype susceptibility effects (Appel, Kronberger, & Aronson, 2011; Inzlicht & Kang, 2010). Based on the suggestion by Shih et al. (1999) that their second study on a Canadian sample of recently immigrated Asian women

DOI: 10.1027/1864-9335/a000184

did not find significant results with the hypothesis that participants were not aware of gender and racial stereotypes about math, we included a measure of stereotype awareness as a basis for exclusion. This measure was adapted from Inzlicht and Kang (2010) who also measured stereotype awareness as a basis for exclusion.

Math Identification

Though it was not part of the original demonstration, subsequent research suggests that domain identification may moderate stereotype susceptibility effects (Shih, Pittinsky, & Ho, 2012; Shih, Pittinsky, & Trahan, 2006). However, Smith and Johnson (2006) observed evidence that strong domain identification may not lead to stereotype boost in response to positive stereotypes. Shih et al. (2012) suggested that further research with domain identification is needed in order to understand how this moderator could affect stereotype susceptibility. As such, in addition to a faithful replication of the original procedure, we added a measure of math identification to test whether the effect is more likely to be observed among the highly identified.

Method

Participants

A total of 164 Asian Female college students participated in this study, with approximately 52 in each condition so as to detect a medium effect size ($r = .35$; Shih et al., 1999, p. 81) with 80% power (Cohen, 1992). Participants were recruited from six universities in the southeast United States. Participants completed the study either individually in a laboratory setting ($n = 19$) or at student centers ($n = 147$). The student centers were relatively quiet and only people sitting alone were invited to participate. So that the recruited Asian women would not be sensitized to having been selected on the basis of that identity, experimenters also asked others nearby to participate. Data collected from non-Asian women were discarded. Experimenters were all White women.

Our registered sampling plan was to preselect Asian American women who were aware of Asian and female stereotypes about math. However, we could not access the necessary sample size through the original data collection plan. So, we altered the sampling plan by adding schools and eliminating the preselection process. Instead, demographic and stereotype awareness measures were included as a basis for exclusion post data collection. Additionally, to maintain consistency with the in-lab setting, distraction during the course of the math test was added as basis for exclusion prior to observing the results. We collected data from 168 participants. Ten participants were excluded: six for distraction, two because they had completed the test at a previous time, and two because they did not speak enough English to complete the survey. Thus, 158 people comprised the sample, consistent with the original sampling plan.

Of the 158 participants, 32 participants were not aware of either or both of the stereotypes of Asians being better than Caucasians and men being better than women at math. Thus, for tests limited to participants aware of the stereotypes, only 127 participants were included.

Design and Materials

Design and materials are the same as the materials used in Shih et al. (1999). This included an identity prime, a 12-item math test with questions from the Canadian Math Competition, and a manipulation check with questions about enjoyment, the perceived difficulty of the test, and hypotheses of what the experiment was assessing. We added a math identification questionnaire with 16 questions (Smith & White, 2001; e.g., "Math is one of my best subjects") with a response scale from 1 = Strongly Disagree to 5 = Strong Agree. All participants also completed a 7-item stereotype awareness survey assessing awareness of cultural stereotypes including men are better than women at math and Asians are better than Caucasians at math (adapted from Inzlicht & Kang, 2010). All materials and data are available on the Open Science Framework (http://osf.io/8q769).

Procedure

For in-lab recruitment, participants received the stereotype awareness survey via e-mail. Those reporting awareness of both stereotypes received a second e-mail inviting them to the laboratory. For out-of-lab recruitment, participants were approached in common areas and asked if they would like to take a math test for $5 and a candy bar. Participants were also offered a chance to enter a drawing to win $500. After agreeing to participate, an experimenter provided the informed consent. Study packets were randomized prior to data collection and began with a blank paper on top that assured the experimenter's blindness to condition. Participants first completed the manipulation making their Asian or female identity salient (or control). In the female-priming condition, participants ($n = 54$) answered questions concerning coed versus single sex living arrangements. In the Asian-priming condition, participants ($n = 52$) responded about their family and ethnic involvement. Those in the control condition ($n = 52$) answered four questions about their lives unrelated to gender or ethnicity (e.g., how often they eat out).

Participants then completed the 12-question math test from the original study and were told they would have 20 min to finish it. After the math test, participants answered a math identification questionnaire, a stereotype awareness survey (except for in-lab participants), and a final set of questions including a manipulation check, participants' self-reported math skill, enjoyment, and difficulty of the test. These latter items were not analyzed for the confirmatory report, but summary analyses are available in supplementary materials.

The original study had not included a manipulation check, but we added one to see if participants had knowledge

Table 1. Means and standard deviations for the linear comparison between the replication and original. Standard deviations are in parentheses. Means for accuracy and number of correct responses are included

	Asian		Control		Female	
	N	Mean	*N*	Mean	*N*	Mean
Current *N* = 158 accuracy	52	0.59 (0.23)	52	0.55 (0.26)	54	0.53 (0.23)
Current *N* = 127 aware only, accuracy	40	0.63 (0.22)	44	0.55 (0.25)	43	0.51 (0.23)
Moon and Roeder *N* = 139 accuracy	53	0.46 (0.17)	48	0.50 (0.18)	38	0.43 (0.16)
Moon and Roeder *N* = 106 aware only, accuracy	42	0.47 (0.18)	37	0.50 (0.17)	27	0.43 (0.16)
Shih et al. (1999) accuracy	16	0.54 (0.17)	16	0.49 (0.20)	14	0.43 (0.16)
Current *N* = 158 correct responses	52	6.52 (2.73)	52	5.75 (2.80)	54	5.72 (2.57)
Current *N* = 127 aware only, correct responses	40	6.93 (2.68)	44	5.73 (2.60)	43	5.60 (2.66)
Moon and Roeder *N* = 139 correct responses	53	4.75 (2.16)	48	5.21 (1.91)	38	4.50 (1.96)
Moon and Roeder *N* = 106 aware only, correct responses	42	4.83 (NR)	37	5.19 (NR)	27	4.30 (NR)
Shih et al. (1999) correct responses	16	5.37 (NR)	16	5.31 (NR)	14	4.71 (NR)

of the theme of the questions in the manipulation that preceded the math test. Many of the participants did not pass the manipulation check (34 people passed the manipulation check, 14 people did not answer, 108 did not pass). The theme of the manipulation questions was a subtle prime, so it is conceivable that people would not realize that there was a theme, but still be influenced by the prime.

Experimenters discreetly observed out-of-lab participants during the math test so as to ensure the conditions were similar to an in-lab session. Participants were excluded if they became distracted during the 20 min of the math test. Distraction criteria were defined by circumstances that would not normally happen in the laboratory. This included talking to another person and leaving the seat during the test. Experimenters were blind to condition and could not assess performance, thus these issues could not bias the decision to exclude participants.

Known Differences From Original Study

Above, we described exclusion criteria for stereotype awareness, and measurement of math identity. These were not included in the original research, but were identified as possible constraints on observing stereotype susceptibility effects in the original article or subsequent reports.

One of the differences between this replication and the original was the region in which it took place (Northeast/Harvard vs. Southeast/Emory, Georgia Tech, UGA, Armstrong-Atlantic, Georgia Southern, and UAB). There could be differences between these samples on average SAT scores and average GPAs that may be important for observing the effect.

Another difference is that the original study used an Asian female experimenter and this study used Caucasian female experimenters. However, Shih et al. (2002) used a Caucasian female experimenter and obtained effects similar to Shih et al. (1999). Even so, we minimized contact between participant and experimenter in case this could have an effect.

Like the original study, some of the participants were run in individual laboratory-based sessions, but most participants completed the study in quiet common areas around campus.

Results

Two separate analyses were conducted. First, per the registered sampling plan, we analyzed the data of 158 participants excluding only those unable to properly complete the experiment. Second, per the registered sampling plan, we analyzed the data of only the 127 participants that were aware of both stereotypes.

Primary Comparisons

Following the original study, data were submitted to a linear contrast analysis. This analysis tested the hypothesis that Asian-primed participants would have the best math performance, control participants would perform in the middle, and the female-primed participants would perform worst.

First, the analysis tested the effects of condition on accuracy. As in Shih et al. (1999), accuracy was calculated by dividing questions answered correctly by questions attempted (see Table 1). There was no significant difference between groups on accuracy, $t(155) = 1.31$, $p = .19$, $\eta^2 = .01$, 95% CI [.00, .06] (*Note*: contrast analysis effect sizes and confidence intervals are reported as η^2, and η^2 cannot be less than zero). Next, as in the original study, a two-tailed independent-samples *t*-test analyzed differences in accuracy between female-primed and Asian-primed conditions, but did not show a significant difference, $t(104) = 1.37$, $p = .18$, $d = .27$, 95% CI [−.11, .65].

A second analysis tested the effects of identity salience on accuracy, but included only participants who were aware of the race and gender stereotypes ($N = 127$). With only this subset, we observed a significant difference between groups on accuracy, $t(124) = 2.30$, $p = .02$, $\eta^2 = .04$, 95% CI [.01, .17], and the means followed the predicted pattern, Asian ($M = .63$), Control ($M = .55$), and Female ($M = .51$). Likewise, an independent samples *t*-test showed that female-primed participants performed significantly worse than Asian-primed participants, $t(81) = 2.40$, $p = .02$, $d = .53$, 95% CI [.09, .97]. There were no significant differences between the Asian-primed group and the control group, $t(82) = 1.51$, $p = .13$, $d = .33$, 95% CI [−.10, .76], or the female-primed group and the control group, $t(85) = .77$, $p = .44$, $d = .17$, 95% CI [−.25, .59].

Shih et al. (1999) also used total number of questions answered correctly as a dependent variable, but did not observe significant differences. In a linear contrast with the full sample, there was no significant difference between groups using this dependent variable, $t(155) = 1.52$, $p = .13$, $\eta^2 = .02$, 95% CI [.00, .07] (see Table 1). An independent-samples *t*-test showed no difference between Asian-primed and female-primed group on correct responses, $t(104) = 1.55$, $p = .13$, $d = .30$, 95% CI [−.08, .68].

Including only those aware of both stereotypes, there was a significant difference between groups on number of questions answered correctly, $t(124) = 2.27$, $p = .03$, $\eta^2 = .05$, 95% CI [.01, .17], and the means followed the expected pattern: Asian-primed ($M = 6.93$), Control ($M = 5.73$), and female-primed ($M = 5.60$). Asian-primed participants had significantly more correct responses than female-primed participants, $t(81) = 2.25$, $p = .03$, $d = .50$, 95% CI [.06, .94]. There was also a significant difference between the Asian-primed group and the control, $t(82) = 2.08$, $p = .04$, $d = .46$, 95% CI [.02, .89], but not between the female-primed group and the control, $t(85) = .22$, $p = .83$, $d = .05$, 95% CI [−.37, .47].

Math Identification as a Moderator

To analyze math identification as a moderator, an ANCOVA analyzed math identification as a covariate. Math identification did not interact with condition and is therefore not considered a moderator, $F(2, 151) = 1.63$, $p = .20$. Math identification was also not a moderator when analyzing the data only with participants aware of both stereotypes, $F(2, 120) = 2.06$, $p = .13$.

Discussion

Shih et al. (1999) found that priming Asian women's racial identity or gender identity led to somewhat better and worse math performance respectively, with a control condition in between. Their linear contrast ($N = 46$) elicited an effect size of $\eta^2 = .07$. We conducted a high-powered replication ($N = 156$) and found the same ordinal relations among the conditions, but a much smaller effect size of $\eta^2 = .01$ and a nonsignificant effect ($p = .18$). However, after excluding participants that were unaware of either the gender or racial stereotypes ($N = 127$) the effect was significant ($p = .02$) and the effect size was larger $\eta^2 = .04$. Beilock, Rydell, and McConnell (2007) determined that a causal mechanism of stereotype threat occurs when one is aware of the negative stereotype, such that it causes an explicit expectation of performance. We used this finding and the precedent set by Inzlicht and Kang (2010) to select participants based on their awareness of the stereotypes in question.

When the original and current study's effect sizes are combined, we estimate an effect size of $\eta^2 = .03$ combined with the full replication sample and $\eta^2 = .05$ combined with the reduced replication sample. Shih et al. (1999) included a second study, but they argued that the study ($N = 19$) was a comparison (Canadian sample) in which the effect was not expected to occur, so that study is not included in the aggregate result.

Caution

We replicated the Shih et al. (1999) effect after excluding participants who did not report awareness of the Asian or female stereotypes about math ability. However, most of the participants completed this stereotype measure at the end of the study. As a consequence, it is conceivable that the manipulation differentially affected responses to the stereotype awareness measure among low and high performers. While we perceive this as unlikely, additional research is needed to rule it out conclusively.

There are some curiosities in the present results that deserve further attention. For one, most participants failed the manipulation check suggesting that they were not aware of the priming theme. This could be news that the prime is influential outside of awareness, or it could be a cause of concern for the interpretability of the results. Had the results shown no evidence for an effect, the lack of effect on the manipulation check might have been taken as cautionary evidence against interpreting the results. We did not consider the manipulation check to be critical in our preregistered plan, but that does not prevent opportunistic interpretation and inflation of alpha levels once the results are known. This suggests that additional evidence for this effect in other contexts will be useful. Also notable is the fact that math identification did not moderate the result despite some prior suggestion that this could be a moderator. These observations suggest that the conditions necessary to obtain the present results are not yet well understood.

Finally, a second replication attempt using our registered protocol was conducted at University of California, Berkeley with a total of 139 participants (Moon & Roeder, 2014). In their study, the three priming groups did not significantly differ on accuracy or number of correct answers, and the pattern of means did not fit the hypothesis. Furthermore, analysis with only participants aware of the stereotypes showed no significant differences between groups. These observations suggest that the conditions necessary to obtain the stereotype susceptibility effect are not yet completely understood.

Conclusion

The present research provided some additional support for the hypothesis that priming Asian women with social identities associated with math stereotypes can influence their performance in mathematics. The results additional suggest that this effect is contingent on awareness of these stereotypes, but not on identification with mathematics. Conducted with a preregistered confirmatory design, these results provide additional empirical value for this finding, and research on stereotype threat and stereotype boost more generally. Studies examining stereotype threat in standardized testing have inconsistent results with some showing an effect of stereotype susceptibility with field methods (Cohen, Garcia, Apfel, & Master, 2006; Good, Aronson, & Inzlicht, 2003; Wei, 2012) and others not (Ganley, Mingle, Ryan, Ryan, & Vasilyeva, 2013; Stricker & Ward, 2004). Given the additional open questions about this effect, future studies may benefit from strong confirmatory designs to help clarify the conditions under which these effects can be observed.

Acknowledgments

The authors contributed equally to this article. Thus, authors are listed alphabetically. Designed research: C. G., J. L., C. V.; Performed Research: C. G., J. L., C. V.; Analyzed Data: C. G., J. L., C. V.; Wrote paper: C. G., J. L., C. V. The authors would like to thank Katharine Gaskin for her help collecting data, and Sean Mackinnon and Daniël Lakens for contributing expertise to analyses. The authors would also like to thank Amy Hackney and Brian Nosek for their helpful comments and support throughout the process. We report all data exclusions, manipulations, and measures, and how we determined our sample sizes. The authors declare no conflict-of-interest with the content of this article. The project was funded by a grant from the Center for Open Science. All materials, data, and the preregistered design are available at: http://osf.io/5itqa.

PREREGISTERED OPEN DATA OPEN MATERIALS

References

Appel, M., Kronberger, N., & Aronson, J. (2011). Stereotype threat impairs ability building: Effects on test preparation among women in science and technology. *European Journal of Social Psychology, 41*, 904–913.

Beilock, S. L., Rydell, R. J., & McConnell, A. R. (2007). Stereotype threat and working memory: Mechanisms, alleviations, spillover. *Journal of Experimental Psychology: General, 136*, 256–276.

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafo, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience, 14*, 1–12.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*, 155–159.

Cohen, G. L., Garcia, J., Apfel, N., & Master, A. (2006). Reducing the racial achievement gap: A social-psychological intervention. *Science, 313*(5791), 1307–1310.

Ganley, C. M., Mingle, L. A., Ryan, A. M., Ryan, K., & Vasilyeva, M. (2013). An examination of stereotype threat effects girls' mathematics performance. *Developmental Psychology, 49*, 1886–1897.

Good, C., Aronson, J., & Inzlicht, M. (2003). Improving adolescents' standardized test performance: An intervention to reduce the effects of stereotype threat. *Applied Developmental Psychology, 24*, 645–662.

Inzlicht, M., & Kang, S. K. (2010). Stereotype threat spillover: How coping with threats to social identity affects aggression, eating, decision making, and attention. *Journal of Personality and Social Psychology, 99*, 467–481.

Moon, A., & Roeder, S. S. (2014). A secondary replication attempt of stereotype susceptibility (Shih, Pittinsky, & Ambady, 1999). *Social Psychology*, 45.

Shapiro, J. R., & Williams, A. M. (2012). The role of stereotype threats in undermining girls' and women's performance and interest in STEM fields. *Sex Roles, 66*, 175–183.

Shih, M., Ambady, N., Richeson, J. A., Fujita, K., & Gray, H. M. (2002). Stereotype performance boosts: The impact of self-relevance and the manner of stereotype activation. *Journal of Personality and Social Psychology, 83*, 638–647.

Shih, M., Pittinsky, T. L., & Ambady, N. (1999). Stereotype susceptibility: Identity salience and shifts in quantitative performance. *Psychological Science, 10*, 80–83.

Shih, M., Pittinsky, T. L., & Ho, G. C. (2012). Stereotype boost: Positive outcomes from the activation of positive stereotypes. In M. Inzlicht & T. Schmader (Eds.), *Stereotype threat: Theory, process, and application* (pp. 141–158). New York, NY: Oxford University Press.

Shih, M., Pittinsky, T. L., & Trahan, A. (2006). Domain specific effects of stereotypes on performance. *Self and Identity, 5*, 1–14.

Smith, J. L., & Johnson, C. S. (2006). A stereotype boost or choking under pressure? Positive gender stereotypes and men who are low in domain identification. *Basic and Applied Social Psychology, 28*, 51–63.

Smith, J. L., & White, P. H. (2001). Development of the domain identification measure: A tool for investigating stereotype threat effects. *Educational and Psychological Measurement, 61*, 1040–1057.

Steele, C. M., & Aronson, J. (1995). Stereotype threat and intellectual test performance of African Americans. *Journal of Personality and Social Psychology, 69*, 797–811.

Stoet, G., & Geary, D. C. (2012). Can stereotype threat explain the gender gap in mathematics performance and achievement? *Review of General Psychology, 16*, 93–102.

Stricker, L. J., & Ward, W. C. (2004). Stereotype threat, inquiring about test takers' ethnicity and gender, and standardized test performance. *Journal of Applied Social Psychology, 34*, 665–693.

Tsui, M. (2007). Gender and mathematics achievement in China and the United States. *Gender Issues, 24*, 1–11.

Wei, T. E. (2012). Sticks, stones, words, and broken bones: New field and lab evidence on stereotype threat. *Educational Evaluation and Policy Analysis, 34*, 465–488.

Received February 28, 2013
Accepted January 22, 2014
Published online May 19, 2014

Carolyn E. Gibson

Department of Psychology
Georgia Southern University
PO Box 8041
USA
Statesboro, GA 30460
E-mail cegibson504@gmail.com

A Secondary Replication Attempt of Stereotype Susceptibility (Shih, Pittinsky, & Ambady, 1999)

Alice Moon and Scott S. Roeder

University of California, Berkeley, CA, USA

Abstract. Prior work suggests that awareness of stereotypes about a person's in-group can affect a person's behavior and performance when they complete a stereotype-relevant task, a phenomenon called stereotype susceptibility (Shih, Pittinsky, & Ambady, 1999). In a preregistered confirmatory design, we found that priming Asian women with social identities associated with math stereotypes did not influence their performance on a subsequent mathematics exam, and hypothesized moderators did not account for the effect. The conditions necessary to obtain the original results are not yet fully understood.

Keywords: replication, stereotype susceptibility, gender, ethnicity

The goal of this project is to offer additional insight into the original Shih, Pittinsky, and Ambady (1999) finding that group-related stereotype awareness can affect performance and behavior on tasks related to that stereotype. We conducted a replication following Gibson, Losee, and Vitiello's (2014) registered protocol.

Method

Participants

The registered protocol called for 156 participants for a very highly powered test. We aimed for that number but obtained a total of 139 Asian female undergraduates because data collection took longer than anticipated (i.e., recruitment of participants was slow). Participants were recruited from The University of California, Berkeley for either course credit or $7. We randomized to condition via Qualtrics, but nevertheless ended up with mismatched sample-sizes between conditions (i.e., Female-prime, *n* = 38; Asian-identity prime, *n* = 53; Control, *n* = 48).

Procedure

After agreeing to participate, subjects signed up for a 30-min in-laboratory timeslot and immediately completed an online pretest (Phase 1) that assessed demographic information, domain identification, and stereotype awareness. Standard consent procedures applied.

The procedure followed the Gibson et al. protocol closely. When participants arrived for their scheduled laboratory session (Phase 2), they were led to an individual cubicle where a laptop, pen, and blank piece of paper were visible. Once seated, participants were instructed to follow the instructions on the online survey and inform the experimenter when they had finished. During the first portion of the Phase 2 survey, participants were randomly assigned (via Qualtrics) to be presented with one of the manipulation questionnaires (Asian-identity, female-identity, or control) before taking the 20-min, 12-question Canadian Math Competition test used in the original study. All experimenters were blind to condition. Participants then reported their scores on the SAT or ACT (standardized tests used for college admissions in the United States), their skill level with mathematics, and how enjoyable they found the experiment. A final question examined whether or not participants picked up on the manipulation or hypothesis.

Differences From Gibson et al.

Several notable deviations from Gibson et al. are reported. First, compensation for completion was either $7 or 1 hr of research participation course credit for the 30-min task rather than entry into a $500 draw. Second, due to space constraints, the study was conducted at three different locations on campus depending on the time of the semester in which they signed up. The experimental location was either the Psychology department's research laboratory, the Behavioral Laboratory at the Haas School of Business, or the Haas Computer Laboratory. In each case, participants were seated at individual cubicles. Third, the experimenter

DOI: 10.1027/1864-9335/a000193

was a white male, a white female, or an Asian female rather than exclusively a white female (as in Gibson et al., 2014). Finally, participants came from one of the Haas Behavioral Laboratory pool (paid or credit) or the Psychology Research Participation Pool at UC Berkeley rather than at a collection of universities in the Southern US.

Results

We followed Gibson et al.'s confirmatory analysis plan. No participants were excluded. For the primary analysis, we analyzed the data of all 139 participants. Following that, as per the registered plan, we analyzed the data of only the 106 participants that were aware of both gender and racial stereotypes.

We first analyzed the effects of condition on accuracy (i.e., the number of math questions that the participant correctly answers divided by the number attempted). We tested the prediction that the Asian-identity primed group would have the best performance, the female-identity primed group would have the worst performance, and the control group would perform in between. Analyzing the entire sample, those primed with their Asian-identity scored (n = 53) .46 (SD = 0.17), those in the control condition (n = 48) scored .50 (SD = 0.18), and those primed with their female-identity (n = 38) scored .43 (SD = 0.16). Using a linear contrast analysis, there was no significant difference between groups on accuracy, $t(136) = .78$, $p = .44$, $\eta^2 = .004$, 95% CI [.00, .05]. Next, as in the original study, a two-tailed independent-samples t-test analyzed differences in accuracy between female-primed and Asian-primed conditions, but did not show a significant difference, $t(89) = .79$, $p = .43$, $d = .17$, 95% CI [−.25, .59].

Analyzing just those who were aware of the stereotype, those primed with Asian-identity (n = 42) had an accuracy score of .47 (SD = 0.18), those in the control condition (n = 37) had an accuracy score of .50 (SD = 0.17), and those primed with their female-identity (n = 27) scored .43 (SD = 0.16). There was no significant difference between groups on accuracy, $t(103) = 1.10$, $p = .28$, $\eta^2 = .012$, 95% CI [.00, .08]. Likewise, an independent samples t-test showed no difference between female-primed participants and Asian-primed participants, $t(67) = 1.09$, $p = .28$, $d = .27$, 95% CI [−.22, .75].

Shih et al. (1999) also used the total number of questions answered correctly as a dependent variable but did not observe significant differences. We likewise did not observe significant differences with this dependent variable. Those primed with Asian-identity answered 4.75 (SD = 2.16) questions correctly, those in the control condition answered 5.21 (SD = 1.91) questions correctly, and those primed with their female-identity answered 4.50 (SD = 1.96) questions correctly. In a linear contrast with the full sample, there was no significant difference between groups, $t(136) = .59$, $p = .55$, $\eta^2 = .003$, 95% CI [.00, .04]. An independent-samples t-test showed no difference between Asian-primed and female-primed group on correct responses, $t(89) = .58$, $p = .57$, $d = .12$, 95% CI [−.29, .54].

Including only those aware of both stereotypes, there was no significant difference between groups on number of questions answered correctly, $t(103) = 1.06$, $p = .29$, $\eta^2 = .011$, 95% CI [.00, .08], and the means followed the expected pattern: Asian-primed (M = 4.83), Control (M = 5.19), and female-primed (M = 4.30). Asian-primed participants did not have more correct responses than female-primed participants, $t(67) = 1.02$, $p = .31$, $d = .25$, 95% CI [−.23, .74].

Math Identification as a Moderator

The original analysis plan also suggested analyzing math identification as a possible moderator. Using the entire sample, we ran an ANCOVA with math identification as a covariate to test math identification as a moderator. This test failed to reveal a significant identity salience by math identification interaction, $F(2, 133) = .08$, $p = .92$. Math identification was also not a moderator when analyzing data from participants who were aware of both stereotypes, $F(2, 100) = .27$, $p = .76$.

In sum, scores between conditions did not significantly differ. Importantly, Asian-identity salient participants' scores on the math test did not significantly differ from female-identity participants' scores. In addition, math identification did not moderate accuracy on the test.

Discussion

We did not observe significant differences between Asian-primed, female-primed, and control conditions on math performance by Asian women. Further, the ordinal relations between conditions did not follow the original results – control participants performed nonsignificantly better than both Asian-primed and female-primed participants. Finally, neither awareness of these stereotypes nor identification with mathematics moderated this effect.

There were some differences between our replication attempt and that of Gibson et al. that could be moderators for observing this effect. However, none of those differences are presently part of the theoretical expectations of when stereotype susceptibility results are understood to occur. Another limitation of our study is that we were unable to collect the target number of participants for an extremely high-powered test. Thus, our study may have been underpowered relative to expectation. Nevertheless, the study was highly powered (n = 139) relative to the original study (n = 46).

These studies were conducted with a preregistered confirmatory design providing greater confidence in the interpretability of the reported p-values. They suggest that priming Asian women with social identities associated with math stereotypes does not influence their performance on a subsequent mathematics exam adding to the cumulative results investigating the possibility of this interesting phenomenon.

Note From the Editors

A commentary and a rejoinder on this paper are available (Moon & Roeder, 2014; Shih & Pittinsky, 2014; doi: 10.1027/1864-9335/a000207).

Acknowledgments

The authors contributed equally to this article and are listed alphabetically. Designed research: A. M., S. R., Performed research: A. M., S. R., Analyzed data: A. M., S. R., Wrote paper: A. M., S. R. We report all data exclusions, manipulations, and measures, and how we determined our sample sizes. The authors declare no conflict-of-interest with the content of this article. The project was funded by a grant from the Center for Open Science. All materials, data, and the preregistered design are available at: osf.io/b9pkj/

PREREGISTERED OPEN DATA OPEN MATERIALS

References

Gibson, C. E., Losee, J., & Vitiello, C. (2014). A replication attempt of stereotype susceptibility: Identity salience and shifts in quantitative performance. *Social Psychology, 45*, 194–198.

Moon, A., & Roeder, S. S. (2014). The effect of positive stereotypes on performance: An open question (a response to Shih and Pittinsky). Commentary and rejoinder on Moon and Roeder (2014). *Social Psychology*. Advance online publication. doi: 10.1027/1864-9335/a000207

Shih, M., & Pittinsky, T. L. (2014). Reflections on positive stereotypes research and on replications. Commentary and rejoinder on Moon and Roeder (2014). *Social Psychology*. Advance online publication. doi: 10.1027/1864-9335/a000207

Shih, M., Pittinsky, T. L., & Ambady, N. (1999). Stereotype susceptibility: Identity salience and shifts in quantitative performance. *Psychological Science, 10*, 80–83.

Received October 8, 2013
Accepted February 7, 2014
Published online May 19, 2014

Alice Moon

Psychology Department
University of California
Berkeley, CA 94720
USA
E-mail alicemoon@berkeley.edu

Sex Differences in Distress From Infidelity in Early Adulthood and in Later Life

A Replication and Meta-Analysis of Shackelford et al. (2004)

Hans IJzerman, Irene Blanken, Mark J. Brandt, J. M. Oerlemans, Marloes M. W. Van den Hoogenhof, Stephanie J. M. Franken, and Mathe W. G. Oerlemans

Tilburg University, The Netherlands

Abstract. Shackelford and colleagues (2004) found that men, compared to women, are more distressed by sexual than emotional infidelity, and this sex difference continued into older age. We conducted four high-powered replications (total $N = 1{,}952$) of this effect and found different results. A meta-analysis of original and replication studies finds the sex difference in younger samples (though with a smaller effect size), and no effect among older samples. Furthermore, we found attitude toward uncommitted sex to be a mediator (although not consistently in the same direction) between participant sex and relative distress between sexual and emotional infidelity. We hypothesize that the discrepancies between the original and replication studies may be due to changing cultural attitudes about sex across time. Confirming this speculative interpretation requires further investigation.

Keywords: evolutionary psychology, human nature, sex differences, cultural differences, replication

The idea that males are more distressed by sexual infidelity than females is perhaps one of the most widely known theoretical contributions from evolutionary psychology (Buss, Larsen, Westen, & Semmelroth, 1992; Buss et al., 1999; Buunk, Angleitner, Oubaid, & Buss, 1996; Shackelford et al., 2004). These four papers by Buss and colleagues have been cited 1,247 times (Google Scholar, March 2013). For example, Buss and colleagues (1999) compared men to women's responses to which of two events they found more distressing, such as the following example: "(A) Imagining your partner enjoying passionate sexual intercourse with that other person" or "(B) Imagining your partner forming a deep emotional attachment to that other person." Men were more likely to select A than were women.

Buss and his colleagues have replicated the effect in several samples from multiple nations (e.g., American, Dutch, German, Korean, and Japanese). In a 2004 paper by Shackelford and colleagues, the authors reused data from a 1999 student sample (Buss et al.'s, 1999, Study 2; $M_{\text{age}} = 20.2$), and compared it to an older sample ($M_{\text{age}} = 67.1$ years) that they collected themselves. They found that men in early adulthood are more distressed with sexual infidelity than women, and replicated this effect in their older sample, albeit with a smaller effect size. This latter effect is an important component of the theory that sexual jealousy is evolutionarily prepared, as the theory predicts sex differences for both age groups.

The two samples examined by Shackelford and colleagues (2004) varied in terms of achieved statistical power and their estimated effect sizes. In the young sample, the estimated Cohen's d of the sex effect was 1.29, and the 95% confidence interval suggests that the real effect size could range from Cohen's $d = 1.01$ to Cohen's $d = 1.57$ (Cohen, 1992).[1] Consistent with large effects, the post hoc achieved power of this study was larger than .99, indicating a well-powered study. In their older sample, the estimated Cohen's $d = .53$, and the 95% confidence interval this time suggests that the effect size could range from .21 to .80. The achieved post hoc power in this sample is .57. In other words, the effect size in this older sample

[1] All effect sizes and confidence intervals in this paper were obtained using Wuensch' (2012) SPSS script to calculate effect sizes with a 95% confidence interval. All Power calculations were conducted with G*Power (Erfelder, Faul, & Buchner, 1996).

DOI: 10.1027/1864-9335/a000185

may range from being very small to being large. Beyond this replication's theoretical importance, well-powered replications will help clarify the true effect size and narrow the confidence intervals for both age groups around the estimate of sex effect on sexual jealousy.

Methods

We ran four replication studies. We also included people's attitudes toward uncommitted sex as a potential moderator, but discovered in the initial studies that it may be a mediator instead. We updated our predictions throughout the research process (described in detail below). Because the methods of the studies are largely the same, we describe them together and note deviations.

Participants

Our first study was based on a convenience sample in which we collected as many participants as we could get from a convenience sample from a first year Introductory Social Psychology course (87 of 310 students completed the questionnaire). Sampling for Studies 2–4 were based on an *a priori* power calculation with G*Power (Erfelder, Faul, & Buchner, 1996; see Supplementary Materials). Studies 1, 2, and 4 included relatively younger samples; Studies 3 and 4 included relatively older samples. The Study 2 sample was recruited opportunistically through social connections around Tilburg University (105 males, 94 females). For Study 3, we planned to collect an older adult sample from city council members in the Netherlands, but we needed to supplement data collection with additional locations, such as the "50 + Beurs" (http://www.seniorenbeurstilburg.nl/) and at meeting places in Brabant where elderly often meet (like sports clubs, bars, choir practice, and personnel in care homes). We aimed to collect data from 94 older females and 105 older males, but fell short of that goal (72 men, 68 women). Study 4 was conducted using Amazon MTurk (participants received $0.30 in exchange for participation) with eligibility restricted to United States residents that had MTurk approval rates greater than 80%.[2] We aimed for 591 younger males, 591 older males, 591 younger females, and 591 older females (a total sample of 2,364 participants), but ended up collecting a total sample of 644 younger males, 577 younger females, 77 older males, and 168 older females because of far lower representation of older adults on MTurk.

Materials

In all studies, participants answered eight dilemmas in regard to what they find more disturbing in relation to their partner cheating on them. All items compared sexual versus emotional infidelity with a forced-choice format. In line with the original research, we included measures of education, relationship status, age, country of birth, country in which participants were raised, ethnicity and sex, all asked prior to the questions on sexual infidelity. We used the methodological technique of translation and backtranslation to keep the translated questionnaire loyal to the meaning of the original (for information on backtranslation see Brislin, 1970).

Additional Variables

In order to provide us with the maximal chance to obtain the effect and account for any differences with the original research, we included a measure of sociosexual orientation (SOI-R; Penke & Asendorpf, 2008). SOI-R is a measure that assesses orientation toward uncommitted sex, past sexual experiences, attitude toward uncommitted sex, and sociosexual desire. If effects have changed over time or differ by cultural context, this measure may account for those differences. The SOI-R was included after all the original study variables.

Known Differences Between Original and Replication Studies

There are several known differences between our replication attempts and the original studies; however, we do not believe that these differences are substantively relevant for the comparability of the replication. Notable differences:

- The original samples of both young and older people was conducted in English (our Studies 1–3: Dutch) with the young sample comprising of undergraduate students at a large university in the Midwestern United States (our Studies 1–3: The Netherlands) and the older sample from retirement communities (our Study 3: Community Councils and local meeting places; Study 4: Online).
- The original study was conducted with students from Introductory to Psychology classes (our Study 1: Introduction to Social Psychology; Study 2: Coauthors' social connections).
- We presume that the original study was conducted via paper-and-pencil questionnaires (our Studies 1 and 4: Qualtrics; Studies 2 and 3: Paper-and-pencil).
- The questionnaire reported in the 2004 paper included six dilemmas (ours, based on a measure provided by the original authors, included eight).
- We added the sociosexual orientation measure at the end.

[2] MTurkers with a higher approval rate generally provide better quality data than those with lower approval ratings. Although some authors advocate a 95% approval rate (e.g., Pe'er, Vosgerau, & Acquisti, 2013), we chose an 80% approval rate, because of the low representation of older participants in this sample.

Results

Studies 1 and 2

We first report our Study 1 and 2 analyses together because these studies were completed prior to the peer review of the registered report for this journal, and these results informed our hypotheses for Studies 3 and 4. For all studies, prior to analysis, we excluded participants who did not complete the entire questionnaire (*N*'s = 20, 0, 41, 80). Our confirmatory tests were chi-square analyses in SPSS, examining percentages of males and females that chose sexual infidelity as the most disturbing option for each individual scenario (see Table 1). We then calculated a composite score over the 8 dilemmas for a second confirmatory test with an independent samples *t*-test. Finally, we conducted exploratory tests: One multiple regression analysis including age and sex as factors, and another with SOI-R and sex as factors, then including their interaction terms as additional steps.

Confirmatory Results Studies 1 and 2

In Studies 1 and 2, all effects were in the predicted direction. Men were more troubled than women by their partner's sexual than emotional infidelity. In our first study of young people (M_{age} = 20.3, SD_{age} = 3.6), five out of eight of the dilemmas reached conventional significance levels. In our more highly powered second study of young people (M_{age} = 21.9, SD_{age} = 2.8), only one of the differences was significant.

Like the original studies, we averaged all scores into a composite sexual dilemma score (hereafter referred to as SDS). With SDS, in Study 1, men (M = 1.30, SD = 0.32) found sexual infidelity more distressing than women (M = 1.64, SD = 0.31), d = 1.10, $t(85)$ = 4.18, $p < .01$. Likewise, in Study 2 men (M = 1.38, SD = 0.29) found sexual infidelity more distressing than women (M = 1.47, SD = 0.35), d = 0.28, $t(182.74)$ = 2.11, p = .03.

Exploratory Results Study 2

We explored our Study 2 dataset using both SOI-R and age as possible moderating influences of the sex difference. We included SOI-R in Study 2 to assess attitudes toward casual sex and sexually promiscuous behavior (uncommitted sex). Age was theoretically relevant a priori and our second sample was slightly older than Shackelford and colleagues' first study. There was no moderation by SOI-R (t(197 = .76, p = .45), but there was one of age. The moderated regression analysis revealed a significant interaction effect between sex and mean-centered-age on SDS, $sr = -.17$, $t(197) = -2.44$, p = .02, $\beta = -.18$. Analyzing age in our regression analysis, with younger (−1 SD) versus older (+1 SD) samples estimated at 19.35 years and 24.71 respectively, we detected a significant age effect for females, $sr = -.14$, $t(197) = -2.44$, p = .05, $B = -.08$, meaning that older females in our sample were more distressed by sexual infidelity than younger females. There was no significant effect of age for males, $t(197)$ = 1.39, p = .17. Importantly, the younger men found sexual infidelity more distressing than younger women, sr = .22, $t(197)$ = 3.18, $p < .01$, B = .20, whereas the comparison between relatively older men and women in our sample yielded no significant effects, $t < 1$. This latter nonsignificant effect is different than Shackelford and colleagues (2004) who found a significant effect in their older sample, despite our sample still being much younger than their older sample.

Finally, while there was no moderation by SOI-R, we did find exploratory support that it could be a mediator. The effect of sex onto SOI ($sr = -.45$, $t(199) = -7.04$, $p < .01$, $B = -.90$), and the effect of SOI onto SDS were both significant ($sr = -.07$, $t(199) = -3.20$, $p < .01$, $B = -.07$), and the addition of SOI-R rendered the impact of sex onto SDS nonsignificant (p = .41 compared to p = .03). This meets all three requirements for full mediation (Sobel's Z = 2.93, $p < .04$). This mediation indicated that men were more likely to be open to uncommitted sex than women, and this difference accounted for greater distress by sexual infidelity.

Confirmatory Results Studies 3 and 4

In Study 3, with just older individuals (M_{age} = 58.7, SD_{age} = 6.9), none of the dilemmas reached significance (all ps > .12). In addition, the averaged SDS was not significantly different when comparing men and women, d = 0.09, $t(138)$ = .51, p = .61. In Study 4, we split the sample into younger (18–30; M_{age} = 24.5, SD_{age} = 3.4) and older (50–70; M_{age} = 55.4, SD_{age} = 4.7) participants. All dilemmas showed a significant sex difference for our younger sample (all ps < .01), with men being more distressed than women by sexual than emotional infidelity. None of the dilemmas showed a significant effect for the older sample (one p = .07, all other ps > .38). For our younger sample, SDS was significant, d = .44, $t(1520.79)$ = 8.59, $p < .01$, with no significant effect for our older sample, −0.05, $t(243) = -.371$, p = .71.

Confirmatory Mediation Analyses Studies 3 and 4

Based on the exploratory results from Study 2, we conducted confirmatory tests of whether SOI-R mediated the effect of sex onto SDS (Studies 3 and 4) and whether this mediation effect was moderated by age (Study 4). Despite no direct effect of sex on SDS above, a bootstrap analysis of Study 3 with 1,000 resamples (Preacher & Hayes, 2008) revealed a negative effect of sex onto SOI-R ($b = -.46$, SE = .20, $t = -2.29$, p = .02) and a positive effect of SOI-R onto SDS (B = .11, SE = .03, t = 4.05, $p < .01$) when including both sex and SOI-R in the equation, suggesting that SOI-R may have suppressed the effect

Table 1. Within-study sex differences in jealousy

	Percent selecting sex as more distressing				
Item #	Men	Women	χ^2 (1 *df*)	Cohen's *d*	*p*
Completed replication Study 1 (N = 87)					
3	***77.8%***	***13.0%***	***30.761***	***1.48***	***< .001***
4	***77.8%***	***42.0%***	***7.299***	***0.59***	***.008***
7	***77.8%***	***40.6%***	***7.911***	***0.63***	***.007***
10	66.7%	43.5%	3.074	0.38	.112
13	72.2%	49.3%	3.026	0.38	.112
18	67.7%	50.7%	1.461	0.38	.292
21	***77.8%***	***36.2%***	***9.963***	***0.72***	***.003***
22	***44.4%***	***8.7%***	***13.512***	***0.86***	***.001***
Overall sex effect, Cohen's *d* = 1.10, *t*(85) = 4.18, *p* < .001					
Completed replication Study 2 (N = 199)					
3	58.9%	46.8%	3.192	0.25	.074
4	69.2%	58.5%	2.181	0.21	.140
7	62.6%	58.5%	0.462	0.10	.562
10	61.7%	48.9%	3.295	0.26	.088
13	65.4%	62.8%	0.331	0.15	.331
18	66.4%	59.6%	1.179	0.15	.305
21	71%	61.7%	1.958	0.20	.179
22	***41.1%***	***22.3%***	***7.528***	***0.40***	***.007***
Overall sex effect, Cohen's *d* = 0.28, *t*(182.74) = 2.11, *p* = .03					
Completed replication Study 3 (Community sample replication, N = 143)					
3	33.3%	41.2%	0.595	0.13	.441
4	34.7%	41.2%	0.442	0.11	.506
7	40.3%	41.2%	0.013	0.02	.910
10	44.4%	41.2%	0.214	0.02	.644
13	48.6%	52.9%	0.281	0.08	.596
18	50.0%	60.3%	2.373	−0.27	.123
21	51.4%	50.0%	0.026	0.03	.872
22	26.4%	27.9%	0.043	0.04	.836
Overall sex effect, Cohen's *d* = 0.08, *t*(124), *p* = .673					
Completed replication Study 4 (MTurk replication; Full sample; N = 1,523)					
3	**58%**	**37.0%**	**67.264**	**0.43**	**< .001**
4	**63.5%**	**46.8%**	**42.710**	**0.34**	**< .001**
7	**64.6%**	**51.1%**	**28.321**	**0.28**	**< .001**
10	**65.1%**	**48.5%**	**42.666**	**0.28**	**< .001**
13	**63.4%**	**51.1%**	**23.405**	**0.25**	**< .001**
18	**63.8%**	**52.0%**	**21.616**	**0.25**	**< .001**
21	**64.4%**	**48.9%**	**37.384**	**0.31**	**< .001**
22	**50.5%**	**28.5%**	**77.337**	**0.46**	**< .001**
Overall sex effect, Cohen's *d* = 0.44, *t*(1,520.788) = 8.589, *p* < .001					
Completed replication Study 4 (MTurk replication, 18–30 Group; N = 1,221)					
3	**58.1%**	**34.8%**	**65.966**	**0.52**	**<.001**
4	**65.1%**	**46.8%**	**41.309**	**0.37**	**<.001**
7	**66.3%**	**51.6%**	**27.108**	**0.30**	**<.001**
10	**66.9%**	**48.5%**	**42.355**	**0.38**	**<.001**
13	**65.1%**	**49.9%**	**28.652**	**0.31**	**<.001**
18	**65.1%**	**51.5%**	**23.171**	**0.28**	**<.001**
21	**65.7%**	**48.7%**	**35.954**	**0.35**	**<.001**
22	**51.9%**	**26.7%**	**80.379**	**0.53**	**<.001**
Overall sex effect, Cohen's *d* = 0.50, *t*(1,187.128) = 8.732, *p* < .001					
Completed replication Study 4 (MTurk replication, 50–70 Group; N = 245)					
3	**54.5%**	**42.3%**	**3.206**	**0.23**	**.073**
4	50.6%	47.6%	0.194	0.06	.660
7	50.6%	48.2%	0.125	0.05	.723
10	50.6%	50.6%	< 0.001	0.00	.994
13	49.4%	55.4%	0.766	0.11	.382
18	51.9%	53.0%	0.022	0.02	.881
21	53.2%	50.0%	0.223	0.06	.637
22	36.4%	34.5%	0.078	0.04	.779
Overall sex effect, Cohen's *d* = 0.05, *t*(243) = .371, *p* = .711					

Note. Bold rows highlight effects consistent with Shackelford and colleagues (2004), in accordance with conventional significance levels.

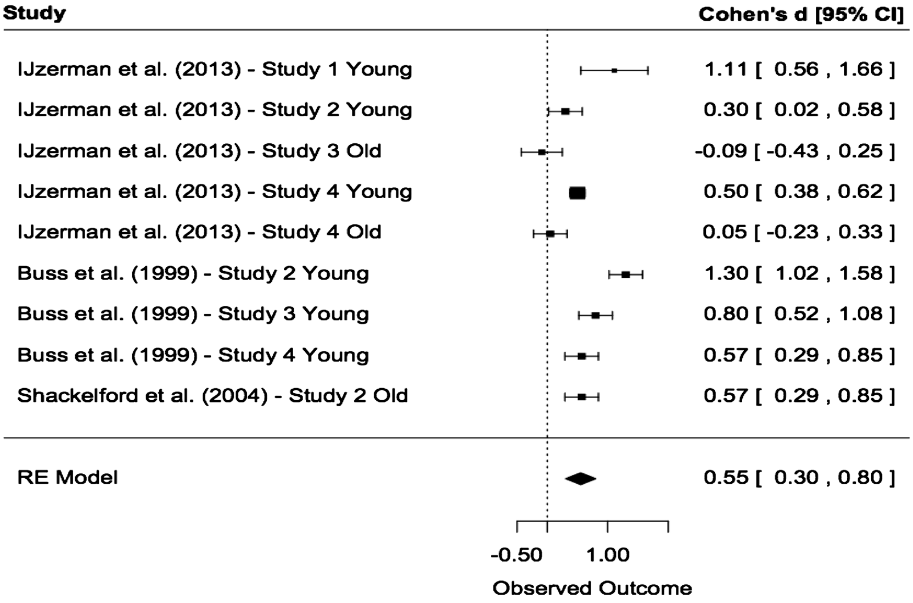

Figure 1. Forest plot of our replication studies and the original studies by Buss and colleagues (1999) and Shackelford and colleagues.

of sex on SDS. Consistent with a suppression effect, the CI of the mediated effect did not overlap with 0 (indirect effect = −.05, 95% CI −.10, −.004). Our female participants were less in agreement with uncommitted sex, and participants scoring higher on SOI-R were more disturbed by emotional infidelity of their partners. Contrary to Study 2, SOI-R served as a suppressor for the relationship between sex and the SDS and the association between SOI-R and SDS was in the opposite direction.

For Study 4, we ran a moderated mediation using PROCESS and 5,000 bias-corrected bootstrapped resamples (Hayes, 2013). For our younger sample, there was a direct positive effect of sex onto SDS (B = .18, SE = .02, t = 8.85, p < .01), and a direct negative effect of sex onto SOI-R (B = −.98, SE = .09, t = −11.06, p < .01). When adding SOI-R into the model, the effect of sex onto SDS strengthened (B = .20, SE = .02, t = 9.14, p < .01), with a direct positive effect of SOI-R onto SDS (B = .02, SE = .01, t = 2.27, p < .01). The indirect effect of sex onto SDS did not include 0 in the CI (indirect effect = −.02, 95% CI −.03, −.002). This replicates the mediation and apparent suppressor effects from Study 3.

For our older sample, there was no effect of sex onto SDS (B = .02, SE = .05, t = .37, p = .71), but there was again a negative effect of sex onto SOI-R (B = −1.21, SE = .23, t = −5.27, p < .01). When adding SOI-R to the model, the effect of sex onto SDS was again strengthened (B = .07, SE = .06, t = 1.17, p = .25), although it did not reach conventional significance levels. The (positive) effect of SOI-R onto SDS was again significant (B = .04, SE = .01, t = 2.53, p = .01). The indirect effect of sex onto SDS did not include 0 (indirect effect = −.05, 95% CI −.09, −.01), which again suggests that SOI-R serves as a suppressor.

Age did not moderate the association between sex and SOI-R (B = −.23, SE = .23, t = −.99, p = .32), nor SOI-R and SDS (B = .02, SE = .02, t = 1.41, p = .16). Finally, we observed the expected interaction effect of age and sex onto SDS (B = −.13, SE = .06, t = −2.34, p = .02).

Meta-Analysis

We conducted a meta-analysis with the metafor package in R (Viechtbauer, 2010) to derive the overall mean effect size of sex on the SDS scores (N = 9). We included the present studies, Shackelford et al. (2004), and Buss et al. (1999).[3] For each study, one aggregate effect size for SDS was calculated by averaging the items. The random effects meta-analysis produced a mean effect size of d = 0.55, 95% CI .30, .80). This aggregate result shows strong support that men are more distressed by sexual than emotional infidelity than are women (z = 4.34, p < .01). This effect of sex on SDS scores was larger among young samples (Md = 0.74, SE = 0.15, p < .01) than among older samples (Md = 0.18, SE = 0.21, p = .39), QM(2) = 24.43, p < .01. Further, the effect of sex differences in SDS was larger for the original studies (Md = 0.63, SE = 0.18, p < .01) than for our replication attempts (Md = 0.39, SE = 0.27, p = .15), QM(2) = 24.27, p < .01. Finally, the effect was larger for American (Md = 0.60, SE = 0.26, p = .02) than Dutch samples (Md = 0.40, SE = 0.32, p = .20), QM(2) = 6.86,

[3] The meta-analysis did not include Study 1 by Buss et al. (1999), and Study 1, 2, and 3 by Buunk et al. (1996) because for these studies it was not possible to retrieve the data necessary to include the studies in the meta-analysis.

p = .03. Figure 1 provides a forest plot of the effect sizes across studies. We did not use country (Japan, Korea, Netherlands, & USA) as an overall moderator, as we had too little power to detect country differences between these four countries.[4]

Discussion

We conducted four replications of Shackelford and colleagues (2004). We replicated the effect that males are more distressed by sexual infidelity than females in younger samples but we observed an effect size 50% of Shackelford et al's original studies (their d = 1.04, our Md = .52). We did not replicate the effect in older samples. Both our effect being smaller and the effect not replicating in older samples were confirmed in our meta-analysis. Together, the present findings can be considered a successful replication of the original results with two important qualifications: Size of the effect and whether the effect extends to older adults.

What accounts for these two qualifiers? Multiple factors could be influential. The original effects could be false positives because of random error, demand or instructions, recruitment strategy, or simple mistakes. We think this is unlikely given that the effect has been replicated in earlier research, sometimes with large samples. Another possibility is that cultural attitudes regarding sex may be changing over time, which seems to reduce the overall sex difference and has completely eliminated the effect among older adults. We did find a suppressor through people's attitudes toward uncommitted sex, meaning that people who were more in agreement with uncommitted sex were less distressed from sexual infidelity. The mediation of sex onto the sexual dilemma scores via people's attitudes toward uncommitted sex provides a hint for this possibility, although the mediation effect was not consistent across studies.

A third factor to consider is methodological differences between the original and replication data collections. For example, we conducted Study 4 via the Internet. However, the results are comparable to effects from our paper-and-pencil samples. And, thus far, there is little support that this difference in data collection format should matter for this kind of effect (e.g., Buhrmester, Kwang & Gosling, 2011; Klein et al., 2014). It is true that in Study 3, 57 questionnaires were not returned making it possible that differential self-selection biased the result estimates. Finally, our older samples were not as large as planned because of recruitment challenges. However, the meta-analysis was still very highly powered and at least in our studies we can conclude that the sex effect of distress from sexual versus emotional infidelity was not present in our older samples.

Conclusion

At present, we have detected the basic sex effect in distress from infidelity, showing that men (as compared to women) are more distressed from sexual (as compared to emotional) infidelity. These effects are in line with the existing reasoning. However, we found this effect to be smaller than previously suggested. It could be that the overall effect is smaller than previously suggested (given the existing confidence intervals), or that the effects have become smaller over time due to changing sexual attitudes, possibly pointing to the role of culture in (partly) determining these effects. This last suggestion is speculative, as we did not directly investigate changing attitudes across time. A promising direction for theory and research is to clarify whether and how attitudes toward uncommitted sex facilitate effects toward distress from sexual versus emotional cheating (Study 2), or repress the same effect (Studies 3 and 4).

In order to examine temporal changes and cultural differences, we think this work should be further investigated across different contexts. One option would be to incorporate this study into student replication projects (for an example, see Grahe, Brandt, IJzerman, & Cohoon, 2013). Importantly, by using replications in this special issue as a first stepping-stone on how to conduct convincing and maximally informative replications (see also, Brandt et al., 2013), our field can gain a greater theoretical understanding of the evolutionary and cultural components of the present effect.

Acknowledgments

We thank Rob Nelissen for his excellent suggestion for the present paper and the original authors Buss and Shackelford for their cooperation (and patience). We report all data exclusions, manipulations, and measures, and how we determined our sample sizes. Designed research: H. I. J., Performed research: H. I. J., I. B., J. M. O., M. M. W. vd. H., S. J. M. F., M. W. G. O.; Analyzed data: H. I. J., I. B., M. J. B.; Wrote paper: H. I. J., I. B., M. J. B. This research was partially supported by a Veni grant of the Netherlands Organization for Scientific Research (NWO) (016.145.049) and a grant from the Center for Open Science, awarded to Hans IJzerman. The authors declare no conflict-of-interest with the content of this article. All studies were preregistered at the Open Science Framework and materials, power calculations, and hypotheses can be found at https://osf.io/nrbjz/.

PREREGISTERED OPEN DATA OPEN MATERIALS

[4] In our registered analysis plan, we indicated that we would take number of items in the questionnaire as covariate. Given that all our studies included 8 items, and all the original studies 6 items, we dropped this redundant analysis.

References

Brandt, M. J., IJzerman, H., Dijksterhuis, A., Farach, F. J., Geller, J., Giner-Sorolla, R., Grange, J. A., … Van't Veer, A. (2013). The replication recipe: What makes for a convincing replication? *Journal of Experimental Social Psychology, 50*, 217–224.

Brislin, R. W. (1970). Back-translation for cross-cultural research. *Journal of Cross-Cultural Psychology, 1*, 185–216.

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

Buss, D. M., Larsen, R. J., Westen, D., & Semmelroth, J. (1992). Sex differences in jealousy: Evolution, physiology, and psychology. *Psychological Science, 3*, 251–255.

Buss, D. M., Shackelford, T. K., Kirkpatrick, L. A., Choe, J. C., Lim, H. K., Hasegawa, M., … Bennett, K. (1999). Jealousy and the nature of beliefs about infidelity: Tests of competing hypotheses about sex differences in the United States, Korea, and Japan. *Personal Relationships, 6*, 125–150.

Buunk, A. P., Angleitner, A., Oubaid, V., & Buss, D. M. (1996). Sex differences in jealousy and evolutionary and cultural perspective: Tests from the Netherlands, Germany, and the United States. *Psychological Science, 7*, 359–363.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*, 155–159.

Erfelder, E., Faul, F., & Buchner, A. (1996). GPOWER: A general power analysis program. *Behavior Research Methods, Instruments, & Computers, 28*, 1–11.

Grahe, J., Brandt, M., IJzerman, H., & Cohoon, J. (2013). *Collaborative Replications and Education Project (CREP).* Retrieved from Open Science Framework, osf.io/wfc6u.

Hayes, A. F. (2013). *Introduction to mediation moderation, and conditional process analyses*. New York, NY: Guilford Press.

IJzerman, H., Brandt, M. J., & van Wolferen, J. (2013). Rejoice! In replication. *European Journal of Personality, 127*, 128–129.

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B. Jr., Bahník, Š., Bernstein, M. J., … Nosek, B. A. (2014). Investigating variation in replicability: A "many labs" replication project. *Social Psychology, 45*, 142–152.

Pe'er, E., Vosgerau, J., & Acquisti, A. (2013). Reputation as a sufficient condition for data quality on Amazon Mechanical Turk. *Behavior Research Methods*,1–9. Retrieved from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2363822

Penke, L., & Asendorpf, J. B. (2008). Beyond global sociosexual orientations: A more differentiated look at sociosexuality and its effects on courtship and romantic relationships. *Journal of Personality and Social Psychology, 95*, 1113.

Preacher, K. J., & Hayes, A. F. (2008). Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. *Behavior Research Methods, 40*, 879–891.

Shackelford, T. M., Voracek, M., Schmitt, D. P., Buss, D. M., Weekes-Shackelford, V. A., & Michalski, R. L. (2004). Romantic jealousy in early adulthood and in later life. *Human Nature, 15*, 283–300.

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software, 36*, 1–48.

Wuensch, K. L. (2012). *Using SPSS to obtain a confidence interval for Cohen's d.* Retrieved from http://core.ecu.edu/psyc/wuenschk/SPSS/CI-d-SPSS.pdf

Received February 28, 2013
Accepted December 17, 2013
Published online May 19, 2014

Hans IJzerman

Tilburg University
School of Social and Behavioral Sciences
P64
Warandelaan 2
5037 AB Tilburg
The Netherlands
E-mail h.ijzerman@uvt.nl

Replication

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)

DOI: 10.1027/1864-9335/a000186

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, $M_{\text{age}} = 19.5$ years, $SD = 2.4$ (compare to SBH's Exp. 1: 75% female, $M_{\text{age}} = 20.0$ 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 *p*s > .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 *d*s = −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, $M_{\text{age}} = 20.5$ years, $SD = 3.6$ (compare to SBH's Exp. 2: 73% female, $M_{\text{age}} = 22.2$ 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 handwashing 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.

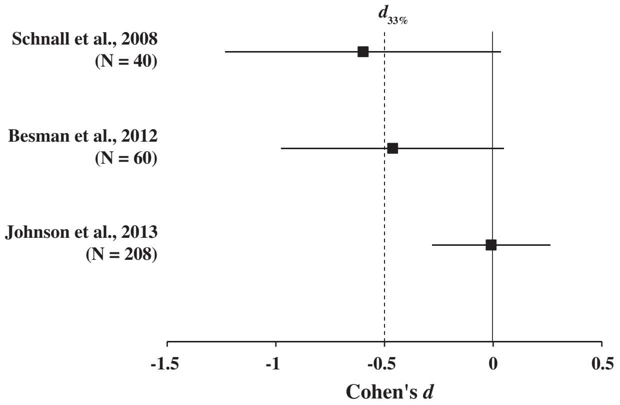

Figure 1. Results from Study 1 by Schnall, Benton, and Harvey and its replications. Markers report effect size (Cohen's d) and horizontal bars their 95% confidence intervals. The dashed line indicates the effect size ($d = -0.50$) that would give the original study, with $N = 40$, 33% power. According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

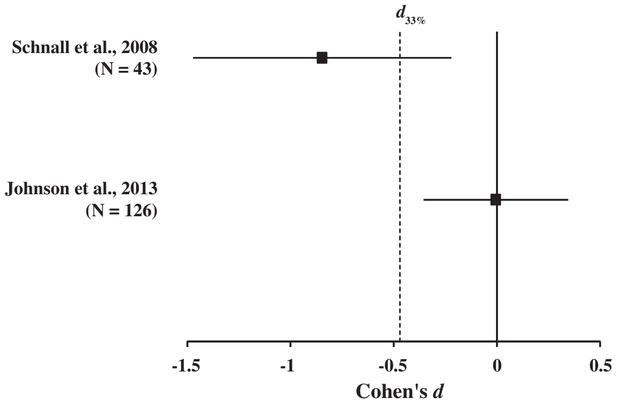

Figure 2. Results from Study 2 by Schnall, Benton, and Harvey and our replication. Markers report effect size (Cohen's d) and horizontal bars their 95% confidence intervals. The dashed line indicates the effect size ($d = 0.47$) that would give the original study, with $N = 43$, 33% power. According 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/.

PREREGISTERED OPEN DATA OPEN MATERIALS

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

Replication of "Experiencing Physical Warmth Promotes Interpersonal Warmth" by Williams and Bargh (2008)

Dermot Lynott,[1] Katherine S. Corker,[2] Jessica Wortman,[3] Louise Connell,[1] M. Brent Donnellan,[3] Richard E. Lucas,[3] and Kerry O'Brien[4,5]

[1]Lancaster University, UK, [2]Kenyon College, Gambier, OH, USA, [3]Michigan State University, USA, [4]Monash University, USA, [5]University of Manchester, UK

Abstract. We report the results of three high-powered, independent replications of Study 2 from Williams and Bargh (2008). Participants evaluated hot or cold instant therapeutic packs before choosing a reward for participation that was framed as a prosocial (i.e., treat for a friend) or self-interested reward (i.e., treat for the self). Williams and Bargh predicted that evaluating the hot pack would lead to a higher probability of making a prosocial choice compared to evaluating the cold pack. We did not replicate the effect in any individual laboratory or when considering the results of the three replications together (total *N* = 861). We conclude that there is no evidence that brief exposure to warm therapeutic packs induces greater prosocial responding than exposure to cold therapeutic packs.

Keywords: embodied cognition, social cognition, replication, temperature

Williams and Bargh (2008) investigated the effects of physical warmth on interpersonal warmth and prosocial behavior in two studies. Inspired by prior research that underscores the importance of interpersonal warmth on interpersonal judgments (e.g., Asch, 1946; Cuddy, Fiske, & Glick, 2008; but see Nauts, Langner, Huijsmans, Vonk, & Wigboldus, 2014), Williams and Bargh hypothesized that exposure to physically warmer temperatures would lead to more positive judgments of strangers and increase prosocial decision making. In their first study, participants briefly held a coffee cup containing either warm or iced coffee. In line with predictions, participants who held the warm beverage judged a target individual to have a warmer personality (i.e., more generous and caring) compared to participants who held the cold beverage.

The second study involved participants ostensibly conducting a product evaluation and subsequently making choices that could be construed as prosocial or as self-interested. The key manipulation was whether participants evaluated either a hot or cold instant therapeutic gel pack. Following the evaluation, participants made a choice that was framed either as a prosocial gift for a friend or as a personal treat. Williams and Bargh (2008) observed that those who had evaluated the warm heat pad were more likely to make the prosocial choice (OR = 3.52, 95% CI = [1.06, 11.73]). More specifically, 75% of the participants who evaluated a cold pack selected a reward for themselves, whereas 46% of the participants who evaluated a warm pack did the same (analyzed *N* = 50). The conclusion from this work was that experiences of physical warmth unconsciously impact our impressions of others and prosocial behavior. The basic idea is that physical feelings of warmth translate to greater interpersonal warmth.

The Williams and Bargh (2008) paper was published in a prestigious journal (*Science*), and the paper has been highly cited (more than 470 times according to Google Scholar, more than 160 citations in Web of Science). The findings also received coverage in the popular press (e.g., Bartlett, 2013; Tierney, 2008), and, despite not having been directly replicated, has impacted subsequent research investigating how experiences of hot and cold can prime other behaviors (e.g., Bargh & Shalev, 2012; Kang, Williams, Clark, Gray, & Bargh, 2011; Leander, Chartrand, & Bargh, 2012; Williams, Huang, & Bargh, 2009). In this paper, we seek to replicate the findings of Study 2 from Williams and Bargh (2008).

DOI: 10.1027/1864-9335/a000187

Method

Power Analysis and Sampling Plan

Based on the effect size from the original study and requiring statistical power of .95 with an alpha level of .05, we estimated the required sample size to be 300 participants (Epicalc; Chongsuvivatwong, 2012). This is a conservative estimate, allowing for the detection of a smaller effect than that observed in the original study (see Table A1, online supplementary materials). Three independent replications were conducted, each with a target sample size of 300. Replications took place at two US locations (as in the original study): Kenyon College, Michigan State University, and one UK location: University of Manchester, following agreed upon procedures.

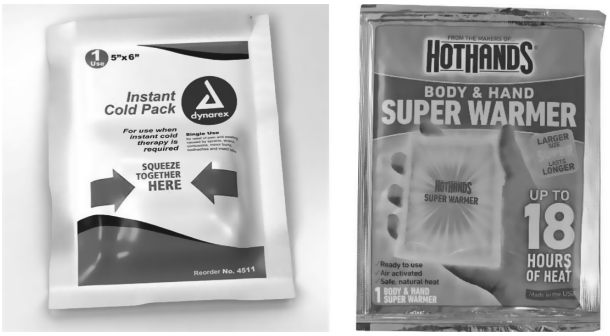

Figure 1. Example of therapeutic packs used in the study. Both packs were 4″ × 5″ size.

Table 1. Demographics for the three study locations

Location	UK	Kenyon	MSU
N (total)	305	306	250
N (analyzed)	282	294	237
Age (*M*)	27.17	41.85	22.11
(*SD*)	(13.65)	(17.16)	(5.30)
Gender (% female)	51.5%	52.9%	58.7%
Native speaker of English	78.4%	98.6%	83.2%
Education			
No high school diploma	0.0%	1.0%	0.0%
High school	21.6%	29.2%	8.1%
Some college	41.3%	38.6%	60.6%
Bachelor's degree or higher	37.1%	31.2%	31.3%

Note. N (analyzed) = number of participants remaining after exclusions.

Materials and Procedure

We preregistered the study proposal on the Open Science Framework (OSF) website and followed the procedures of the original study as closely as possible, with some minor modifications. For example, the choice of rewards offered varied depending on local availability. We also used different brands of therapeutic packs, again due to availability. Additionally, in all three replications, research assistants were blind to participants' assigned temperature conditions, to reduce experimenter expectancy effects. Full details of the procedures at each location can be found in the online supplemental materials.

Researchers set up tables and testing areas at each event, and passersby were approached to take part in a product-evaluation study. Participants were brought to the testing area, where they were separated from the researcher by partitions. Once the consent form was signed, participants were given a questionnaire booklet. The cover page served to hide the second page that instructed the participant which of two boxes in front of them they should open; one box contained a hot pack, and one contained a cold pack. The cover page ensured that researchers were blind to the temperature pack condition to which participants were assigned.

On the questionnaire (see online supplementary materials), participants evaluated the effectiveness of either the hot/cold pack on a scale ranging from 1 = *not at all* to 7 = *extremely* and indicated to what extent they would recommend the product to their family, friends, or strangers on the same 7-point scale. (In the original study, participants indicated whether they would or would not recommend the product to their family, friends, or strangers as a dichotomous choice.) Finally participants estimated the internal temperature of the gel pack in degrees Celsius (UK site) or Fahrenheit (US sites). The first four questions were included in the original study to support the initial cover story, and the final question was intended as a manipulation check.

Once participants completed these questions, the questionnaire directed them to place the therapeutic gel pack product back in its original box. This also served to ensure that researchers remained unaware of the participant's condition. The next page of the questionnaire included the main dependent variable, which consisted of the reward choice.

Each participant then completed a short funneled debriefing questionnaire (see online supplementary materials), which allowed us to evaluate whether the participant was suspicious of the study or guessed the underlying hypothesis. Once the participant had completed the funneled debriefing, they were brought away from the testing area and were given their chosen reward together with a page explaining the true nature of the study.

All three sites used the same 4″ × 5″ hot and cold instant therapeutic packs. Brand names were obscured from the packs with black marker. Hot packs were HeatMax brand hand and body warmers; cold packs were Dynarex brand (see Figure 1).

Participants

Demographic information for all participants is reported in Table 1.

Kenyon College

Participants were 306 individuals from the Mt. Vernon, Ohio community area (*N* = 289) and Kenyon College psychology research pool participants (*N* = 17). Community participants were recruited in the outdoors from a local summertime festival in June and July of 2013. Kenyon College participants completed the study indoors. The rewards for participating in the study were a Snapple beverage (available immediately) or a voucher for a local cupcake shop worth $2 (located within walking distance of the data collection site).

Michigan State University

Participants were 250 individuals recruited at various indoor and outdoor locations on the Michigan State University campus during October and November 2013. Two hundred and fifty was the maximum number of participants that could be collected given the available time for data collection. The rewards for participating in the study were a Snapple beverage (available immediately) or a voucher for an ice cream at the campus dairy store. The voucher was actually worth $2 in US currency but the cash value was not mentioned to participants.

University of Manchester

Participants (*N* = 305) were recruited over several days during September and October 2013, at indoor and outdoor events around the University of Manchester (Open Days, Welcome Week), with a small proportion tested at an army reserve training day (*N* = 13). The rewards for participating in the study were a voucher for either a fruit juice or a fruit smoothie.

Design

The dependent variable was whether participants made a prosocial or selfish choice on the critical reward question. The temperature of the pack (hot/cold) and reward framing (i.e., the counterbalancing of each reward as prosocial or selfish) served as between-participants independent variables.

Results

Data Preparation and Manipulation Check

Each replication study was conducted independently, and there was no discussion of results between the groups until all data were collected. The analyses reported for each study have also been verified by at least one other group. We report all data exclusions, manipulations, measures, and how we determined our sample sizes. All data are available on the OSF project page.

Participants who met any of several a priori agreed upon rules for exclusion were removed prior to analysis. Grounds for exclusion included (1) being ± 3 *SD* away from the mean within each condition for pack temperature estimation, (2) failing to choose a reward for participation (the key dependent measure), and (3) making a connection in the debriefing form between physical and interpersonal warmth. There were 12 Kenyon College, 13 Michigan State University, and 23 University of Manchester participants excluded from the analysis on these grounds. The *N* on which all analyses are based is listed in Table 1. Additional information about excluded participants is in the supplemental materials.

Participants at all three sites rated the hot pack as warmer than the cold pack: *d*s = 2.50 (Kenyon), 2.22 (Michigan State), 2.61 (Manchester), suggesting that the manipulation was effective. These effect sizes are similarly large to those obtained in the original study (*d* = 2.98; Williams & Bargh, 2008). Descriptive statistics are in Table 2. The effectiveness item and the three recommendation items had high intercorrelations (Cronbach's α = .95 (Kenyon), .93 (Michigan State), and .93 (Manchester)) and were averaged together into a scale. There was no evidence of consistent differences across sites regarding this scale (see Table 2). Michigan State participants rated the cold pack as more effective/recommendable than the hot pack (*d* = 0.42); Kenyon College and University of Manchester participants did not distinguish between the packs on this scale (*d*s = 0.05 and 0.20, respectively). Williams and Bargh's participants completed dichotomous recommendation ratings. However, on the single effectiveness item, Williams and Bargh (2008) found that the cold pack was rated more effective than the warm pack (*d* = 0.93).

Kenyon College

A chi-square test of pack temperature (cold vs. hot) on selfish behavior for each reward frame was conducted.[1] The analysis was significant for the "Snapple is selfish" framing though in the opposite direction to that predicted by Williams and Bargh (2008), χ^2 (1) = 5.276, *p* = .022. In this framing condition, 61.3% of participants who evaluated the hot pack made the selfish choice, whereas 42.5% of participants who evaluated the cold pack did the same. The analysis was not significant for the "Cupcake is selfish" framing, χ^2 (1) = 0.259, *p* = .611. In this framing condition, 41.1% of participants who evaluated the hot pack made the selfish choice; 37.0% of participants who

[1] In the preregistered version of the study we indicated that we would use logistic regressions to analyze the data, following the analyses employed by Williams and Bargh (2008). However, we felt that, given the complexities of interpreting the significant interaction of the logistic regression analysis in the 2008 study, chi-square analyses provided a clearer test of the data, with more easily interpretable results. Nonetheless, analyzing the data from the replication studies using logistic regressions yields the same patterns as the chi-square analyses reported.

Table 2. Ratings of hot and cold packs

	Kenyon College	Michigan State	Uni. Manchester
Temperature (hot)	87.53 (26.70)	85.37 (25.29)	77.36 (20.71)
Temperature (cold)	36.00 (11.78)	35.23 (19.42)	35.58 (9.23)
Effectiveness (hot)	4.61 (1.44)	4.80 (1.31)	3.71 (1.48)
Effectiveness (cold)	4.69 (1.53)	4.20 (1.53)	4.00 (1.48)

Notes. Values represent *M* (*SD*). Temperatures are in degrees Fahrenheit.

evaluated the cold pack did the same. Collapsing across framing conditions, the analysis yielded no evidence that participants exposed to cold packs were more selfish than participants exposed to warm packs. The chi-square value was statistically significant, χ^2 (1) = 4.00, *p* = .045, OR = 0.61, 95% CI = [0.38, 0.98], though it was in the opposite direction of the original Williams and Bargh (2008) prediction with 51.4% of hot pack participants making the selfish choice compared to 39.7% of cold pack participants.[2]

An additional set of exploratory analyses were conducted to probe the robustness of the results under different selection conditions. First, the analysis was repeated with an additional 10 participants' data removed. These participants experienced a variety of procedural problems that made us question whether they should stay in the analysis (see supplementary materials). The previously reported findings hold with these 10 participants removed (they also hold with all available data included – i.e., no exclusions). Further, we tested whether restricting the participants to only the community participants would matter; the findings held with just this subset of participants. Finally, we examined whether June 2013 participants (taken on an unseasonably cool evening) would differ from July 2013 participants (a much warmer evening). The findings did not achieve statistical significance in either month examined in isolation, but the pattern of results appeared similar in both months.

A chi-square analysis predicting prosocial behavior from pack temperature separately for men and women was conducted. Among women, 38.0% made the selfish choice after exposure to the cold pack, compared to 53.8% after exposure to the warm pack, OR = 0.53, 95% CI = [0.28, 0.99], a statistically significant difference. Among men, 41.8% made the selfish choice after exposure to the cold pack, compared to 47.8% after exposure to the warm pack, which was not statistically significant, OR = 0.78, 95% CI = [0.40, 1.55]. Thus the pattern of the results appeared similar for men and women, but perhaps the pattern was somewhat more pronounced for women participants.

Overall, these analyses indicate that we failed to replicate the findings of Williams and Bargh (2008). In fact, in one framing condition, the predicted effect was in the opposite direction, such that participants who held the warm pack were actually more selfish than participants who held the cold pack.

Michigan State University

An identical set of analyses was conducted. The results revealed that there was no effect of pack temperature on selfishness of choice in either framing condition (''Snapple is Selfish'': χ^2 (1) = 0.234, *p* = .629; ''Ice Cream is Selfish'': χ^2 (1) = 0.019, *p* = .889) or collapsed across framing, χ^2 (1) = 0.039, *p* = .842, OR = 0.92, 95% CI = [0.56, 1.53]. In the ''Snapple is selfish'' framing, 44.1% of participants in the hot pack condition and 39.7% of participants in the cold pack condition made the selfish choice. In the ''ice cream is selfish'' framing, 62.7% of participants in the hot pack condition and 63.9% of participants in the cold pack condition made the selfish choice. Collapsed across framing conditions, 53.4% of hot pack and 52.1% of cold pack participants made the selfish choice.

The inclusion of the data from the 13 removed participants did not change this result. Furthermore, the hypothesis was not supported for either male or female participants examined in isolation. Thus, these data also failed to replicate the original Williams and Bargh (2008) results.

University of Manchester

We originally intended to counterbalance which items were framed as the self-interested and prosocial options, but due to a printing error, fruit juice was always the self-interested option in the warm condition, whereas fruit smoothie was always the self-interested option in the cold condition. This modification should not impact our results as (1) the items (fruit juice, smoothie) were chosen for their similarity, (2) the specifics of the reward item are not theoretically relevant, only whether participants make the prosocial or the self-interested choice, and (3) no effect of temperature condition on type of reward was observed in the results of the other two replication sites reported in this paper.

[2] For comparison purposes, in Williams and Bargh's (2008) data the effect of the temperature pack manipulation was statistically significant in the ''ice cream is selfish'' framing condition, χ^2 (1) = 6.032, *p* = .014 (% selfish in the cold condition = 92.9%; % selfish in the hot condition = 50.0%). The analysis was not significant in the ''Snapple is selfish'' framing condition, χ^2 (1) = .729, *p* = .527 (% selfish in the cold condition = 50.0%; % selfish in the hot condition = 42.9%). The analysis was statistically significant collapsing across framing, χ^2 (1) = 4.327, *p* = .038 (% selfish in the cold condition = 75.0%; % selfish in the hot condition = 46.2%).

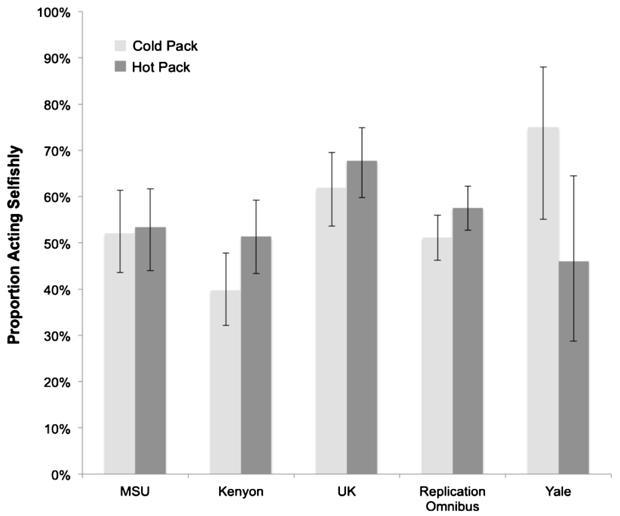

Figure 2. Proportion of participants who made the selfish choice across studies. Error bars represent 95% CI. MSU = Michigan State University.

We therefore proceeded with a chi-square test of pack temperature on prosociality of choice.

Analysis revealed that 64.2% of people made the selfish choice overall, but that this choice was not significantly related to the temperature priming condition: χ^2 (1) = 1.10, p = .295, OR = 0.77, 95% CI = [0.47, 1.26]. Specifically, 61.2% (N = 85) chose the selfish response in the cold condition, and 67.1% (N = 96) chose the selfish option in the warm condition. The inclusion of the data from the 23 removed participants did not change this result. Thus, the overall result did not replicate the original finding. Instead, the observed numerical trend was slightly in the opposite direction to that predicted.

Given the large sample it is possible that subsets of the participants displayed the pattern predicted by the original study, but that the pattern did not generalize to the overall sample. With this in mind, we examined whether significant effects of the temperature manipulation could be observed if we divided the sample by whether participants took part indoors (N = 157) or outdoors (N = 125), were native (N = 220) or non-native (N = 62) speakers of English, or were male (N = 139) or female (N = 143). There were no significant effects of temperature condition for any of these binary groupings.

Omnibus Analysis

The data from the three replication sites were combined into one chi-square analysis to determine the impact of pack temperature on reward choice (selfish vs. prosocial). There was no significant effect, although the result approached statistical significance in the opposite direction of that predicted by Williams and Bargh (2008), χ^2 (1) = 3.402, p = .065, OR = 0.77, 95% CI = [.58, 1.02]. The results are displayed in Figures 2 and 3.

Discussion

Williams and Bargh (2008) found that participants who previously held a hot pack made a more prosocial choice than participants who previously held a cold pack. We attempted three high-powered, independent replications of this original study, but we did not replicate the original result. We found no indication that participants who held warm packs were more prosocial than participants who held cold packs when prosocial actions were defined as opting for a token reward gift for a friend as opposed to a treat for the self. In our samples, the effect was (not significantly) in the opposite direction, such that participants who evaluated a cold pack were marginally more prosocial than participants who evaluated a hot pack, but this effect did not reach statistical significance at the $p < .05$ level. To summarize, we did not replicate the original result.

There may be several reasons for why we did not observe significant effects in these replications. One possibility is expectancy effects, which have previously been suggested as explanations following failures to replicate other social priming effects (see e.g., Klein et al., for a recent discussion of expectancy effects in experimental studies). The effect observed by Williams and Bargh may have been due, in part, to unconscious cues given by the researcher. In the original study, the researcher interacted directly with participants as they received their hot/cold packs, and so it is possible that unplanned cues were transmitted during this brief exchange (e.g., giving cues to behave more prosocially if participants were in the hot condition). In our study, the researchers were not aware of the condition the participants were in, at least until the debriefing procedure took place (and only then if participants verbally revealed details of their condition), and so could not provide unconscious cues consistent with the study predictions.

Of course there are many other possible explanations for why effects were found in the original study and not in the replication attempts (e.g., small sample sizes of original studies, random variations in the data, influence of unknown moderators). However, it is important to emphasize that the current results do not suggest that there are no influences of temperature on people's behavior or that the current and related effects in the hot and cold priming literature are false positives. In the first case, there are many other examples demonstrating links between temperature change and behavioral outcomes, although the general tendency has been to find links between increased aggression and higher temperatures (e.g., Anderson, 2001), rather than higher temperatures being associated with more prosocial behaviors. In the second case, while it is clear that the temperature priming effect observed by Williams and Bargh (2008) cannot be reliably observed using highly similar procedures, it is important that evidence for any given priming effect in the literature should be considered on its own merits; effects should be investigated and replicated independently and not automatically dismissed beforehand. In short, we suggest more work is needed on this topic and conclude that the current results suggest some degree

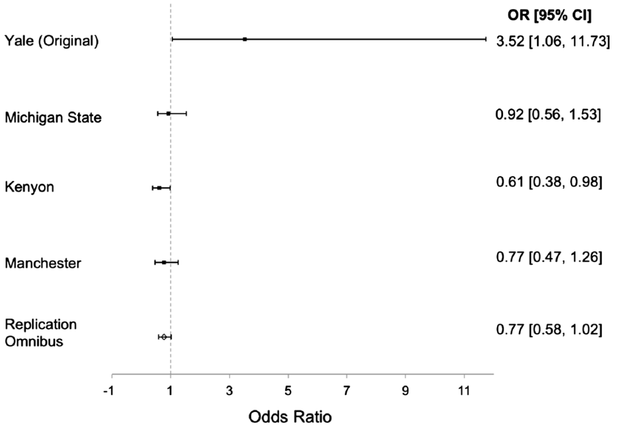

Figure 3. Odds ratios (OR) for tendency to make the selfish choice after exposure to cold versus warm temperatures. Values on the right hand side represent ORs and 95% CIs (lower and upper values). An OR of 1.0 indicates a null effect (i.e., even odds of selfish responding).

of added caution is needed when considering whether exposure to hot or cold temperatures impacts prosocial behavior. More broadly, there is a need for greater specification of the theoretical underpinnings and limitations of priming effects by researchers (Cesario, 2014) and more details of experimental procedures and analyses conducted (see, e.g., Klein et al., 2012; Nosek, Spies, & Motyl, 2012 for detailed suggestions in this vein). In this way, we can look forward to building a more robust social psychology for the future.

Note From the Editors

A commentary and a rejoinder on this paper are available (Lynott et al., 2014; Williams, 2014; doi: 10.1027/1864-9335/a000205).

Acknowledgments

The first and second authors contributed equally to this manuscript. This research was funded by a grant to the first and second authors from the Center for Open Science. Designed research: D. L., K. S. C., B. D., R. E. L., K. O.; Performed research: D. L., K. S. C., J. W., L. C.; Analyzed data: D. L., K. S. C., J. W., B. D.; Wrote initial draft: D. L., K. S. C.; Provided editorial comments: D. L., K. S. C., L. C., M. B. D., R. E. L., K. B., J. W. The authors declare no conflict of interest with the content of this article. The authors would like to give enormous thanks to Pat Watkinson and Jackie Harte for help with data collection at Manchester University and to Melanie Crank and Manchester University Student Union for help in finding space to run the study. We also thank Simon Garcia at Kenyon College for his assistance in evaluating the temperature of the therapeutic packs, as well as Katie Finnigan, Melek Spinel, Olivia Siulagi, and Maureen Hirt for their assistance in data collection. Finally, we would like to thank Lawrence Williams for helpful comments on the replication proposal and for making his data freely available. All materials, data, images of the procedure, and the preregistered design are available at https://osf.io/gjt37/.

PREREGISTERED OPEN DATA OPEN MATERIALS

References

Anderson, C. A. (2001). Heat and violence. *Perspectives on Psychological Science, 10*, 33–38.

Asch, S. E. (1946). Forming impressions of personality. *The Journal of Abnormal and Social Psychology, 41*, 258–290.

Bargh, J. A., & Shalev, I. (2012). The substitutability of physical and social warmth in daily life. *Emotion, 12*, 154.

Bartlett, T. (2013, January 30). Power of suggestion. *The Chronicle of Higher Education*. Retrieved from http://chronicle.com/article/Power-of-Suggestion/136907

Cesario, J. (2014). Priming, replication, and the hardest science. *Perspectives on Psychological Science, 9*, 40–48.

Chongsuvivatwong, V. (2012). epicalc: Epidemiological calculator, [Computer software manual]. Retrieved from http://cran.r-project.org/web/packages/epicalc/epicalc.pdf (R package version 3.0.0)

Cuddy, A. J., Fiske, S. T., & Glick, P. (2008). Warmth and competence as universal dimensions of social perception: The stereotype content model and the BIAS map. *Advances in Experimental Social Psychology, 40*, 61–149.

Harris, C. R., Coburn, N., Rohrer, D., & Pashler, H. (2013). Two failures to replicate high-performance-goal priming effects. *PloS One, 8*, e72467.

Kang, Y., Williams, L. E., Clark, M. S., Gray, J. R., & Bargh, J. A. (2011). Physical temperature effects on trust behavior: The role of insula. *Social Cognitive and Affective Neuroscience, 6*, 507–515.

Klein, O., Doyen, S., Leys, C., 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.

Leander, N. P., Chartrand, T. L., & Bargh, J. A. (2012). You give me the chills: Embodied reactions to inappropriate amounts of behavioral mimicry. *Psychological Science, 23*, 772–779.

Lynott, D., Corker, K. S., Wortman, J., Connell, L., Donnellan, M. B., Lucas, R. E., & O'Brian, K. (2014). High quality direct replications matter: Response to Williams (2014). *Social Psychology*. Advance online publication. doi: 10.1027/1864-9335/a000205

Nauts, S., Langner, O., Huijsmans, I., Vonk, R., & Wigboldus, D. H. J. (2014). A replication and review of Asch's (1946) evidence for a primacy-of-warmth effect in impression formation. *Social Psychology, 45*, 153–163.

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.

Tierney, J. (2008, October 23). Heart-warming news on hot coffee. *New York Times*. Retrieved from http://tierneylab.blogs.nytimes.com/2008/10/23/heart-warming-news-on-coffee/?ei=5070&emc=eta1

Williams, L. E. (2014). Improving psychological science requires theory, data, caution: Reflections on Lynott (2014). Commentary and rejoinder on Lynott et al. (2014). *Social Psychology*. Advance online publication. doi: 10.1027/1864-9335/a000205

Williams, L. E., & Bargh, J. A. (2008). Experiencing physical warmth promotes interpersonal warmth. *Science, 322*, 306–307.

Williams, L. E., Huang, J. Y., & Bargh, J. A. (2009). The scaffolded mind: Higher mental processes are grounded in early experience of the physical world. *European Journal of Social Psychology, 39*, 1257–1267.

Received February 28, 2013
Accepted December 5, 2013
Published online May 19, 2014

Dermot Lynott

Department of Psychology
Lancaster University
Lancaster LA1 4YF
UK
E-mail d.lynott@lancaster.ac.uk

Katherine S. Corker

Department of Psychology
Kenyon College
Gambier, OH, 43022
USA
E-mail corkerk@kenyon.edu

Replication of Experiments Evaluating Impact of Psychological Distance on Moral Judgment

(Eyal, Liberman & Trope, 2008; Gong & Medin, 2012)

Iris L. Žeželj, and Biljana R. Jokić

Department of Psychology, University of Belgrade, Serbia

Abstract. Eyal, Liberman, and Trope (2008) established that people judged moral transgressions more harshly and virtuous acts more positively when the acts were psychologically distant than close. In a series of conceptual and direct replications, Gong and Medin (2012) came to the opposite conclusion. Attempting to resolve these inconsistencies, we conducted four high-powered replication studies in which we varied temporal distance (Studies 1 and 3), social distance (Study 2) or construal level (Study 4), and registered their impact on moral judgment. We found no systematic effect of temporal distance, the effect of social distance consistent with Eyal et al., and the reversed effect of direct construal level manipulation, consistent with Gong and Medin. Possible explanations for the incompatible results are discussed.

Keywords: psychological distance, moral judgment, construal level theory, replication

Most of the time people act and feel toward the construal of an object, rather than object itself. Construal Level Theory (CLT; Trope & Liberman, 2010) addresses universally relevant issues of mental representations and thus runs across the subdisciplines of psychology. CLT has been widely tested in different domains, and seems to have received relatively robust empirical confirmation. It is established as an influential theory of social cognition: A chapter is regularly devoted to it in contemporary handbooks (Shapira, Liberman, Trope, & Rim, 2012; Trope & Liberman, 2011), and the seminal article on Temporal construal (Trope & Liberman, 2003) has been highly cited (656 citations in Scopus, 689 in APA Psych Net, and 1,324 times in Google scholar[1]).

CLT's major premise is that mental representations of objects and situations vary depending on their distance from the perceiver: Distant objects tend to be represented by few essential characteristics (high-level construal), whereas close objects tend to be represented by detailed and contextual information (low-level construal). High-level construals are important for regulation toward distal objects, when it is important to focus on their invariant, central features, while low-level construals are more important for regulation toward close objects, as we need to act toward them immediately. Given that psychological distance is egocentric, with its reference point being here and now, it can be social (self vs. other), temporal (present self vs. future or past self), spatial (here vs. some other place), or hypothetical (highly likely vs. unlikely event).

CLT proposes that moral values have stronger impact on judgment from a greater psychological distance, while contextual information is more relevant when at a close distance. In other words, the same act will be evaluated differently depending on the perceivers' distance from the presented event. In a series of experiments, Eyal, Liberman, and Trope (EL&T, 2008) tested how temporal and social distance impact the evaluation of moral acts (both transgressions and virtuous acts), followed by contextual information that intended to attenuate their severity. Temporal distance was manipulated by the instruction to imagine the situation was taking place now or later in time. Social distance was manipulated by the instruction to imagine the situation from one's own or from a third person's perspective. Four experiments in their article confirmed CLT propositions: Both moral transgressions and virtuous behaviors were evaluated differently depending on the psychological distance. When the distance was higher, values (as high-level construal) had a greater impact on evaluation. When the distance was smaller, contextual information (as low-level construal) was of greater concern. This study has been cited (33 times in Scopus, 76 in Google scholar), as supportive of CLT predictions within the moral evaluation domain. In a similar vein, Agerström and Bjorklund (2009) examined the impact of temporal distance on

[1] Number of citations retrieved in October 2013.

DOI: 10.1027/1864-9335/a000188

moral concerns (selfish vs. altruistic considerations) and reported that moral concerns were higher for temporarily distant situations. In a series of experiments, Lammers' results (2012) mirrored those of EL&T: Subjects reacted more negatively to others' morally questionable behaviors when adopting an abstract rather than concrete perspective.

A following project by Gong and Medin (G&M, 2012), however, yielded some directly conflicting results. The authors used the same stimulus material as EL&T (vignettes about moral transgressions and virtuous acts), but instead of manipulating distance, they directly manipulated construal level by priming tasks designed to stimulate either the abstract or concrete mindset. Contrary to expectations, when participants were primed with a concrete mindset, they showed more extreme moral judgments of both moral transgressions and virtuous behaviors. In order to test the possible impact of a different procedure (manipulation of distance vs. manipulation of construal level), G&M directly replicated one of four experiments from EL&T's study. Again, G&M observed the opposite results. When EL&T tried to replicate G&M's experiments (using the priming tasks), their results were not significant. Baring in mind these contradictory findings and the importance of the topic (impact of construal level on moral judgment), we proposed a two-stage replication project:

1. Direct replication of EL&T's Studies 2, 3, and 4 and G&M's Study 1 in the Social Cognition Laboratory at Belgrade University, Serbia. The main authors of both studies agreed to provide us with the materials (vignettes, instruments, and instructions) used in their experiments.
2. Aggregation of databases from three laboratories: Ben-Gurion University (Israel), Northwestern University (Ilinois, USA), and Belgrade University (Serbia). This provided us with the opportunity to directly compare: (a) The effects obtained by three laboratories (Experiment 2 from EL&T's study, and Experiment 1 from G&M's study); (b) the effects obtained by two laboratories (Israeli and Serbian), from Experiments 3 and 4 of EL&T's.

Materials, data, and the preregistered proposal are available on the project page on the Open Science Framework (https://osf.io/Z5TE6/).

We report all data exclusions, manipulations, and measures, and how we determined our sample sizes. The planned sample size[2] was based on the effect size from each of the original studies, so as to achieve .95 power (see Table 1 for details).

Whenever we had an opportunity, we tested up to 5% respondents over the planned sample size, in case we had to omit some from further analysis. Participants were randomly assigned to experimental groups. They were recruited from a pool of psychology students from Belgrade University, in exchange for course credits. The same recruiting method was applied in EL&T's original study, while in G&M's study participation was on a voluntary basis. Gender information was registered, although it was not found to have a significant effect in either study.

Table 1. Effect sizes from original studies and planned sample size

Original study	Original sample size	Cohen-s' *d* from original study	Planned sample size
Exp 2 (EL&T)	58	.68	114 (57 per group)
Exp 3 (EL&T)	40	.72	102 (51 per group)
Exp 4 (EL&T)	47	.81	80 (40 per group)
Exp 1 (G&M)	34	1.07	48 (24 per group)

We used the vignettes from EL&T's studies, with the authors' permission. The vignettes were translated into Serbian by two independent bilingual translators, and then translated back into English in order to provide maximum correspondence (as suggested in Brislin, 1970, 1976). As was the procedure in the original experiments, the respondents filled in booklets in paper format.

The experimenters were PhD students from the Social Cognition Laboratory of the Department of Psychology, Belgrade University, blind to the study hypothesis. Participants were tested in groups no larger than ten. At the beginning of each session, the experimenter presented themselves to the participants and explained the research purpose: "This study is about a judgment of different people's actions. After reading each story, please provide your opinion below it. There are no right or wrong answers. You are expected to evaluate just as you think. Examination is anonymous, so you do not need to provide any personal information." The experimenter was present until the end of the experiment, but not allowed to give any additional instructions, except to encourage participants to give their own opinions if they had any questions. After hearing the instructions, participants read the vignettes and evaluated the wrongness (in the first two studies), virtuousness (in the third study), or moral acceptability (in the fourth study) of the actions presented. We used the same scales as in corresponding studies of EL&T and G&M.

Our analysis was planned in a confirmatory fashion.[3] We performed a series of standard *t*-tests, as well as default Bayesian *t*-tests, as proposed by Rouder, Speckman, Sun, Morey, and Iverson (2009), in order to test the differences:

1. In perceived wrongness of the actions between high and low temporal distance primed groups (Study 1);
2. In perceived wrongness of the actions between high and low social distance primed groups (Study 2);
3. In perceived virtuousness of the actions between high and low temporal distance primed groups (Study 3);

[2] Estimations were calculated using Lenth, R. V. (2006–2009). Java Applets for Power and Sample Size [Computer software]. Retrieved from http://www.stat.uiowa.edu/~rlenth/Power

[3] https://osf.io/Z5TE6/files/proposal_for_replication_zezelj_final.pdf/

4. In perceived moral acceptability of the transgressions between high and low construal level primed groups (Study 4).

In a next step we performed an analysis on the aggregated database (with provided data from two or three laboratories), and at the end we introduced "laboratory" as a factor.

The main known difference from the original study was the cultural and linguistic backgrounds of the samples (Serbian vs. Israeli/Hebrew vs. American/English). One minor difference was that our sample consisted exclusively of students, while in one of EL&T's Study 3 participants were workers from security service organizations. Apart from that, a full methodological and procedural equivalence was set up.

Study 1

The aim of the Study 1 was to replicate the findings of EL&T (Study 2, 2008) indicating that people would judge immoral acts more harshly if presented to them as temporally distant rather than presented as temporally close. Participants judged the wrongness of moral transgressions as expected to occur either the next day (near future condition) or next year (distant future condition).

Method

Participants

Participants in our study were 116 undergraduate students from University of Belgrade, Serbia, who participated in exchange for course credit. Our aggregated database included 58 participants from the original Israeli study and 36 from American replication; a total of 210 participants.

Procedure

Participants read three vignettes (adopted from Haidt 2001; Haidt, Koller, & Dias, 1993; as in EL&T, 2008) describing a moral transgression followed by situational details that moderated the offensiveness of the action, for example "sister and brother had a sexual intercourse" (violating a widely accepted moral rule), but "they used birth control and agreed to do it only once" (contextual information that was supposed to attenuate the severity of the act). They were asked to imagine that the events would happen the next day or the next year. After reading each vignette, participants evaluated the wrongness of the actions on a scale ranging from -5 (= *very wrong*) to +5 (= *completely acceptable*).

Results for Serbian Replication

A mixed ANOVA with temporal distance (near vs. distant future) as between-subject factor and story (eating one's dead pet, sexual intercourse with sibling, dusting with national flag) as within-subject factor yielded a main effect of story, $F(2, 115) = 167.29$, $p < .001$, $\eta^2 = .59$, indicating that the wrongness of events were judged differently ("incest": $M = -3.93$[4], $SD = 0.21$; "dog": $M = -3.82$, $SD = 0.19$; "flag": $M = 0.59$, $SD = 0.32$).

More importantly and in contrast to original EL&T's study (Study 2, EL&T, 2008), there was no main effect of temporal distance, $F(1, 115) = 0.11$, $p = .746$, $\eta^2 = .001$, $g = -.06$, [CI = -0.428, 0.310], meaning that distant future transgressions were judged just as unacceptable ($M = -2.32$; $SD = 2.13$) as near future transgressions ($M = -2.44$; $SD = 1.91$). Across three stories, there was a marginally significant effect of distance only on one ("dog," $M_{\text{close}} = -4.19$, $SD = 1.51$; $M_{\text{dist}} = -3.44$, $SD = 2.48$, $F(1, 115) = 3.87$, $p = .052$, $\eta^2 = .03$).

Scaled JZS Bayes Factor (1.12) supported the null hypothesis, indicating that two temporal distance groups did not differ from each other.[5]

Results for Aggregated Data From Three Laboratories

A mixed ANOVA with temporal distance (near vs. distant future) as between-subject factor and story (1–3) as within-subject factor again yielded a strong main effect of story, $F(2, 209) = 186.61$, $p < .001$, $\eta^2 = .47$ ("incest": $M = -4.06$, $SD = 0.14$; "dog": $M = -3.93$, $SD = 0.13$; "flag": $M = -0.50$, $SD = 0.24$), meaning that wrongness of different transgressions was judged differently. As for the central research hypothesis, we ended up with conflicting findings: The original study found the expected impact of temporal distance to the wrongness assessment; the direct replication with the American sample yielded the opposite pattern of results, whereas direct replication with the Serbian sample yielded no significant effect. Analysis on the integrated sample revealed no effect of temporal distance $F(1, 209) = 0.25$, $p = .618$ $\eta^2 = .001$, $g = -0.069$, [CI = -0.323, 0.186] (distant future transgressions: $M = -2.76$; $SD = 2.03$; near future transgressions: $M = -\ 2.89$; $SD = 1.72$).

Further analysis revealed the significant effect of "laboratory" as a factor, $F(2, 209) = 11.80$, $p < .001$, $\eta^2 = .10$ and a significant interaction between story and

[4] We did not multiply the wrongness assessments by -1, as in in Eyal et al. study, since we obtained larger dispersion of raw scores, out of which some were positive and in the vignette "flag" the mean score was even positive.

[5] Bayes factors were calculated using online calculator provided by University of Missouri. It can be retrieved at http://pcl.missouri.edu/bayesfactor

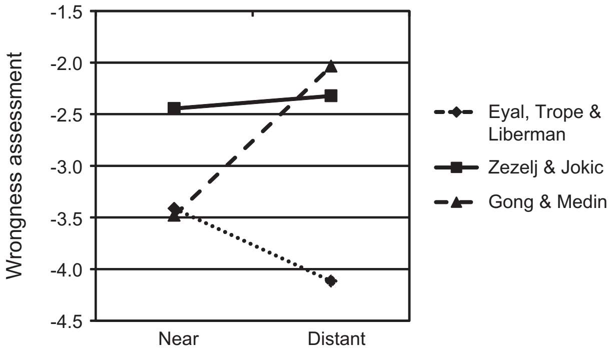

Figure 1. Acceptability of moral transgressions by temporal distance in three studies.

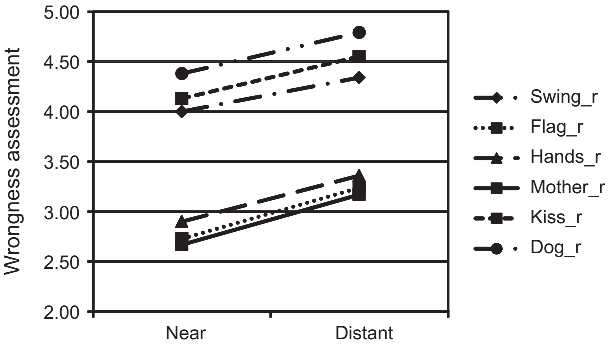

Figure 2. Wrongness of moral transgressions by social distance across stories.

laboratory, $F(4, 209) = 10.76$, $p < .001$, $\eta^2 = .09$, indicating cultural differences in assessment of wrongness of different immoral acts.[6] More importantly, there was also an interaction between temporal distance and "laboratory", $F(2, 209) = 4.08$, $p = .018$, $\eta^2 = .04$. A post hoc Tukey test showed that the Israeli laboratory differed significantly from both Serbian one (at $p < .001$) and American one (at $p = .022$), while Serbian and American laboratories did not differ from one another (see Figure 1).

Study 2

Study 2 was designed to build upon the results of Study 1 by manipulation of social distance (self vs. other) rather than temporal, and to extend the number of acts to be evaluated (six instead of three). It presents a direct replication of EL&T's study 3.

Method

Participants

Participants were 105 undergraduate students from University of Belgrade, Serbia. The original Israeli study included 40 participants, so aggregated database consisted of 145 participants.

Procedure

Participants read six vignettes. Out of those, two were the same as in Study 1 (eating a dog, dusting with a national flag) and the additional four were adopted from Haidt et al. (1993) (a girl pushing another kid off a swing, cousins kissing each other on the mouth, a man breaking a promise to his dying parent, and a man who ate with his hands in public). Participants were asked to think about a specific person they knew (high social distance condition) or to focus on their own feelings and thoughts (low social distance condition), and to judge the events either from a third person viewpoint or from a first person viewpoint. They evaluated the wrongness of the actions on a scale ranging from 1 (*not acceptable*) to 5 (*completely acceptable*).

Results for Serbian Replication

In order to unify data from replication and original databases, we reversed the participants' ratings so that higher ratings indicated higher wrongness scores. A mixed ANOVA with social distance (self vs. other) as between-subject factor and story (1–6) as within-subject factor yielded a main effect of story, $F(5, 104) = 52.01$, $p < .001$, $\eta^2 = .34$, indicating that the wrongness of events were judged differently. More importantly, there was also a main effect of social distance, $F(1, 104) = 9.90$, $p = .002$, $\eta^2 = .09$, $g = 0.615$, [CI = 0.479, 0.751]. All actions were judged as more wrong from a third person perspective ($M = 3.91$, $SD = 0.61$) then from a first person perspective ($M = 3.47$, $SD = 0.80$), which was in accordance with CLT predictions and the results of EL&T's original Study 3. Across six scenarios, the effect was significant on two ("dog", $F(1, 104) = 5.17$, $p = .025$, $\eta^2 = .05$; "broken promise", $F(1, 104) = 4.73$, $p = .032$, $\eta^2 = .04$), and marginally significant on three ("swing", $F(1, 104) = 3.39$, $p = .068$, $\eta^2 = .03$; "flag", $F(1, 104) = 3.34$, $p = .070$, $\eta^2 = .03$; "kiss", $F(1, 104) = 2.89$, $p = .092$, $\eta^2 = .03$). The pattern of means of six stories across social distance is presented in Figure 2.

Scaled JZS Bayes Factor which was lower than one (0.07) strongly supported the alternative hypothesis, indicating that two social distance groups significantly differed from each other.

[6] This difference largely stems from different assessment in the vignette "flag": while Israeli and American participants thought this was an immoral act ($M = -2.48$, $SD = 2.35$; $M = -0.83$, $SD = 3.45$), Serbian participants even viewed it as somewhat positive ($M = 0.59$, $SD = 3.49$). As we anticipated this might happen, we made a note explaining potential reasons in our preregistration proposal (https://osf.io/qhagec).

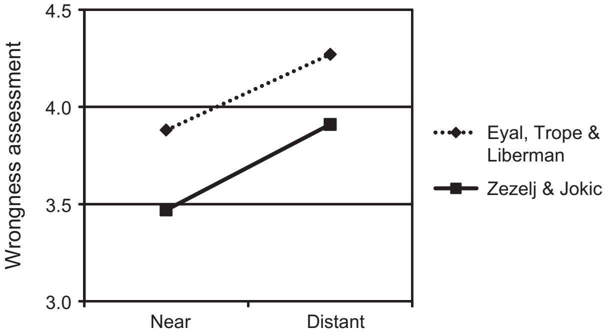

Figure 3. Wrongness of moral transgressions by social distance in two studies.

Results for Aggregated Data From Two Laboratories

As expected, mixed ANOVA with social distance (self vs. other) as between-subject factor and story (1–6) as within-subject factor again yielded a main effect of story, $F(5, 144) = 68.83$, $p < .001$, $\eta^2 = .33$. The main effect of social distance was once again significant, $F(1, 144) = 13.77$, $p < .001$, $\eta^2 = .09$, $g = 0.624$ [CI = 0.513, 0.736] (means are detailed in Figure 3).

Further analysis revealed the main effect of laboratory as a factor, $F(1, 144) = 9.36$, $p = .003$, $\eta^2 = .06$. There was no interaction between laboratory and social distance factors, and as in previous study, there was a significant interaction between story and laboratory that emerged, $F(5, 144) = 9.05$, $p < .001$, $\eta^2 = .06$.

Study 3

The main objective of Study 3 was to explore the effects of temporal distance on moral judgment of virtuous acts performed under attenuating circumstances (e.g., a fashion company donates to the poor and it positively affects its sales rate). In EL&T's original Study 4, higher distance lead to more positive virtuousness ratings, that is ascribing less weight to attenuating contextual information.

Method

Participants

Participants were 84 undergraduate students from Belgrade University, Serbia. They were randomized into "near" or "distant" future condition. As the original Israeli study included 47 participants, the aggregated database comprised of 131 in total.

Procedure

Participants were presented with three vignettes describing virtuous acts followed by extenuating contextual information. They were asked to imagine a described event occurring the next day (low temporal distance) or in a year (high temporal distance). After that they evaluated the virtuousness of each act on a scale anchored with 1 (*not at all virtuous*) to 7 (*extremely virtuous*).

Results for Serbian Replication

We conducted a mixed ANOVA with virtuousness ratings as a dependent variable, temporal distance (near vs. distant future) as a between-subject factor and story (1–3) as a within-subject factor. Results yielded a main effect of story, $F(2, 83) = 16.68, p < .001, \eta^2 = .17$, indicating that the virtuousness of events was judged differently. Same as in our Study 1 and in contrast to EL&T's original Study 4, there was no main effect of temporal distance, $F(1, 83) = 1.46$, $p = .23$, $\eta^2 = .02$, $g = -0.261$, [CI = −0.521, −0.002], meaning that the virtuousness of distant future acts ($M = 4.55$; $SD = 1.27$) was judged the same as the virtuousness of near future acts ($M = 4.87$; $SD = 1.15$), with no significant differences across the stories.

Scaled JZS Bayes Factor higher than one (3.04) supported the null hypothesis, indicating that two temporal distance groups did not differ from one another.

Results for Aggregated Data From Two Laboratories

A mixed ANOVA with temporal distance (near vs. distant future) as between-subject factor, and story (1–3) as within-subject factor again yielded a main effect of story, $F(2, 130) = 14.97$, $p < .001$, $\eta^2 = .10$. There was no main effect of temporal distance, $F(1, 130) = 0.11$, $p = .737$, $\eta^2 = .001$, $g = 0.061$, [CI = −0.136, 0.257] meaning that the virtuousness of distant future acts ($M = 4.61$; $SD = 1.13$) was judged the same as the virtuousness of near future acts ($M = 4.54$; $SD = 1.17$).

Further analysis revealed the marginally significant effect of laboratory as a factor, $F(1, 130) = 3.31$, $p = .071$, $\eta^2 = .025$ (see Figure 4). There was also an interaction between temporal distance and laboratory, $F(1, 130) = 6.79$, $p = .01$, $\eta^2 = .05$, and as in previous studies, an interaction between story and laboratory, $F(2, 130) = 4.80$, $p = .009$, $\eta^2 = .04$.

Study 4

In this study we introduced a direct manipulation of construal level as employed in G&M's (2012) first study. We primed our participants with a series of "how-and-why" questions that were expected to directly activate either a low or high construal mindset. By asking subjects to generate subordinate goals, we led them to adopt an instrumental, lower-level perspective. In contrast, generating superordinate goals led them to adopt a higher-level perspective. We aimed to explore if priming a high-or-low construal

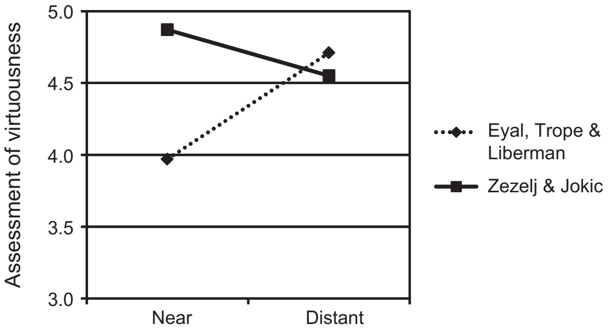

Figure 4. Virtuousness of moral acts by temporal distance in two studies.

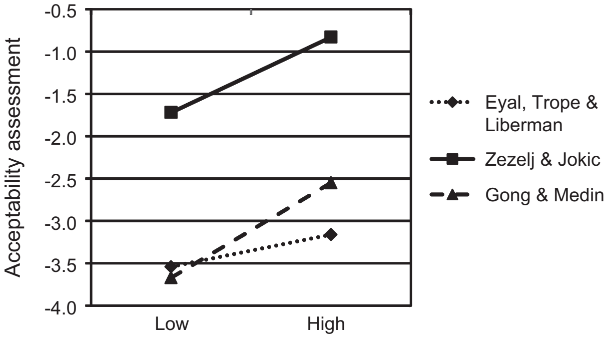

Figure 5. Acceptability of moral transgressions by construal level in three studies.

level would have an impact on the severity of our participants' moral judgment. In G&M's study, this manipulation yielded results contradictory to those of EL&T: Low-level construals led to harsher condemnation of moral transgressions in comparison to high-level construals.

Method

Participants

Participants were 48 undergraduate students from the University of Belgrade, Serbia, randomly assigned to either how or why priming condition. There were 34 participants in G&M's original study, while in EL&T's replication study there were 81, which left our aggregated database with 163 participants.

Procedure

Participants completed both priming and evaluation task in one session, ostensibly as two independent studies. Half of the participants was asked *how* they could improve and maintain health, while the other half was asked *why* they should improve and maintain health. After stating the first reason, they were asked to respond to that reason in the same vain (i.e., how or why). They repeated this process four times, filling in a diagram. Upon finishing this task, participants were presented with four scenarios of moral violation (three vignetess were the same as in Study 1 of this paper, plus one additional of a student cheating in an exam). Respondents rated moral acceptability of each act on an 11-point scale ranging from −5 (= *extremely unacceptable*) to 5 (= *extremely acceptable*).

Results for Serbian Sample

We conducted a two (high vs. low level construal) by four (story 1–4) mixed design ANOVA on moral acceptability ratings, with construal as between-subjects and story as within-subjects factor. Results yielded a main effect of story, $F(3, 47) = 85.31$, $p < .001$, $\eta^2 = .65$. More importantly, there was also a main effect of construal level, $F(1, 47) = 5.68$, $p = .021$, $\eta^2 = .11$, $g = -0.681$, [CI = -1.045, -0.317], meaning that the wrongness of transgressions was judged in accordance to G&M's results. Participants judged the acts more harshly after low-level construal priming ($M = -1.72$, $SD = 1.55$), then after high-level construal priming ($M = -0.83$, $SD = 0.95$). Across four stories, the effect was significant in one ("flag," $M_{\text{low}} = 1.25$, $SD = 3.54$; $M_{\text{high}} = 3.50$, $SD = 2.00$, $F(1, 47) = 7.34$, $p = .009$).

Scaled JZS Bayes Factor (0.37) indicated scarce evidence for the alternative hypothesis.

Results for Aggregated Data From Three Laboratories

A mixed ANOVA with construal level (low vs. high) as between-subject factor and story (1–4) as within-subject factor again yielded a main effect of story, $F(3, 162) = 54.56$, $p < .001$, $\eta^2 = .25$. There was also a main effect of the construal level, $F(1, 162) = 8.81$, $p = .003$, $\eta^2 = .05$, $g = -0.338$, [CI = -0.564, -0.113] in contrast to CLT prediction: Low level priming led to harsher wrongness assessment ($M = -3.03$, $SD = 1.41$) than high level priming ($M = -2.35$, $SD = 1.53$).

Further analysis revealed the significant effect of the laboratory, $F(2, 162) = 51.78$, $p < .001$, $\eta^2 = .40$. A post hoc Tukey test demonstrated that the Serbian laboratory differed significantly from both Israeli and American, at $p < .001$), whereas Israeli and American laboratories did not differ from one another (as can be seen in Figure 5). Once more we discovered an interaction between story and laboratory, $F(6, 162) = 26.98$, $p < .001$, $\eta^2 = .26$. The difference was mainly due to the fact that Serbian participants judged two transgressions less harshly than the Israeli and American: "flag" and "cheat" (both at $p < .001$). We have already addressed the former; the latter might be because the concept of academic honesty is more vague in Serbian than in Israely/American university setting, with no explicit (written) ethical guidelines.

Table 2. Results summary of multiple studies

Study	Eyal, Liberman, and Trope, Hedge's *g* ± CI[7]	Zezelj and Jokic, Hedge's *g* ± CI	Gong and Medin, Hedge's *g* ± CI	Do the studies agree about the direction of the effect?	What is the pattern of statistical significance?	Is the effect size from original study within the CI of the Zezelj and Jokic study?
1 (EL&T Study 2, 2008)	0.66 ± 0.27	−0.06 ± 0.37	−0.78 ± 0.59	No	Zezelj and Jokic not, other two significant	No
2 (EL&T Study 3, 2008)	0.71 ± 0.17	0.61 ± 0.14		Yes	Both significant	Yes
3 (EL&T Study 4, 2008)	0.80 ± 0.26	−0.26 ± 0.26		No	Eyal et al. significant, Zezelj and Jokic not	No
4 (G&M Study 1, 2012)	−0.34 ± 0.24	−0.68 ± 0.36	−1.06 ± 0.35	Yes	Eyal et al. not, other two significant	No

Notes. As in some of the original studies samples were relatively small ($n < 20$), and Cohen's *d* gives a biased estimate of the population effect size especially for small samples, we opted for *corrected effect size*, Hedges's *g* (recommended in Cumming, 2013; Lakens, 2013).

Following the recommendations from Valentine et al. (2011) and Lakens (2013), we summarized the results of four experiments in Table 2.

Discussion

To set the right tone for this discussion, we must first acknowledge that there is no such thing as an exact replication – there are always known and unknown factors that possibly contribute to a certain outcome that might not be identical between two studies (from characteristics of participants to physical conditions and time of day). We share the view of other social scientists (e.g., Asendorpf et al., 2013; Cacioppo & Cacioppo, 2013; Spellman, 2013) that this fact, however, should not discourage researchers from performing replications, as it is a necessary step for further generalization and/or establishing the limits of a certain phenomenon. In our replication study, we put every effort to have identical stimuli, procedure and participants as in original studies. We were fortunate enough to have the full cooperation from the other two laboratories – they were fast in sharing materials, instruments, and databases upon our request. The only known difference between the studies was the fact that samples were drawn from different cultures. Strictly speaking, this should not be decisive to the main manipulation effect: Cognitive construals should be sensitive to psychological distance, and this mechanism should serve self-regulatory needs, that are culturally invariant. However, EL&T application of CLT was in the domain of moral reasoning, which has proven to be culturally sensitive (e.g., Boyes & Walker, 1988; Snarey, 1985; Tsui & Windsor, 2001). Moreover, there were attempts to attribute current inconclusive results to cultural differences (G&M, 2012), which is why a direct replication of the original experiments in different cultural setting seemed to be an appropriate starting point.

The original study's hypotheses were derived from the premises of Construal Level Theory (Trope & Liberman, 2010). CLT proposes that distant events are represented more abstractly, globally than events that are psychologically closer. Therefore, a distant event should be evaluated in terms of more primary, high-level features. Applied in the domain of moral reasoning, this translates to expectation that one's reliance on universal moral principles and neglect to the detail should be enhanced in evaluating psychologically distant events.

We begun our investigation following two conflicting sets of findings: One supporting the conclusion that high-level construal leads to less sensitivity to context – therefore to harsher judgment of moral transgressions and more appreciation to virtuous behavior (EL&T, 2008), and the other supporting the conclusion that low construal leads to more sensitivity to context and therefore to an opposite moral evaluation of acts (G&M, 2012). Given that the cooperative efforts of the two research groups did not lead

[7] Effect sizes were calculated using De Fife (2009). Effect size calculator, Emory University. Retrieved from http://www.psychsystems.net/Manuals/

to a resolution, it was necessary to further disentangle this puzzle.

Our four attempts to replicate experiments investigating the impact of psychological distance on moral judgment yielded three different outcomes: No systematic effect of temporal distance (regardless of the nature of the act: Transgression or virtuous), the effect of social distance compatible with CLT predictions, and the reversed effect of direct construal level manipulation.

There is accumulated empirical evidence demonstrating that both temporal and social distance indeed affects moral judgment. For example, research done in Sweden (Agerström & Björklund, 2009; Agerström, Björklund, & Carlsson, 2013) concluded that people make more extreme moral judgments of behavior from a distant than from a near time (or visual) perspective, and that this effect was driven by the level of construal. Lammers (2012) demonstrated in four studies that subjects reacted more negatively to *others*' morally questionable behaviors when they took an abstract (high-level) perspective rather than a concrete (low-level) perspective. However, they were inclined to react less negatively to *their own* moral transgressions.

CLT does not assume different impact of temporal and social distance – in fact, it was initially a theory of temporal construal, exploring the effects of time perspective on mental representation (Liberman, Sagristano, & Trope, 2002; Liberman & Trope, 1998; Trope & Liberman, 2003). Only later it was generalized to other forms of psychological perspectives, namely social, spatial, and hypothetical (Trope & Liberman, 2010). It is therefore difficult to speculate the reasons as to why the temporal distance manipulation showed no consistent impact on morality judgments.

The effects of direct priming of construal level within the "how-and-why" task were in line with the results obtained by G&M. This task was designed to activate a low construal mindset (through a series of "how questions" emphasizing the means by which activities are carried out) or a high construal mindset (through a series of "why questions" emphasizing the end state activities lead to). The priming procedure was developed by Freitas, Gollwitzer, and Trope (2004), and successfully implemented for this purpose in other experiments within CLT framework (e.g., Wakslak & Trope, 2009). In our research it was expected to shift respondents' focus in evaluation of immoral acts from moral universalities to contextual information that attenuated the severity of the act. However, it could be the case that if a person was prompted to think about the means of immoral acts, they tend to focus on the act itself and thus represent it very concretely and vividly. This representation could invoke strong emotional responses. Alternatively, it could be argued that the details provided with intention to mitigate the severity of the act could have been perceived as implausible excuses and therefore dismissed. Both could paradoxically lead to harsher instead of leaner judgments.

However, the obtained pattern of results is not easily attributable to cultural differences or to unexpected consequences of provided contextual information. Had we discovered, for example, that a low construal mindset consistently leads to a harsher judgment (as G&M had), we could have speculated that attenuating information would not be plausible to our respondents so that focusing on them did not have the expected effect. Yet we had the strongest effect of social distance manipulation, in which the judgment was harsher for a high construal mindset (i.e., judging moral behavior from a third person perspective).What our replication venture seems to show is that there is a complex interplay between (A) domain of judgment (moral judgment seems to be very specific), (B) procedures employed to invoke a specific mindset (direct priming, temporal distance, and social distance manipulation yield different results in morality assessment, maybe because they do not all necessarily lead to focus on moral universalities or context, as expected), and (C) the ethnicity/culture/society that respondents are recruited from.

Future research could benefit from developing clear manipulation checks aiming to assess if different priming techniques really lead to different levels of mental construal. It could also seek to directly compare the effects of different procedures (social, temporal distance, and direct priming) on moral judgments. This in turn could help in establishing the limits of CLT's generalizability as one of the most promising and influential theories in the field.

Note From the Editors

Commentaries and a rejoinder on this paper are available (Eyal, Liberman, & Trope, 2014; Gong & Medin, 2014; Žeželj & Jokić, 2014; doi: 10.1027/1864-9335/a000206).

Acknowledgments

This research was supported by a grant from Center for Open Science (Grant No. 2013005). We are very grateful to Tal Eyal and Han Gong for their full cooperation during all phases of the research, to Dunja Anzelm and Ivan Grahek for their help with data collection and to Natasha Indiana Cordeaux for editing the manuscript. Designed research: I. Ž., J. B.; Performed research: J. B., I. Ž.; Analyzed data: J. B., I. Ž.; Wrote paper: I. Ž., J. B. The authors declare no conflict-of-interest with the content of this article. Materials, data, and the preregistered proposal are available on the project page on the Open Science Framework: https://osf.io/Z5TE6/.

PREREGISTERED OPEN DATA OPEN MATERIALS

References

Agerström, J., & Bjorklund, F. (2009). Moral concerns are greater for temporally distant events and are moderated by value strength. *Social Cognition, 27*, 261–282.

Agerström, J., Björklund, F., & Carlsson, R. (2013). Look at yourself! Visual perspective influences moral judgment by level of mental construal. *Social Psychology, 44*, 42–46.

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

Boyes, M. C., & Walker, L. J. (1988). Implications of cultural diversity for the universality claims of Kohlberg's theory of moral reasoning. *Human Development, 31*, 44–59.

Brislin, R. W. (1970). Back-translation for cross-cultural research. *Journal of Cross-Cultural Psychology, 1*, 185–216.

Brislin, R. W. (1976). Comparative research methodology: Cross-cultural studies. *International Journal of Psychology, 11*, 215–229.

Cacioppo, J.T., & Cacioppo, S. (2013). Minimal replicability, generalizability and scientific advances in psychological science. *European Journal of Personality, 27*, 121–122.

Cumming, G. (2013). The new statistics: A how-to guide. *Australian Psychologist, 48*, 161–170.

Eyal, T., Liberman, N., & Trope, Y. (2008). Judging near and distant virtue and vice. *Journal of Experimental Social Psychology, 44*, 1204–1209.

Eyal, T., Liberman, N., & Trope, Y. (2014). A comment on Žeželj and Jokić replication. Commentaries and rejoinder to Žeželj and Jokić (2014). *Social Psychology*. Advance online publication. doi: 10.1027/1864-9335/a000206.

Freitas, A., Gollwitzer, P., & Trope, Y. (2004). The influence of abstract and concrete mindsets on anticipating and guiding others'self-regulatory efforts. *Journal of Experimental Social Psychology, 40*, 739–752.

Gong, H., & Medin, D.L. (2012). Construal levels and moral judgment: Some complications. *Judgment and Decision Making, 7*, 628–638.

Gong, H., & Medin, D. L. (2014). Commentary on Žeželj and Jokić (2014). Commentaries and rejoinder to Žeželj and Jokić (2014). *Social Psychology*. Advance online publication. doi: 10.1027/1864-9335/a000206

Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review, 108*, 814–834.

Haidt, J., Koller, S., & Dias, M. (1993). Affect, culture, and morality, or is it wrong to eat your dog? *Journal of Personality and Social Psychology, 65*, 613–628.

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for *t*-tests and ANOVAs. *Frontiers in Psychology, 4*, 863–900.

Lammers, J. (2012). Abstraction increases hypocrisy. *Journal of Experimental Social Psychology, 48*, 475–480.

Liberman, N., & Trope, Y. (1998). The role of feasibility and desirability considerations in near and distant future decisions: A test of temporal construal theory. *Journal of Personality and Social Psychology, 75*, 5–18.

Liberman, N., Sagristano, M. D., & Trope, Y. (2002). The effect of temporal distance on level of mental construal. *Journal of Experimental Social Psychology, 38*, 523–534.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian *t* tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review, 16*, 225–237.

Shapira, O., Liberman, N., Trope, N., & Rim, S. (2012). Levels of mental construal. In S. Fiske & N. McRae (Eds.), *Sage handbook of Social Cognition* (pp. 229–250). London: Sage.

Snarey, J. R. (1985). Cross-cultural universality of social-moral development: A critical review of Kohlbergian research. *Psychological Bulletin, 97*, 202.

Spellman, B. A. (2013). There is no such thing as replication, but we should do it anyway. *European Journal of Personality, 27*, 136–137.

Trope, Y., & Liberman, N. (2003). Temporal construal. *Psychological Review, 110*, 403–421.

Trope, Y., & Liberman, N. (2010). Construal-level theory of psychological distance. *Psychological Review, 117*, 440–463.

Trope, Y., & Liberman, N. (2011). Construal level theory. In P. Van Lange, A. W. Kruglanski, & E. T. Higgins (Eds.), *Handbook of theories of social psychology* (pp. 118–135). London, UK: Sage.

Tsui, J., & Windsor, C. (2001). Some cross-cultural evidence on ethical reasoning. *Journal of Business Ethics, 31*, 143–150.

Valentine, J. C., Biglan, A., Boruch, R. F., Castro, F. G., Collins, L. M., Flay, B.R., … Schinke, S. P. (2011). Replication in prevention science. *Prevention Science, 12*, 103–117.

Wakslak, C., & Trope, Y. (2009). The effect of construal level on subjective probability estimates. *Psychological Science, 20*, 52–58.

Žeželj, I. L., & Jokić, B. R. (2014). A rejoinder to comments by Eyal, Liberman & Trope and Gong & Medin. Commentaries and rejoinder to Zezelj and Jokic (2014). *Social Psychology*. Advance online publication. doi: 10.1027/1864-9335/a000206

Received February 21, 2013
Accepted November 27, 2013
Published online May 19, 2014

Iris Žeželj

Faculty of Philosophy
University of Belgrade
Čika Ljubina 18-20
11000 Belgrade
Serbia
E-mail izezelj@f.bg.ac.rs

Original Article

Three Attempts to Replicate the Moral Licensing Effect

Irene Blanken,[1] Niels van de Ven,[1] Marcel Zeelenberg,[1] and Marijn H.C. Meijers[2]

[1]Tilburg University, The Netherlands, [2]University of Amsterdam, The Netherlands

Abstract. The present work includes three attempts to replicate the moral licensing effect by Sachdeva, Iliev, and Medin (2009). The original authors found that writing about positive traits led to lower donations to charity and decreased cooperative behavior. The first two replication attempts (student samples, 95% power based on the initial findings, $N_{\text{Study1}} = 105$, $N_{\text{Study2}} = 150$), did not confirm the original results. The third replication attempt (MTurk sample, 95% power based on a meta-analysis on self-licensing, $N = 940$) also did not confirm the moral licensing effect. We conclude that (1) there is as of yet no strong support for the moral self-regulation framework proposed in Sachdeva et al. (2009) (2) the manipulation used is unlikely to induce moral licensing, and (3) studies on moral licensing should use a neutral control condition.

Keywords: moral licensing, moral cleansing, self-regulation, replication

People like to present themselves as good people, both to themselves and to others, to maintain a positive self-image and to feel like a moral person (Aronson, Cohen, & Nail, 1999; Schlenker, 1980; Steele, 1988). Furthermore, central theories of human behavior highlight humans' desire for cognitive consistency in their thoughts, feelings, and behavior (Festinger, 1957; Heider, 1946). Intriguing research on *moral licensing* qualifies this desire for consistency by suggesting that individuals who behave in a morally laudable way, later feel more justified to perform a morally questionable action (Merritt, Effron, & Monin, 2010; Miller & Effron, 2010). Moral licensing is found to lead to a broad spectrum of undesirable behaviors. For example, after (reminders of) prior moral or socially desirable behavior people displayed more prejudiced attitudes (Effron, Cameron, & Monin, 2009; Monin & Miller, 2001), cheated more (Jordan, Mullen, & Murninghan, 2011; Mazar & Zhong, 2010), displayed a preference for hedonic over utilitarian products (Khan & Dhar, 2006), and indulged more in highly palatable foods (Mukhopadhyay, Sengupta, & Ramanathan, 2008).

An important contribution to the literature on moral licensing examines how writing about one's own positive or negative traits can influence donations to charity and cooperative behavior in a commons dilemma (Sachdeva, Iliev, & Medin, 2009). In just 4 years since publication, this paper has been cited 129 times (Google Scholar, November 27, 2013). Based on their findings, the authors argued that this moral licensing effect can best be interpreted as part of a larger moral self-regulation framework where internal balancing of moral self-worth and the costs associated with prosocial behavior determine whether one will display (im)moral behavior. When the moral image of oneself is established, an immoral action is allowed without the fear of losing that moral image (moral licensing). However, when one appears immoral to others, positive actions are needed to restore the moral image (moral cleansing). The studies of Sachdeva et al. (2009) comparing licensing with neutral control conditions show medium-sized effect sizes ($d = 0.62$ ([CL_{95}] -0.11 to 1.35) for Study 1 and $d = 0.59$ ([CL_{95}] -0.12 to 1.30) for Study 3).[1] However, note that because of the small sample sizes ($N = 14$ to 17 per condition), the obtained effects have large variances, implying that the true effect sizes could range from very small to very large.

There are no published direct replication attempts of the methodologies of Sachdeva et al.'s (2009) studies. Conway and Peetz (2012) conducted a study that was similar to Sachdeva et al.'s Study 1. However, this was not a direct replication because they adapted the procedure and added extra manipulations. We sought to replicate the studies by Sachdeva et al. to obtain additional insight in the complete moral self-regulation framework by testing for both moral licensing and moral cleansing effects contrasted to a neutral control condition.

We conducted high-powered replications of Sachdeva et al.'s (2009) Study 1 and Study 3 in Dutch student samples with 95% statistical power based on the effect size of the original studies. We did a third study with a US sample via Amazon's MTurk with 95% power based on

[1] Note that the overall differences between the three conditions (moral licensing, moral cleansing, and the neutral control condition) of Sachdeva et al.'s Study 1 and Study 3 were significant. For Study 1, no statistics on post hoc comparisons were reported. When calculating the Cohen's *d* effect sizes comparing the moral licensing with the neutral control conditions, we found that for both studies, the confidence intervals included zero, indicating marginally significant moral licensing effects.

DOI: 10.1027/1864-9335/a000189

the effect size that we obtained in our meta-analysis on self-licensing ($d = 0.26$; Blanken, Van de Ven, & Zeelenberg, 2014). This study examined both dependent variables of original Studies 1 and 3 in a counterbalanced order. For all studies, we report how we determined our sample sizes, all data exclusions, all manipulations, and all measures.

Study 1 – Replication of Sachdeva et al.'s (2009) Study 1

Participants

Using G*Power (Faul, Erdfelder, Buchner, & Lang, 2009) we calculated that at least 63 participants were needed to achieve 95% power for the effect size of Sachdeva et al.'s Study 1 (2009; $N = 46$). We planned to collect data for one full week, and our sample consisted of 106 undergraduate students who participated for course credit. One was removed because this participant indicated a willingness to donate €100 to charity, more than 32 standard deviations from the mean donation response. The remaining 105 (25 males, 78 females, 2 unknown, $M_{\text{age}} = 19.58$) participants included native Dutch students (83.3%), non-native Dutch students (9.5%), and foreign students (4.8%). Participants were randomly assigned to either the positive trait ($N = 35$), negative trait ($N = 34$), or neutral control condition ($N = 36$).

Materials and Procedure

Participants completed the study as the first of a series of experiments behind separate desks in the laboratory. The experimenter in the laboratory was blind to condition. Prior to the experiment, participants provided their informed consent. The experimenter guided participants to their desks and instructed them to complete the paper-and-pencil questionnaire.

We obtained the original paper-and-pencil questionnaire from Sachdeva et al. (2009) and translated these materials into Dutch (for all materials see supplements section). The cover story indicated that the study was about handwriting styles. Depending on the assigned condition, participants were exposed to either nine positive trait words, nine negative trait words, or nine neutral words and were asked to copy each word four times and think about each word for 5–10 s. Next, participants were asked to write a short story about themselves including the words they just copied.

After this manipulation, participants responded to some filler items. Subsequently, the main dependent variable was presented. Participants read that the laboratory, in an effort to increase social responsibility, asked all participants whether they would like to contribute to a worthy cause. If they would like to do so, they could pledge to make a small donation to any good cause of their choice. They were told that they would be reminded of their choice at a later time via a confirmation e-mail from the experimenter. Participants could select to which cause(s) they would like to donate (cancer research, animal rights, ending world hunger, environmental preservation, human rights, veteran's affairs, or other) and how much they would be willing to donate (from €0 up to €10 or another specified amount). Finally, participants completed seven self-presentation items from the Self-Monitoring scale (Lennox & Wolfe, 1984) and a set of demographic measures.

Known Differences From Original Study

The only known difference between our replication and the original Studies 1 and 2 of Sachdeva et al. (2009) was that we ran this study in a laboratory at a Dutch university, while the original study was conducted in a laboratory at a USA university. When participants were asked to write about the positive trait, neutral, or negative trait words, we used the exact instruction of the original Study 2, which explicitly stated that participants should use the nine given words to *write a story about themselves*. This was not done in Sachdeva et al.'s Study 1, although it was intended that participants would do so. As such, for this replication, we combined the best of Sachdeva et al.'s Study 1 (including a control condition) and Study 2 (the manipulation with the clearest instruction).

Results

Following our confirmatory analysis plan, we conducted Sachdeva et al.'s (2009) analysis to test the effect of writing about one's own positive traits, negative traits, or neutral words on donation amount. Table 1 contains the mean responses per condition and statistical tests. There were no significant differences between the moral identity conditions on donation amount.[2] The results of an additional regression model including gender, age, and ethnicity indicated that none of these factors significantly predicted donation amount (all ps $\geq .321$). A reviewer suggested that self-monitoring might moderate the observed effects. It did not, $p = .086$. Analysis details for all studies are available in the supplements.

Exploratory Analysis

When reading the recalled stories, we noticed that 55.7% of the participants violated the instructions by not writing about themselves or by using the words in a negating way (for instance, "Alyssa is a generally friendly person

[2] A nonparametric independent-samples Kruskal-Wallis test (which controls for the skewness of the data), also found a nonsignificant effect, $H(2) = 0.36$, $p = .837$.

with a caring and compassionate disposition" or "I am neither a very caring nor compassionate individual"). When we only used a post hoc selection of those that wrote about their own positive traits ($N = 28$) and compared it to the neutral control condition, there was still no difference on donation amount ($p = .756$).

Study 2 – Replication of Sachdeva et al.'s (2009) Study 3

Participants

Using G*Power (Faul et al., 2009) we calculated that we should include at least 96 participants in our study to achieve 95% power for the effect size that Sachdeva et al. (2009) obtained in their Study 3 (the original used $N = 46$). We planned to collect data for one full week, and our sample consisted of 150 undergraduate students who participated for course credit (27 males, 122 females, 1 unknown, $M_{\text{age}} = 20.34$) and included native Dutch students (87.3%), non-native Dutch students (7.3%), and foreign students (4.7%). All participants were randomly assigned to either the positive trait condition ($N = 49$), the negative trait condition ($N = 52$), or the neutral control condition ($N = 49$).

Materials and Procedure

Participants first provided informed consent, and then completed the study as the first of a series of experiments. The laboratory experimenter was blind to condition. The experimenter led participants to a separate cubicle and instructed them to complete the paper-and-pencil questionnaire.

The materials were the same as those in Study 1 except that the dependent variable was a hypothetical commons dilemma. In this commons dilemma, participants imagined a scenario in which they were the manager of a midsized industrial manufacturing plant. They read that all manufacturers reached an agreement to install filters to eliminate toxic gasses and to run these filters 60% of the time. Running a filter was costly for the manufacturing plant, but would be beneficial to society. To measure cooperative behavior, participants were asked to indicate what percentage of time they would operate the filters, indicated on an 11-point scale from 0 (labeled 0%) to 10 (labeled 100%).

After the main dependent variable, participants explained their decision and completed three secondary measures; they estimated (1) the percentage of other managers who would not cooperate, on the same 11-point scale; (2) the amount of environmental damage expected when the filters would be run less than the agreed 60% on an 11-point scale from 0 (none) to 10 (a great amount); and (3) the likelihood of getting caught when operating the filters less than 60% of the time on an 11-point scale from 0 (= *impossible*) to 10 (= *certain*). Finally, participants completed the seven self-presentation items from the Self-Monitoring scale (Lennox & Wolfe, 1984) and a set of demographic measures.

Known Differences From Original Study

The only known difference compared to the original study is that we ran this study in a laboratory at a Dutch university instead of a USA university.

Results

Following our confirmatory analysis plan, we conducted Sachdeva et al.'s (2009) analysis to test the effect writing about one's own (im)moral traits on cooperation (the amount of time participants were willing to run the filters). There were no significant differences between the conditions on cooperative behavior (Tabel 1).[3] Furthermore, there were no effects on the secondary variables (Table 2). The results of an additional regression model including gender, age, and ethnicity indicated that none of these demographic variables predicted cooperative behavior (all ps $\geq .257$). Self-monitoring did not moderate the observed effects ($p = .787$).

Exploratory Analysis

We noticed that 48.5% of the participants violated the recall instructions and did not write about their own traits or used the words in a negating way. When we only used a post hoc selection of those who actually wrote about their own positive traits ($N = 42$), there was still no difference on cooperative behavior between the positive trait stories about oneself and the neutral control condition ($p = .197$).

Study 3 – Replication of Sachdeva et al.'s (2009) Study 1 and Study 3 With a General US Population Sample on MTurk

Participants

Whereas in Study 1 and Study 2, we based our sample size on a power analysis using the original studies, for Study 3 we did so based on the effect size of self-licensing that we obtained in the preliminary data of our meta-analysis ($d = 0.26$) (Blanken et al., 2014). We calculated with G*Power (Faul et al., 2009) that we would need at least 918 participants in our study to achieve 95% power to find a self-licensing effect. The sample was recruited on Mturk.

[3] A nonparametric independent-samples Kruskal-Wallis test (which controls for the not normally distributed data), also showed no effect, $H(2) = 2.87$, $p = .238$.

Table 1. Means, standard deviations, sample sizes, and test statistics for dependent variables in all studies

Dependent variable	Positive trait *M* (*SD*)	Neutral trait *M* (*SD*)	Negative trait *M* (*SD*)	*F*	*p*	η_p^2
Study 1, *N* = 105	35	36	34			
Donation amount[a]	2.89 (3.64)	2.78 (3.83)	2.35 (3.28)	0.21	.810	.004
Study 2, *N* = 150	49	49	52			
Cooperative behavior	6.29 (1.14)	5.88 (1.49)	5.84 (1.37)	1.70	.187	.023
Study 3, *N* = 940	306	326	308			
Donation amount	4.52 (2.91)	4.60 (3.11)	5.10 (3.18)	3.20	.041	.007
Cooperative behavior	6.27 (1.58)	6.39 (1.78)	6.21 (1.60)	0.78	.457	.002

Notes. For donation amount the answers could range from €0 to €10 (or participants could indicate another amount). For cooperative behavior, the answers indicate how long participants would choose to do the costly cooperative act in the scenario (run filters), on a range from 0 (*the least cooperative*) to 10 (*the most cooperative*). The statistical test of Study 3 is the main effect, controlling for order effects. Effect sizes and Confidence Intervals for the moral licensing and moral cleansing effects can be found in the Forest plots (Figures 1 and 2). [a]For the main dependent variable, participants were asked to indicate how much money they wanted to donate to a good cause. If they did not answer this question, we interpreted their response as €0. When these participants were excluded from the analysis, the results did not differ, $F(2, 81) = 0.29$, $p = .747$, $\eta_p^2 = .007$.

Table 2. Means, standard deviations, and test statistics for secondary measures in Studies 2 and 3

	Positive trait *M* (*SD*)	Neutral trait *M* (*SD*)	Negative trait *M* (*SD*)	*F*	*p*	η_p^2
Study 2, *N* = 150	49	49	52			
Expected cooperative behavior of others	4.64 (2.56)	4.57 (2.55)	4.28 (2.74)	0.26	.770	.004
Estimated likelihood of getting caught	5.92 (1.80)	5.61 (1.85)	5.62 (1.75)	0.47	.624	.006
Negative consequences for the environment	7.22 (1.62)	6.90 (1.95)	6.62 (1.88)	1.41	.248	.019
Study 3, *N* = 940	306	326	308			
Expected cooperative behavior of others	4.72 (2.91)	4.89 (3.08)	4.51 (2.86)	1.30	.273	.003
Estimated likelihood of getting caught	6.91 (2.07)	7.09 (2.14)	7.05 (2.02)	0.63	.535	.001
Negative consequences for the environment	5.84 (2.04)	5.98 (2.11)	5.93 (1.99)	0.33	.721	.001

Notes. For expected cooperative behavior of others, answers could range from 0 (the least cooperative) to 10 (the most cooperative). For estimated likelihood of getting caught, answers could range from 1 (impossible) to 10 (certain). For negative consequences for the environment, answers could range from 1 (none) to 10 (a great amount).

We included an instructional manipulation check to prevent inattentive participants from starting the study (see Oppenheimer, Meyvis, & Davidenko, 2009). Participants were asked to provide an answer to three neutral questions about stories and were explicitly instructed to answer "five" on the first question, and "seven" on the second and third question. Participants who did not follow these instructions (*N* = 160) could not participate in our study. Our final sample consisted of 940 participants (449 males and 491 females, $M_{age} = 33.41$) who participated in exchange for $1.80.[4] All participants were randomly assigned to the positive trait condition (*N* = 306), the negative trait condition (*N* = 308), or the neutral control condition (*N* = 326).

Materials and Procedure

Participants completed the study materials via the Qualtrics survey program. Participants could subscribe to participate in our study entitled "writing style and several questions" if they had an MTurk approval rate that was higher than 95% and if they lived in the US.

After finishing writing the stories with the positive traits, negative traits, or neutral words, participants answered the filler questions and both dependent measures from Sachdeva et al.'s (2009) Study 1 (donation amount) and Study 3 (cooperative behavior) in a counterbalanced order. Subsequently, participants completed the self-presentation items from the Self-Monitoring scale and a set of demographic measures.

Known Differences From Original Study

The study was conducted online. We made two slight changes to these materials to increase the credibility of the online study. First, for the cover story, we instructed participants that the study was about general writing styles instead of handwriting, as the latter would not be believable in an online study. Second, we changed the donation

[4] We set the target higher than 918 to ensure a minimum of 918 valid participants after data exclusion.

measure. We told participants that 10 of them would be randomly selected to win an additional $10 MTurk worker bonus. They were then asked that if they were one of the winners, would they be willing to donate a portion of this bonus to a cause of their choice from a list (cancer research, animal rights, ending world hunger, environmental preservation, human rights, veteran's affairs, or other). Participants selected a cause and indicated the amount they would donate ranging from $0 to $10 (or more).

Results

Donations

Following our confirmatory analysis plan, we conducted Sachdeva et al.'s (2009) analyses to test the effect of writing about (im)moral traits on how much participants would want to donate to a good cause. We controlled for order effects by including the order in which the two dependent variables were presented as a separate independent variable in the model. Order did not affect the donation amount, $F(1, 934) = 0.78$, $p = .378$, $\eta_p^2 = .001$, nor was there an interaction effect of order with the manipulation of what words participants wrote about, $F(2, 934) = 0.42$, $p = .656$, $\eta_p^2 = .001$.

As Table 1 shows, there was a main effect of moral identity condition on donation amount.[5] Post hoc Tukey tests indicated that participants in the negative trait condition donated more money than participants in the positive trait condition ($p = .044$) and participants in the neutral control condition ($p = .020$). There was no difference in donation amount between participants in the positive trait condition and participants in the neutral control condition ($p = .729$). Thus, we did not find a moral licensing effect, but we did observe a moral cleansing effect – the recall of negative traits increased subsequent moral behavior. Self-monitoring did not moderate the observed effects.

Of the demographic variables gender, age, education level, family income, and ethnicity, only age significantly influenced donation amount ($\beta = .11$, $t(930) = 3.36$, $p < .001$). When we included age as a covariate to the effect of the manipulation on donation amount, the effect of the manipulation remained significant, $F(2, 932) = 3.15$, $p = .043$, $\eta_p^2 = .007$.

Cooperative Behavior

Next, we conducted Sachdeva et al.'s (2009) analyses to test the effect of moral identity condition on cooperation in a hypothetical commons dilemma. The order in which the dependent variables were presented did affect cooperative behavior, $F(1, 934) = 11.20$, $p = .001$, $\eta_p^2 = .012$, with participants who first completed the donation dependent variable displaying slightly more cooperative behavior ($M = 6.47$, $SD = 1.77$) than participants who first completed this cooperative behavior dependent variable ($M = 6.11$, $SD = 1.52$). The interaction between moral identity and order was not significant, $F(2, 934) = 0.83$, $p = .438$, $\eta_p^2 = .002$. We do not know why this order effect exists, but for the current study it is mainly important that we control for this possible influence by adding it as a factor in the analyses. As Table 1 shows, there was no main effect of moral identity on cooperative behavior,[6] nor on the secondary variables (see Table 2). Again, self-monitoring did not moderate the observed effects.

Of the demographic variables, only one of the ethnicity dummy variables significantly influenced cooperation (with African Americans cooperating less than others, $\beta = -.19$, $t(930) = -3.11$, $p = .002$). When including ethnicity as a covariate, there was still no effect of moral identity condition on cooperative behavior, $F(2, 933) = 0.81$, $p = .447$, $\eta_p^2 = .002$.

Exploratory Analyses

We noticed that 43.6% of the participants violated the recall instructions and did not write about their own traits or used the words in a negating way. Using solely the coded stories about oneself in our analyses, there was a main effect of moral identity condition on donation amount ($p = .020$) with participants in the negative trait condition donating more money than participants in the positive trait condition ($p = .017$) and in the neutral control condition ($p = .009$). There was no main effect of moral identity condition on cooperative behavior ($p = .495$).

General Discussion

We made three attempts to replicate the findings of Sachdeva et al. (2009) on moral licensing, with samples based on precalculated power and preplanned analyses. In the first two replication attempts using student samples, the data did not confirm the original results. In our third replication attempt using a general population sample the data did not confirm the moral licensing effect. We did, however, find support for the moral cleansing effect on one of the two dependent variables in Study 3, but not in Studies 1 and 2.

Current Status of the Moral Licensing Effect

We conducted a meta-analysis of this moral licensing effect by including both the original Studies 1 and 3 by Sachdeva et al. and the three current replication attempts, using the metafor package of Viechtbauer (2010). For our Study 3,

[5] A nonparametric independent-samples Kruskal-Wallis test (which controls for the skewness of the data), found a similar effect, $H(2) = 5.85$, $p = .054$.

[6] A nonparametric independent-samples Kruskal-Wallis test (which controls for the not normally distributed data), also found a nonsignificant effect, $H(2) = 2.68$, $p = .713$.

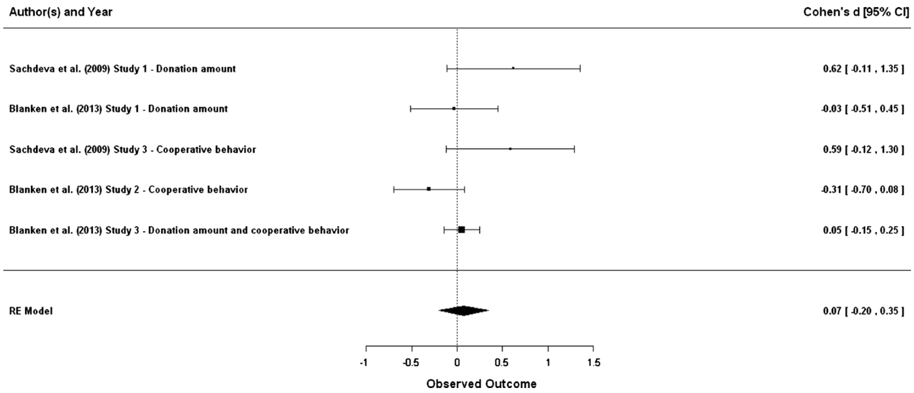

Figure 1. Forest plot including all comparisons between the *moral licensing* and neutral control conditions of the original studies by Sachdeva et al. (2009) and our replication attempts.

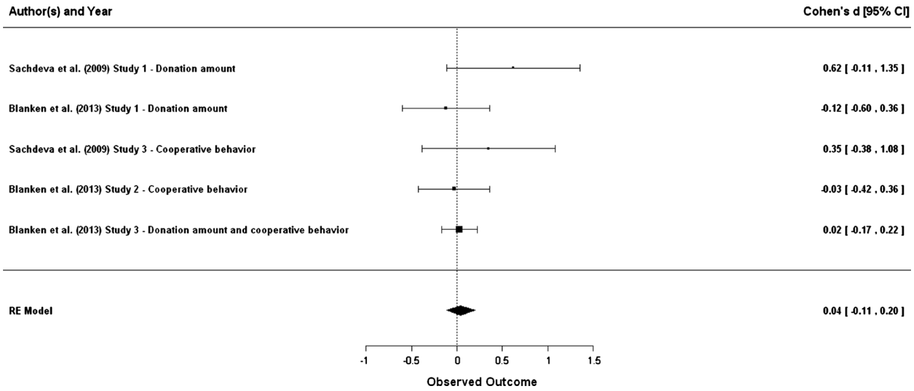

Figure 2. Forest plot including all comparisons between the *moral cleansing* and neutral control conditions of the original studies by Sachdeva et al. (2009) and our replication attempts.

we used the average effect size of the two dependent variables. The random effects meta-analysis including all five studies produced a mean effect size of moral licensing of d = 0.07 ([CL_{95}] −0.20 to 0.35). There was thus no significant moral licensing effect across studies (z = 0.52, p = .603). Figure 1 contains an overview of all moral licensing effect sizes (when compared to the neutral control conditions).

Current Status of the Moral Cleansing Effect

We conducted a meta-analysis of this moral cleansing effect by including both the original Studies 1 and 3 by Sachdeva et al. and the three current replication attempts. The random effects meta-analysis including all five studies produced a mean effect size of moral cleansing of d = 0.04 ([CL_{95}] −0.11 to 0.20). There was thus no significant moral cleansing effect across studies (z = 0.53, p = .593). Figure 2 contains an overview of all moral cleansing effect sizes (when compared to the neutral control conditions). However, note that only a small number of participants in the moral cleansing condition of the replication studies actually wrote about themselves.

Possible Limitations of Our Replication Attempts

Although we did our best to design direct replications of the original studies, differences are inevitable, and some of those may be consequential for moderating the results. First, our Studies 1 and 2 used Dutch students not US students. There is no theoretical reason to expect different licensing effects for Dutch compared to US citizens, but our pilot test (see supplements) suggested that words in the positive moral trait condition were seen to be slightly more positive in the US than in the Netherlands. Even so, the words were evaluated very positively in both national samples. Study 3 used a US based sample, but this study differed on two aspects compared to the original study. It was conducted online instead of in the laboratory, and the manipulation involved donating a part of potential winnings instead of money out-of-pocket. We cannot rule out that these procedural differences were consequential, but there presently exists no theoretical reason or identification of these as boundary conditions on moral licensing.

Conclusion

Although Sachdeva et al. (2009) theorized that moral licensing and moral cleansing should be considered jointly as being part of a moral self-regulation process, our three high-powered studies did not replicate the key moral licensing effect. Further, the meta-analytic result suggests that the present state of evidence with this paradigm is not different from a null effect. Sachdeva et al. (2009, p. 524) suggested that their findings showed that "moral-licensing and moral-cleansing effects can act convergently as part of a moral self-regulation process." Based on the present findings, we do not argue that the theory is incorrect, only that it lacks sufficient empirical support when using the Sachdeva et al. (2009) paradigm.

We suggest three concrete steps to clarify the effects of moral licensing on social judgment. First, the method used by Sachdeva et al. (2009) seems unlikely to elicit moral licensing, especially since many participants violated the recall instructions and did not write about their own traits or used the words in a negating way. This is a procedure-specific issue; it does not invalidate moral licensing more generally. Second, the meta-analysis of all licensing research suggests that the effect is relatively small (Blanken et al., 2014). Therefore, small sample studies are highly inadvisable as they would need to leverage chance to detect a result using null hypothesis significance testing. Third, because moral licensing and moral cleansing are theoretically distinct, it is important to use a neutral control condition to clarify the role of each in social judgment.

Acknowledgments

This work was supported by a grant from the Center of Open Science. The authors gratefully thank Sonya Sachdeva and her co-authors for sharing all materials and information about the methodology and for her helpful comments. The authors declare no conflict-of-interest with the content of this article. Designed research: I.B., N.V., M.Z., M.H.C.M.; Performed research: I.B.; Analyzed data: I.B.; Wrote paper: I.B., N.V., M.Z., M.H.C.M. All materials, data, and the preregistered design are available at: https://osf.io/3cmz4/.

PREREGISTERED OPEN DATA OPEN MATERIALS

References

Aronson, J., Cohen, G. L., & Nail, P. R. (1999). Self-affirmation theory: An update and appraisal. In E. Harmon-Jones & J. Mills (Eds.), *Cognitive dissonance: Progress on a pivotal theory in social psychology* (pp. 127–148). Washington, DC: American Psychological Association.

Blanken, I., Van de Ven, N., & Zeelenberg, M. (2014). *A meta-analytic review of self-licensing* Manuscript under review.

Conway, P., & Peetz, J. (2012). When does feeling moral actually make you a better person? Conceptual abstraction moderates whether past moral deeds motivate consistency or compensatory behavior. *Personality and Social Psychology Bulletin, 6*, 907–919.

Effron, D. A., Cameron, J. S., & Monin, B. (2009). Endorsing Obama licenses favoring Whites. *Journal of Experimental Social Psychology, 45*, 590–593.

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

Festinger, L. (1957). *A theory of cognitive dissonance*. Stanford, CA: Stanford University.

Heider, F. (1946). Attitudes and cognitive organization. *Journal of Psychology, 201*, 107–112.

Jordan, J., Mullen, E., & Murnighan, J. K. (2011). Striving for the moral self: The effects of recalling past moral actions on future moral behavior. *Personality and Social Psychology Bulletin, 37*, 701–713.

Khan, U., & Dhar, R. (2006). Licensing effect in consumer choice. *Journal of Marketing Research, 43*, 357–365.

Lennox, R. D., & Wolfe, R. N. (1984). Revision of the self-monitoring scale. *Journal of Personality and Social Psychology, 46*, 1349–1364. doi: 10.1037/0022-3514.46.6.1349

Mazar, N., & Zhong, C. B. (2010). Do green products make us better people? *Psychological Science, 21*, 494–498.

Merritt, A. C., Effron, D. A., & Monin, B. (2010). Moral self-licensing: When being good frees us to be bad. *Social and Personality Psychology Compass, 4*, 344–357.

Miller, D. T., & Effron, D. A. (2010). Psychological license: When it is needed and how it functions. *Advances in Experimental Social Psychology, 43*, 115–155.

Monin, B., & Miller, D. T. (2001). Moral credentials and the expression of prejudice. *Journal of Personality and Social Psychology, 81*, 33–43.

Mukhopadhyay, A., Sengupta, J., & Ramanathan, S. (2008). Recalling past temptations: An information-processing perspective on the dynamics of self-control. *Journal of Consumer Research, 35*, 586–599.

Oppenheimer, D. M., Meyvis, T., & Davidenko, N. (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology, 45*, 867–872.

Sachdeva, S., Iliev, R., & Medin, D. L. (2009). Sinning saints and saintly sinners: The paradox of moral self-regulation. *Psychological Science, 20*, 523–528.

Schlenker, B. R. (1980). *Impression management: The self-concept, social identity and interpersonal relations*. Monterey, CA: Brooks/Cole.

Steele, C. M. (1988). The psychology of self-affirmation: Sustaining the integrity of the self. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 21, pp. 261–302). New York, NY: Academic Press.

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software, 36*, 1–48.

Received February 28, 2013
Accepted December 3, 2103
Published online May 19, 2014

Irene Blanken

Department of Social Psychology
Tilburg University
PO Box 90153
5000 LE Tilburg
The Netherlands
E-mail i.blanken@tilburguniversity.edu

Replication of the Superstition and Performance Study by Damisch, Stoberock, and Mussweiler (2010)

Robert J. Calin-Jageman and Tracy L. Caldwell

Dominican University, River Forest, IL, USA

Abstract. A recent series of experiments suggests that fostering superstitions can substantially improve performance on a variety of motor and cognitive tasks (Damisch, Stoberock, & Mussweiler, 2010). We conducted two high-powered and precise replications of one of these experiments, examining if telling participants they had a lucky golf ball could improve their performance on a 10-shot golf task relative to controls. We found that the effect of superstition on performance is elusive: Participants told they had a lucky ball performed almost identically to controls. Our failure to replicate the target study was not due to lack of impact, lack of statistical power, differences in task difficulty, nor differences in participant belief in luck. A meta-analysis indicates significant heterogeneity in the effect of superstition on performance. This could be due to an unknown moderator, but no effect was observed among the studies with the strongest research designs (e.g., high power, *a priori* sampling plan).

Keywords: superstition, luck, self-efficacy, replication

Can superstitions actually improve performance? Damisch, Stoberock, and Mussweiler (2010) reported a striking experiment in which manipulating superstitious feelings markedly increased golfing ability. Participants attempted 10 putts, each from a distance of 100 cm. Some participants were primed for superstition prior to the task by being told "Here is the ball. So far it has turned out to be a lucky ball." Controls were simply told "This is the ball everyone has used so far." Remarkably, this manipulation produced a substantial increase in golf performance: Controls made 48% of putts while superstition-primed participants made 65% of putts (d = 0.83, 95% CI [0.05, 1.60]).

This simple experiment suggests a major revision to our notions of superstition. The very definition of a superstition is a belief that is "irrational" which arises due to a "false conception of causation" (*Merriam-Webster*, 2013). Indeed, there has been a long scientific tradition of pointing out the fundamental lack of efficacy of superstitious behavior (e.g., Galton, 1872). The prevalence of superstitious behavior has thus been classically explained as an effect of confirmation bias rather than a true association with reinforcing outcomes (Skinner, 1948). In contrast, the results from Damisch et al. (2010) suggest that superstitions about one's own behavior can be efficacious. If true, this class of superstition is not completely irrational and the prevalence of such behaviors could be explained by their strong positive consequences. Both psychologists and the general public have been quick to recognize the importance of this finding. The original report has been cited 55 times (*Google Scholar*, 2013), was covered extensively in the popular press at the time of publication (e.g., Doheny, 2010; Hutson, 2010), and has even become part of the sales pitch for an online website selling lucky charms (http://www.good-luck-gifts.com, n.d.).

In support of their findings, Damisch et al. (2010) reported three successful conceptual replications. In addition, a dissertation by Damisch (2008) reports an additional two successful conceptual replications. This work is summarized in Table 1 and Figure 1. Integration across results indicates an overall effect size that is at least moderate and possibly very large (unbiased d = 0.82, 95% CI [0.53, 1.11], white diamond in Figure 1).

While these results suggest a robust and powerful effect of superstition on performance, conceptual replications by others show mixed results. Lee, Linkenauger, Bakdash, Joy-Gaba, and Proffitt (2011) found that golfers told they were using a famous golfer's putter performed substantially better on a putting task than controls. In contrast, Aruguete, Goodboy, Jenkins, Mansson, and McCutcheon (2012) found that superstitions related to prayer are not effective at improving performance on a reasoning test. In an additional experiment, priming participants to think about their religious beliefs also failed to improve scores on a reasoning test relative to controls.

The contrasting results of these conceptual replications could be due to a number of factors. It could be that only some types of superstitions are efficacious, perhaps those related to luck rather than religion. Another possibility is that superstition can affect performance on only some types of tasks. Given the uncertainty, it seemed important as a first step to directly confirm the replicability of the original finding. Here we report two high-powered, precise

DOI: 10.1027/1864-9335/a000190

Table 1. Summary of studies examining the effect of superstition on performance

Study	Superstition Prime	Dependent Variable	*n*/group	Power if *d* = 0.83
Damisch et al. (2010)				
Experiment 1	Experimenter labeled ball lucky	Golf task	14	0.55
Experiment 2*	Experimenter wished good luck	Motor-dexterity task	17	0.64
Experiment 3	Presence of self-selected lucky charm	Memory game	20	0.71
Experiment 4	Presence of self-selected lucky charm	Anagram task	15	0.58
Damisch (2008)				
Study 2	Experimenter made shot and labeled ball lucky	Golf task	14	0.55
Study 5	Subliminal priming for word ‘‘luck’’	Tetris	14	0.55
Lee et al. (2011)				
Experiment 1	Told using a professional golfer’s putter	Golf task	20	0.71
Aruguete et al. (2012)				
Experiment 1	Religious beliefs scale	Verbal reasoning test	71	0.99
Experiment 2*	Wrote and prayed for success	Verbal Reasoning test	53	0.99
This manuscript				
Experiment 1	Experimenter labeled ball lucky	Golf task	58	0.99
Experiment 2	Drew ball with clover and experimenter labeled lucky	Golf task	54	0.99

Notes. For *n*/group, smaller group size reported if group sizes were uneven. Power calculated for an effect size of 0.83, the unbiased overall effect size calculated over the six studies by Damisch (2008) and Damisch et al. (2010). *Additional groups run beyond control and superstition prime, but only 1 contrast selected for meta-analysis.

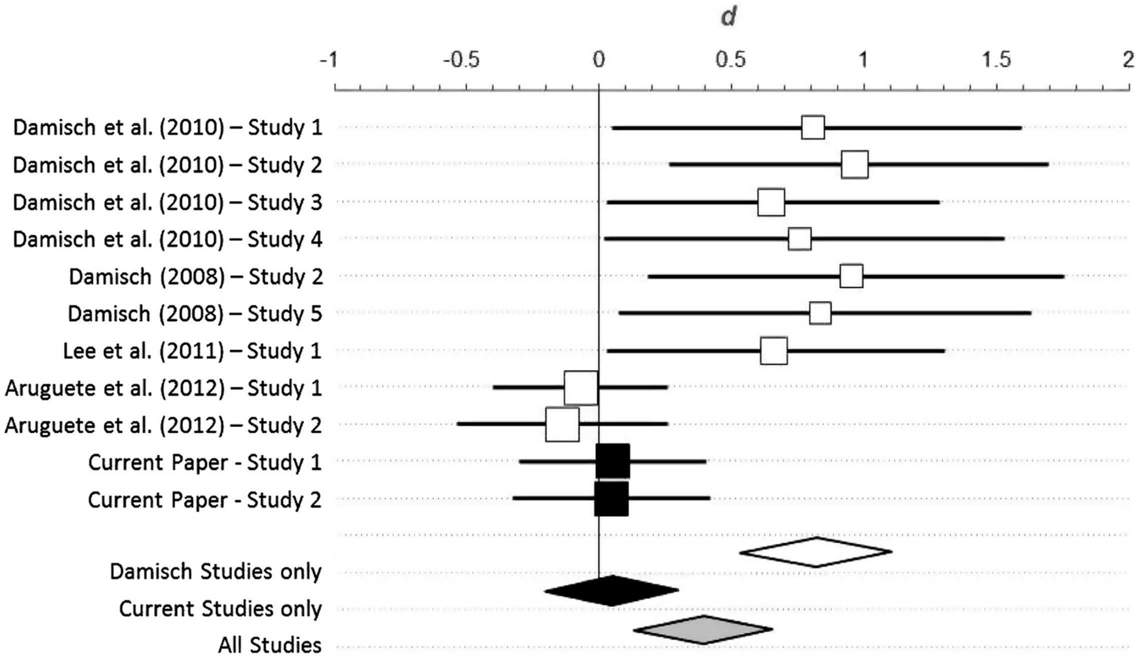

Figure 1. Meta analysis of the effects of superstition on performance. The location of each square represents the observed effect size of a single study. The 95% CI of the effect size is represented by the line extending from the square, and relative sample size is represented by the area of the square. Studies conducted prior to this one are shown with white squares; the two studies reported in this manuscript are shown in black. The diamonds represent unbiased effect sizes estimates over groups of studies, with the center of the diamond marking the point estimate for effect size and the width of the diamond covering the 95% CI. The first overall estimate (white diamond) is for the six studies conducted by Damisch (2008) and Damisch et al. (2010). The black diamond represents the overall effect size estimate from the two studies reported in this manuscript. The gray diamond is for all studies, but note that significant heterogeneity of effect sizes was evident ($Q(10) = 26.5$, $p = .003$). This figure was created using ESCI (Cumming, 2011).

replications of the golf and superstition experiment (Study 1) from Damisch et al. (2010). We focused on this study because it is simple (just two groups), involves no idiomatic language, applies to the general population, involves equipment which can be precisely matched to the original experiment, and has a large effect size.

Experiment 1: Direct Replication

We matched both the materials and procedures of the target study as precisely as possible. This was facilitated by the gracious cooperation of Lysann Damisch (personal communications, 2012–2013) who provided detailed feedback based on a review of our materials and a video of a training session.

The replication was registered in advance of data collection on the Open Science Framework. All materials, data, video of the procedure, and the preregistered design are available at http://osf.io/fsadm/. We report all data exclusions, manipulations, and measures, and how we determined our sample sizes.

Method

Participants

The target study's data came from a convenience sample of German undergraduates, with 57% females and 43% males. Left-handers and psychology majors were excluded (Damisch, personal communication). No compensation was provided. Eighty percent of the participant pool believed in superstition. This was based on responses to a single item "How much do you believe in good luck and bad luck" rated on a scale from 1 (= *not at all*) to 9 (= *very much*), with responses greater than 3 counted as belief in superstition (Damisch, personal communication).

We collected a convenience sample from biology laboratory classes at a private comprehensive university in the United States. The biology classes we targeted were open to non-majors and most fulfilled general education requirements, leading to enrollment from a wide range of majors. Moreover, science majors within our university exhibit similar levels of superstition compared to the participants in the target study (79%, or 27 out of 34 responses to the same item delivered as an online questionnaire to Chemistry, Natural Science, and Mathematics majors at the same university).

We did not exclude left-handers because we did not know about this criterion until after our sampling plan was developed. However, we fortuitously targeted classes with relatively few psychology majors, and tracked major so that these participants could be excluded post hoc. Participants were compensated with an entry into a drawing to receive a $25 Amazon gift card, with odds of winning set at 1:20.

We planned to sample at least 42 but no more than 91 participants per group. The minimum was set to provide 0.95 power for the overall average effect size (0.82) across the six prior superstition and performance studies (Damisch et al., 2008, 2010), the maximum to provide similar power for the lower bound of the 95% CI for the effect (0.53). We collected data until our minimum target had been exceeded and our participant pool was depleted, yielding data from 58 controls and 66 superstition-activated participants (power at 0.83 even for the lower-bound effect size estimate). Our sample consisted of 90 females (73%), 28 males (23%) and 6 participants who did not report their gender (5%). No participants were excluded from initial analysis. Although this overrepresents females relative to the target study, the effects of superstition on performance have been demonstrated with all-female samples (Damisch et al., 2010, Study 2).

Materials

Three female research assistants collected all the data for this study. To ensure smooth and even-handed administration of the experiment, each assistant memorized an experimental script and completed at least five practice sessions prior to collecting data. None of the research assistants had read the target article, but were informed that the manipulation could enhance, impair, or have no effect on performance.

Two approximately identical research rooms were used for data collection. Each contained a personal computer with the monitor surrounded by a study carrel for privacy. Each also had an open space for the putting task. The floor was covered with office-grade brown wall-to-wall carpeting.

Participants completed a computerized questionnaire at the beginning to record consent, gender, major, and school year. In addition, text instructions explained that they would complete a golf task because adapting to a new task is a good predictor of future success. This cover story was provided by Damisch (personal communication).

We acquired the same executive putting set (Shivam Putter Set in Black Travel Zipper Pouch, see source list) used in the original research. The set consists of a metal putter, two standard white golf balls, and a square wooden target with an omega-shaped cutout. We replaced the putter, however, with a similar model made for both left- and right-handed putters (Quolf Two-Way Putter), to accommodate left-handed participants.

Damisch et al. (2010) used a putting distance of 100 cm. In a pilot test, we found that students in our undergraduate population are too good at this task, due either to more golf experience or to a slower "green speed" for the carpeting in our research rooms. Controls (n = 8) averaged 8.25/10, considerably higher than the 4.8/10 reported for controls in the original study. We therefore moved the target back to 150 cm to equate difficulty. In a second round of pilot testing at this distance, controls (n = 19) averaged 5.9/10, much closer to the original study. We used this longer putting distance to achieve similar task difficulty.

The target itself was placed 100 cm from the wall (Damisch, personal communication), and the starting point for the ball was marked with tape.

To ensure and measure the quality of the replication, we added a quality-control task and a manipulation check via a computerized post golfing task questionnaire. Participants were asked "What did the researcher say to you as she handed you the golf ball?" Participants in the lucky condition passed if they mentioned the word *luck* (or any of its variants); participants in the control condition passed if they failed to mention the word *luck* (or any of its variants).

Then, participants completed a two-item manipulation check: "Before starting this task, I believed that the golf ball assigned to me was lucky" and "Now that I have completed this task, I believe that the golf ball assigned to me is lucky." Responses were made on a Likert scale from 1 (= *strongly disagree*) to 5 (= *strongly agree*). Note that these manipulation checks were retrospective. Responses could thus be contaminated by their experience on the golf task. This order was used, however, to avoid altering the original protocol. Furthermore, pilot testing suggested that these measures would still elicit the expected group difference in feelings of luck, $t(41) = 2.23$, $p = .031$, $d = 0.66$.

Score sheets were created in advance for each participant with random assignment to the control or superstition-primed group via a random number generator. Score sheets were then printed, and placed in the research rooms for the research assistants to use in sequential order. Condition was indicated on the score sheet as a "C" or "L" to avoid priming participants should they glance at the score sheet.

Procedure

Participants were recruited during down-time in their laboratory class sessions. Volunteers were escorted to the research room one at a time by a research assistant. Upon arrival, the researcher asked the participant to complete the initial portion of the computerized questionnaire, including informed consent, demographics, cover story, and task explanation. The researcher then explained the task again and handed the participant the golf ball, saying either "Here is the ball. So far it has turned out to be a lucky ball" (superstition-activated group) or "This is the ball everyone has used so far" (control group).

Participants then completed the golf task (10 putts). After each shot, the researcher stated "Ok, now for shot X" where X was the next shot. No other feedback was given. After the golf task, the participants completed the quality-control task and manipulation check. The research assistant stood on the other side of the room during this task.

Differences From the Original Study

We conducted a faithful replication of the golf and superstition study by Damisch et al. (2010). The only differences are that we:

- recruited US college students rather than German college students.
- administered the experimental script in English rather than German (but using the translation provided by Damisch et al. (2010) for the key manipulation).
- recruited a somewhat higher proportion of women.
- collected data with three female research assistants rather than one.
- used a putting distance of 150 cm rather than 100 cm to equalize task difficulty for our population.
- included left-handed golfers as well as right-handed golfers.
- added a quality-control task and manipulation check to the end of the protocol.

Most of these differences are not substantive, with the possible exception of cultural differences between undergraduates from Germany and the US. However, similar results have been reported with students living in the US. (Lee et al., 2011), and our participants were well-matched in terms of their belief in good and bad luck.

Analysis

As in the original report, differences in performance between the superstition and control groups were analyzed with an independent samples *t* test. Effect sizes are estimated using Cohen's *d*. We also report confidence intervals for group differences. Estimates of power were calculated with PS Power and Sample Size Software for Windows (Dupont & Plummer, 1998)

Results and Discussion

We did not observe a strong impact of superstition on golf performance (Table 2). The superstition-activated group performed just 2% better than the control group (compared to 35% improvement in the target study). This difference did not reach statistical significance, $t(122) = 0.29$, $p = .77$.

Participants in the superstition-activated group retrospectively reported themselves to have felt luckier at the start of the golf task compared to those in the control group, $t(115.4) = 4.28$, $p = .00004$. This feeling of luck was also evident after the golf task was complete, $t(112) = 2.02$, $p = .045$. Despite the successful manipulation checks, we did find that many participants in the superstition-activated group failed the quality-control task we designed. Specifically, when asked, "What did the experimenter say when she handed you the golf ball?" only 42 of 66 (63%) participants mentioned "luck." Excluding the participants who failed this task still preserved strong power for the analysis (0.98), but the group difference remained very small (3.6% improvement) and did not reach statistical significance, $t(98) = 0.40$, $p = .69$.

Debriefing provided some clues as to why so many participants in the superstition-activated group failed the quality-control task. Some participants reported that they believed the mention of luck by the experimenter was "off script" and had not wanted to mention it for fear of getting the experimenter in trouble. Thus, some participants

Table 2. Effects of superstition on golf performance

Measure	Control *M(SD)*	Superstition *M(SD)*	Mean difference *M* [95% CI]	Effect size *d*
Golf putts made	4.62 (2.13)	4.73 (1.96)	0.11 [− 0.62, 0.83]	0.05
Manipulation checks:				
Felt lucky prior to task	1.64 (0.93)	2.55 (1.36)	0.91 [0.50, 1.32]	0.47
Felt lucky after task	2.02 (1.26)	2.48 (1.30)	0.47 [0.10, 0.93]	0.36
Quality control:				
Recalled prompt	58/58 (100%)	42/66 (63%)		
Golf putts made	4.62 (2.13)	4.79 (1.97)	0.17 [− 0.66, 0.99]	0.08

Notes. Control group $n = 58$, superstition-activated group $n = 66$. Note that manipulation checks were retrospective ratings made *after* the task. Under quality control, putts made are only for those participants who passed the quality control task.

may have failed this task not due to poor impact but due to highly credulous responses to the manipulation. Exploratory analysis provided some evidence consistent with this interpretation; those in the superstition-activated group who failed the quality-control task actually reported slightly *higher* feelings of luck than those who passed (e.g., $M = 2.88$, $SD = 1.29$ for the 14 participants who failed the task; $M = 2.36$, $SD = 1.38$ for the 42 participants who passed the task, though this difference is not statistically significant, $t(64) = 1.50$, $p = .14$).

We conducted exploratory analyses to try to uncover an effect of superstition on performance. We excluded psychology majors, checked for an interaction by gender, and checked for an interaction by research assistant. No significant effects were observed (see Supplementary Table S1).

Experiment 2: Higher Impact Replication

Although our replication attempt succeeded in having high power and demonstrable impact, we wondered if a stronger superstition prime would produce the expected effect.

Method

Methods were the same as above but with the following modifications.

Participants

We altered our sampling plan to target the general university population to ensure that the results were not idiosyncratic to students enrolled in biology courses. The sample was recruited by advertising on campus bulletin boards and on campus. For this study, participants signed up for appointments and arrived at the research room on their own. Participants were compensated with an experimental participation voucher that could be redeemed in some classes toward course credits.

We collected data for 113 participants, halting data collection when our minimum target was exceeded and only one week remained before the deadline for manuscript submission. One participant in the lucky condition requested at the end of the experiment that his or her data be withdrawn from analysis. Another failed the quality-control task and was removed. Thus, our final sample consisted of 111 participants (28 males and 83 females); these were randomly assigned to either the control ($n = 54$) or superstition-activated ($n = 66$) conditions.

Materials

To enhance impact, participants selected their ball from a velour sack containing eight golf balls: Four regular and four emblazoned with a green clover (Shamrock Golf Ball, see source list). The experimenter's prompt was also enhanced: "This is the ball you will use" for the control group versus "Wow! You get to use the lucky ball" for those in the superstition-activated group.

Table 3. Effects of enhanced superstition activation on golf performance

Measure	Control *M(SD)*	Superstition *M(SD)*	Mean difference *M* [95% CI]	Effect size *d*
Golf putts made	4.02 (2.20)	4.12 (2.01)	0.10 [− 0.69, 0.90]	0.05
Manipulation checks:				
Felt lucky prior to task	2.02 (0.93)	2.86 (1.4)	0.85 [0.38, 1.30]	0.68
Felt lucky after task	2.07 (0.95)	2.46 (1.28)	3.47 [− 0.04, 0.81]	0.34
Quality control:				
Recognized ball	54/54 (100%)	66 of 67 (99%)		

Notes. Control group $n = 54$, superstition-activated group $n = 66$ (one participant who failed quality-control task excluded from data analysis). Note that manipulation checks were retrospective ratings made *after* the task.

To explore possible moderators, we added a measure of belief in luck. This was the same measure described earlier that Damisch et al. had used to measure belief in luck in their participant population (Damisch, personal communication). This item, like the others, was administered via a computerized questionnaire after the golf task was completed.

The quality-control task was modified to a recognition task: Participants were shown an image of both the regular ball and the "lucky" ball and asked to choose which they had received. To avoid contaminating other responses by showing both conditions the "lucky" ball, this task was moved to the end of the questionnaire.

Results and Discussion

We did not observe an impact of superstition on performance (see Table 3). Participants in the superstition-activated group scored just 2.5% higher than those in the control group, a nonsignificant difference, $t(109) = 0.26$, $p = .80$.

Our failure to replicate was not due to insufficient impact, as this study produced an even larger difference in participants' retrospectively reported feelings of luck before the golf task, $t(96) = 3.65$, $p = .004$. The difference in ratings of luck after the golf task was not statistically significant, $t(109) = 1.78$, $p = .08$.

Could these results be due to insufficient superstition in our participants? This seems unlikely. Seventy percent of control and 80% of superstition-activated participants reported a belief in luck, similar to the target study's participant pool. Moreover, excluding participants in both groups who did not believe in luck (using same criterion as Damisch, 2008) did not reveal an effect (see Supplementary Table S2, $t(82) = -0.69$, $p = .49$).

In exploratory analyses, we did not observe interactions by gender or experimenter, and excluding psychology majors did not have an effect (see Supplementary Table S2).

Aggregating the data across the two studies indicates a null effect of superstition on performance: Unbiased overall $d = 0.05$, 95% CI[-0.21, 0.30], as indicated by the black diamond in Figure 1. The confidence interval of this estimate does not overlap with that generated across the studies by Damisch et al. (2010) and Damisch (2008) (white diamond, Figure 1).

Meta-Analysis

To better understand the divergence between our results and those of the target study, we conducted a small-scale meta-analysis. We included the original golf experiment and conceptual replications described in the introduction (Aruguete et al., 2012; Damisch, 2008; Damisch et al., 2010; Lee et al., 2011) and the two attempts reported here (summarized in Table 1 and Figure 1). We conducted the meta-analysis using ESCI (Cumming, 2011), an Excel-based analysis package which includes tools for integrating effects sizes and visualizing their differences across studies.

The meta-analysis provides an overall unbiased estimate of effect size: $d = 0.40$, 95% CI [0.14, 0.65], gray diamond in Figure 1. However, there is significant heterogeneity in the reported effect sizes ($Q(10) = 26.52$, $p = .003$): One subset of studies indicates a strong effect, the remainder indicates little to no effect. This heterogeneity requires caution in interpreting the overall estimated effect size (see Discussion).

General Discussion

Although we took care to precisely replicate the materials and procedures of the target study, we could not replicate the strong effect of superstition on performance consistently observed by Damisch et al. (2010) and Damisch (2008).

What could account for our failed replications? We can rule out a lack of impact: We observed robust effects in a manipulation check, conducted a second replication that achieved an even higher impact, and implemented quality controls that allowed filtering out of any participants not sufficiently engaged in the task. It is possible that the target study achieved even greater impact but no manipulation check was conducted to provide comparison. This seems implausible, however, as Damisch et al. (2010) was able to observe strong effects on performance with subtle manipulations.

Our meta-analysis suggests considerable heterogeneity in observed effects of superstition on performance. Such heterogeneity can indicate the operation of a moderator, perhaps one that differs between the European participants in the target study and the American participants in these replications. Indeed, culture can play a surprisingly large role even in basic psychological phenomena (Henrich, Heine, & Norenzayan, 2010). This seems unlikely, however, as we took care to equate with the original study or monitor key moderators including belief in luck, task difficulty, and sample characteristics. It is notable, though, that in the Damisch et al. studies (2008) performance gains in the superstition group were associated with increased self-efficacy (Study 3 and 4) and task persistence (Study 4). This suggests, then, that strong effects of superstition only emerge when control participants are not confident or motivated enough to perform near their ability, providing "room" for superstition to boost performance through these factors. Indeed, Matute (1994) and others have suggested that superstitions function specifically to maintain performance under adverse conditions.

Heterogeneity of effect sizes can also arise due to substantive differences in research quality. We made every effort to replicate the target study precisely. Further, we developed an *a priori* sampling plan, took steps to minimize expectation effects (e.g., experimental script), and acquired a large enough sample size to provide a relatively precise estimate of effect size. These are all design features recently emphasized for increasing research rigor, especially for ensuring good control of Type I error (e.g., Simmons, Nelson, & Simonsohn, 2011). Along these lines, it is notable that the four studies with these features (our own plus the two from Aruguete et al., 2012) consistently indicate no effect of superstition on performance. The studies that do show an effect of superstition on performance lack some or all of these design features. Moreover, the Damisch studies show a remarkable consistency of result that could occur if Type I error is not well controlled: Given the overall effect size from these studies (0.83, white diamond, Figure 1), the odds of all six of these studies reaching statistical significance is only four in 100.

Ultimately, only further research can determine if the lack of effect we observed is due to moderators, improved rigor, or both. Currently, the studies with the strongest design features do not indicate a robust effect of superstition on performance.

Acknowledgments

We thank Jamie Mussen, Malgorzata Rozko, and Amy Taraszka for collecting the data for this study. We also thank Lysann Damisch for her gracious and extensive cooperation. Funding for materials and participation remuneration was provided by a grant from the Center for Open Science. The authors declare no conflict-of-interest with the content of this article. Designed Research: R. C-J., T. C.; Analyzed Data: R. C-J.; Wrote Paper: R. C-J., T. C. All materials, data, video of the procedure, and the preregistered design are available at http://osf.io/fsadm/.

PREREGISTERED OPEN DATA OPEN MATERIALS

References

Aruguete, M. S., Goodboy, A. K., Jenkins, W. J., Mansson, D. H., & McCutcheon, L. E. (2012). Does religious faith improve test performance? *North American Journal of Psychology, 14*, 185–196.

Cumming, G. (2011). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. New York, NY: Routledge.

Damisch, L. (2008). *Keep your fingers crosssed! The influence of superstition on subsequent task performance and its mediating mechanism (Doctoral dissertation, University of Cologne, Cologne, Germany)*. University of Cologne. Retrieved from http://kups.ub.uni-koeln.de/2471/

Damisch, L., Stoberock, B., & Mussweiler, T. (2010). Keep your fingers crossed! How superstition improves performance. *Psychological Science, 21*, 1014–1020. doi: 10.1177/0956797610372631

Doheny, L. (2010, July 16). Good Luck Charms Might Just Work. *U.S. News Health*. Retrieved from http://health.usnews.com/health-news/family-health/brain-and-behavior/articles/2010/07/16/good-luck-charms-might-just-work

Dupont, W. D., & Plummer, W. D. (1998). Power and sample size calculations for studies involving linear regression. *Controlled Clinical Trials, 19*, 589–601.

Galton, F. J. (1872). Statistical inquiries into the efficacy of prayer. *The Forthnightly Review, 12*, 125–135. doi: 10.1093/ije/dys109

Good-Luck-Gifts.com. (n.d.). *Lucky charms really work.* Retrieved from http://www.good-luck-gifts.com/index.php/lucky-charms-guide/118-lucky-charms-really-work

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *The Behavioral and Brain Sciences, 33*, 61–63. doi: 10.1017/S0140525X0999152X

Hutson, M. (2010, February). *Harder, Better, Faster, Luckier [Web log post]. Psychology Today, Psyched! Blog.* Retrieved from http://www.psychologytoday.com/blog/psyched/201002/harder-better-faster-luckier

Lee, C., Linkenauger, S. A., Bakdash, J. Z., Joy-Gaba, J. A., & Profitt, D. R. (2011). Putting like a pro: The role of positive contagion in golf performance and perception. *PloS One, 6*, e26016. doi: 10.1371/journal.pone.0026016

Matute, H. (1994). Learned helplessness and superstitious behavior as opposite effects of uncontrollable reinforcement in humans. *Learning and Motivation, 25*, 216–232. Retrieved from http://www.sciencedirect.com/science/article/pii/S0023969084710125

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22*, 1359–1366. doi: 10.1177/0956797611417632

Skinner, B. (1948). "Superstition" in the pigeon. *Journal of Experimental Psychology, 38*, 168–172. Retrieved from http://psycnet.apa.org/journals/xge/38/2/168/

Received February 18, 2013
Accepted December 3, 2013
Published online May 19, 2014

Robert J. Calin-Jageman

Department of Psychology
Dominican University
7900W. Division
River Forest, IL 60305
USA
E-mail rcalinjageman@dom.edu

Does Recalling Moral Behavior Change the Perception of Brightness?

A Replication and Meta-Analysis of Banerjee, Chatterjee, and Sinha (2012)

Mark J. Brandt, Hans IJzerman, and Irene Blanken

Tilburg University, The Netherlands

Abstract. Banerjee, Chatterjee, and Sinha (2012) recently reported that recalling unethical behavior led participants to see the room as darker and to desire more light-emitting products (e.g., a flashlight) compared to recalling ethical behavior. We replicated the methods of these two original studies with four high-powered replication studies (two online and two in the laboratory). Our results did not differ significantly from zero, 9 out of 10 of the effects were significantly smaller than the originally reported effects, and the effects were not consistently moderated by individual difference measures of potential discrepancies between the original and the replication samples. A meta-analysis that includes both the original and replication effects of moral recall on perceptions of brightness find a small, marginally significant effect ($d = 0.14$ CL_{95} -0.002 to 0.28). A meta-analysis that includes both the original and replication effects of moral recall on preferences for light-emitting products finds a small effect that did not differ from zero ($d = 0.13$ CL_{95} -0.04 to 0.29).

Keywords: morality, light, grounded cognition, embodiment

One recent addition to the literature on grounded cognition, specifically on conceptual metaphors of morality, examined how reminders of people's own morality or immorality can shape perceptions of lightness and darkness (Banerjee, Chatterjee, & Sinha, 2012; from here on referred to as "BCS"). BCS requested participants to recall a time they engaged in ethical or unethical behavior and then asked how light or dark the room was. They found that participants who recalled three unethical deeds and subsequently wrote about the most unethical deed perceived the room as darker (Study 1, $N = 40$) and being lit with fewer Watts (Study 2, $N = 74$) than participants who recalled three ethical deeds and subsequently wrote about the most ethical deed. These effects provide support for the idea that the abstract target domain of morality influences perceptions in the concrete source domain of light (see Figures 1 and 2 for effect sizes and confidence intervals; see Firestone & Scholl, 2014, for a recent alternative explanation for this effect). An intriguing addition to BCS Study 2 was the finding that people in the unethical condition compared to the ethical condition were more likely to prefer products that convey light (e.g., lamps, flashlights), presumably because they perceive their environment to be darker. That is, these studies have found that abstract thought (morality) shapes concrete experiences (perception of light) (the abstract → concrete causal direction). We aimed to replicate these studies as closely as possible.

The work by BCS builds on other studies that have linked immorality/morality with darkness/lightness (Frank & Gilovich, 1988; Sherman & Clore, 2009; Webster, Urland, & Correll, 2012; Zhong, Bohns, & Gino, 2010). These studies all suggest that the concrete source domain of color and light perception influences the abstract target domain of morality (the concrete → abstract causal direction). These results are consistent with the linguists Lakoff and Johnson's (1999) suggesting that conceptual metaphors are unidirectional. That is, the learning of an abstract concept may co-occur with the concrete experience, but the concrete experience is not necessarily associated with the abstract concept. Concretely, this means that lightness/darkness should lead to an alteration in perceptions of morality, but priming of moral or immoral deeds should not lead to perceptions of lightness/darkness (for skepticism of this argument see IJzerman & Koole, 2011; IJzerman & Semin, 2010; for skepticism of this skepticism, see Lee & Schwartz, 2012; Slepian & Ambady, 2014). The studies by BCS are important for understanding the theoretical link of light and morality because the studies by BCS suggest that the morality-light association goes beyond such conceptual metaphors (cf. Lakoff & Johnson, 1999).

DOI: 10.1027/1864-9335/a000191

They indicate that the abstract concept of morality guides our processing of color/light information and influences our perception of light in our environment, thereby potentially suggesting Conceptual Metaphor Theory is incomplete, or that moral concepts are grounded in basic perceptual simulators (Barsalou, 1999; IJzerman & Koole, 2010; see also Hamlin, Wynn, & Bloom, 2007, for reasoning that may suggest this argument).

Methods

We replicated BCS Studies 1 and 2 using the original methods from BCS that were provided to us by the original authors. The details of our methods, the precise differences between our replications and the original studies, and our sample size justifications and planning can be found in online supplemental material and in the original preregistration of the studies. Because the methods of the two original studies are largely the same, with the exception of the dependent variables, we simultaneously describe our four replication studies and note where they deviate from one another.

Procedure and Measures

All of the participants completed the studies on computers. Participants in our online samples completed the study on their own computer. The laboratory samples completed the study on computers in individual cubicles (see a photo of a cubicle and a video simulation from one week of our replication of Study 2 in the supplemental materials). All of the measures from the original study were included in the replications. In the replication of Study 1, participants described in detail an ethical or an unethical deed from their past, completed filler items about the room they were in, and made judgments of the brightness in the room on a 7-point scale (1 = *not bright at all*, 7 = *very bright*). In the replication of Study 2, the procedure was the same, except following the filler items participants rated their preferences (1 = *not at all desirable*, 7 = *very desirable*) for light-emitting (lamp, candle, flashlight) and filler (jug, crackers, apple) products before estimating the brightness of the room in watts. The brightness judgments (Studies 1 & 2) and preference for light-emitting products (Study 2) were the primary dependent variables. We also included several additional measures of demographic information, religiosity, political ideology, and moral-self identification at the very end of the study to test possible moderators that may explain differences between the original and replication samples, and to maximize chances to obtain an effect at all.

Participants

For each of the original two studies reported by BCS we conducted two replication attempts, one online via MTurk where participants received $0.50 and one in our laboratory at Tilburg University in the Netherlands where participants received course credit or 5 Euros (see below). The final sample sizes and basic demographic information from both the original study and our replication studies are in Table 1. The sample sizes reported in Table 1 are the largest sample sizes available for the study; however, given that some participants did not complete all of the measures the precise degrees of freedom vary depending on the analysis.

In the online studies we aimed for *N*s of 496 and 510 for Studies 1 and 2, respectively. In the laboratory studies we aimed for *N*s of 126 and 130 for Studies 1 and 2, respectively. We aimed for these sample sizes because they would give us 95% power given the effect sizes reported by BCS (and assuming that the effect sizes in the online studies would be 50% weaker than in the original studies that were conducted in the laboratory). Although we followed the data collection protocol and stoppage rules outlined in our preregistration for laboratory Study 1 and online Studies 1 and 2, we fell short of our sample size goals because there were more participants than expected in our online samples who did not follow directions or completed the study outdoors. In our laboratory study, we did not collect the expected sample size because we were unfortunate to collect data during a "slow" laboratory week. For laboratory Study 2, we collected data for 1 week (as specified in our preregistration) and participants were compensated with partial course credit; however, we did not have nearly a sufficient number of participants ($N = 66$). Therefore we collected data for 2 additional weeks and participants were compensated with €5 ($N = 55$). Analyses that take the "week of data collection" into account do not alter the conclusions we report below. Each of our studies still had a high amount of power to detect effects of the size reported by BCS. The achieved power was above typically recommended power levels (e.g., .80 by Cohen, 1988) (Lowest Achieved Power Online Study 1 = .94, Online Study 2 = .93, Laboratory Study 1 = .90, Laboratory Study 2 = .90).

Results

Our confirmatory analyses replicated the analyses reported in BCS (i.e., independent-sample *t*-tests comparing experimental conditions) and can be found in Table 2. In the online studies, the effects of the experimental conditions on all of the primary dependent variables were nonsignificant (all t's < $|1.28|$, all p's > .20). Similarly, in the laboratory studies the effects of the experimental conditions on all of the primary dependent variables were nonsignificant (all t's < $|0.59|$, all p's > .55). With the exception of the estimation of brightness in the online version of Study 1, all of the effect sizes were significantly smaller in the replication studies than the original study (see Table 2).

Exploratory Analyses

We tested to see if age, gender, ethnicity (online studies), education (online studies), income (online studies), importance

Table 1. Final sample sizes and demographic information in the original and replication studies

	BCS		B&I: Online		B&I: Laboratory	
	Study 1	Study 2	Study 1	Study 2	Study 1	Study 2
N	40	74	475	482	100	121
M age (*SD*)	NR	NR	28.8 (9.5)	29.2 (9.5)	19.6 (2.2)	20.0 (2.3)
Gender (M/F)	NR	NR	173/301[a]	169/313[a]	27/73	44/77
Population	"participants at a large public university"	NR. Assumed to be the same as Study 1	MTurk workers	MTurk workers	Dutch university students	Dutch university students
Population location	United States	United States	United States	United States	Netherlands	Netherlands
Study setting	Computer in laboratory	Computer in laboratory	Online	Online	Computer in individual laboratory cubicle	Computer in individual laboratory cubicle

Notes. NR = not reported. [a]1 person did not report their gender.

Table 2. Means, standard deviations, and effect sizes for the original and replication studies

	BCS			Replication: Online			Replication: Laboratory		
	Unethical prime		Ethical prime	Unethical prime		Ethical prime	Unethical prime		Ethical prime
Study 1									
Perceived brightness *M*	4.71		5.3	4.34		4.51	4.79		4.66
Perceived brightness *SD*	0.85		0.97	1.52		1.47	1.09		1.19
Perceived brightness *d*		$.65_{a}$			$.12_{ab}$			$-.11^{b}$	
Achieved power		.52			>.99			.90	
Study 2									
Estimated watts *M*	74.3		87.6	296.12		443.88	130.44		135.91
Estimated watts *SD*	26.85		7.40	1,095.43		3,991.04	152.03		192.64
Estimated watts *d*		$.64_{a}$			$.05_{b}$			$.03_{b}$	
Achieved power		.78			>.99			0.9	
Lamp preference *M*	4.16		2.34	4.02		4.07	3.44		3.62
Lamp preference *SD*	1.70		1.15	1.59		1.64	1.74		1.47
Lamp preference *d*		1.23_{a}			$-.03_{b}$			$-.11_{b}$	
Achieved Power		>.99			>.99			>.99	
Candle preference *M*	3.62		2.37	3.34		3.33	4.30		4.32
Candle preference *SD*	1.83		1.16	1.76		1.80	1.67		1.51
Candle preference *d*		$.79_{a}$			$.003_{b}$			$-.01_{b}$	
Achieved power		.92			>.99			.99	
Flashlight preference *M*	4.33		2.35	3.26		3.43	2.67		2.80
Flashlight preference *SD*	1.71		1.15	1.75		1.78	1.58		1.39
Flashlight preference *d*		1.33_{a}			$-.10_{b}$			$-.09_{b}$	
Achieved power		>.99			>.99			>.99	

Notes. Achieved power for the replications is the achieved power based on the effect sizes reported in BCS (without the adjustments made for online vs. laboratory studies, see main text and preregistration). Effect sizes within the same row with different subscripts are significantly different from one another $p < .05$. BCS effect sizes are the effect sizes reported in the paper. Effect sizes from our studies were calculated with Becker's effect size calculator: http://www.uccs.edu/~lbecker/. Differences between the effect sizes were computed by first computing the equivalent *r*-value for each *d*-value and then computing a *z*-score for differences in correlation coefficients (http://vassarstats.net/rdiff.html).

of morality to the self, religiosity growing up, current religiosity, and political ideology moderated the effect of the experimental manipulations on any of the primary dependent measures. There were 3 of 80 moderation effects were possible. None of the significant differences were observed consistently across studies.

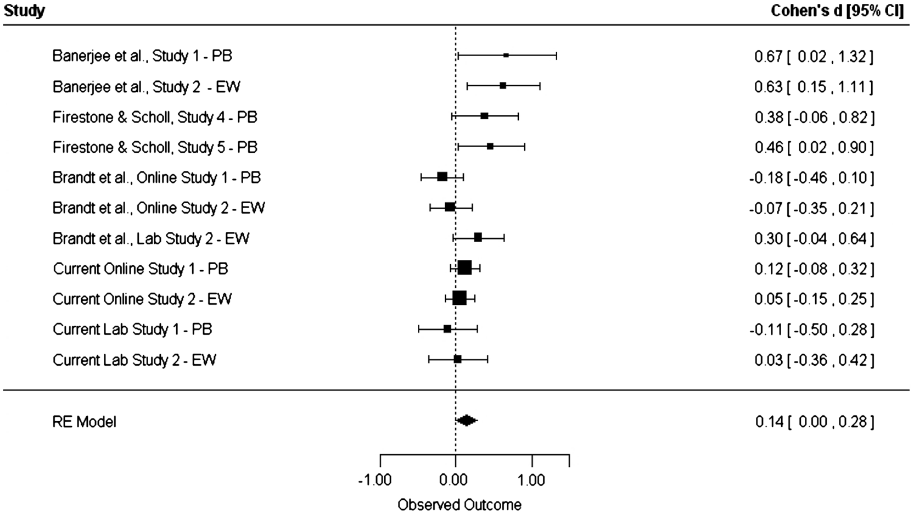

Figure 1. Cohen's *d*, 95% confidence intervals, and estimate of overall effect size from a random effects meta-analytic model for the perceptions of brightness. PB = perceived brightness, EW = estimated watts.

Meta-Analyses

The results of any one study, including high-powered replication studies, could be the result of chance. Similarly, the original studies may have uncovered robust effects, but by chance estimated the effect sizes as much larger than the true effects. Therefore, to gain a more precise understanding of the effects we conducted two meta-analyses (one on the brightness judgments and one on the desirability of light-emitting products) including the original studies, the replication attempts reported here, our own previous replication attempts (Brandt, IJzerman, & Blanken, 2013), and two other recent published replication attempts of Study 1 of BCS (Firestone & Scholl, 2014). With the information we collected, we were also able to test whether the effect was more robust online or in the laboratory and whether it was more likely in the United States or in the Netherlands. Although not specified in our preregistration, we also tested whether the research laboratory where the study was conducted affected the obtained effect sizes. All analyses were conducted using the metaphor package for the R program (Viechtbauer, 2010).

Meta-Analysis on the Effects of Experimental Condition on Brightness Judgments

We first conducted a meta-analysis to derive the overall mean effect size of experimental condition on brightness judgments (N = 11). The random effects meta-analysis produced a mean effect size of d = 0.14 ([CL$_{95}$] −0.002 to 0.28). There was a marginal effect of experimental condition across all the studies on brightness judgments (z = 1.93, p = .054). Figure 1 provides a forest plot of the effect sizes of the brightness judgments across studies. The effect of experimental condition on brightness judgments did not differ for participants from the US (M effect size = 0.17, SE = 0.09) versus participants from the Netherlands (M effect size = 0.08, SE = 0.16), QM(2) = 3.63, p = .16. The effect was larger for studies conducted in the laboratory (M effect size = 0.24, SE = 0.12, p = .05) than for studies conducted online (M effect size = 0.08, SE = 0.12, p = .36), QM(2) = 4.85, p = .04.

Our exploratory analysis on research laboratory where the study was conducted yielded an overall significant effect, QM(4) = 18.45, p = .001, with studies conducted in the Banerjee Laboratory (M effect size = 0.64, SE = 0.20) and in the Firestone Laboratory (M effect size = 0.42, SE = 0.16) showing significant effects in the positive direction (p = .001 and p = .1, respectively) and studies conducted in the cubicles by the Brandt Laboratory (M effect size = −0.04, SE = 0.14) and online by the Brandt Laboratory (M effect size = 0.04, SE = 0.05) showing no significant effects (p = .8 and p = .3, respectively).

Meta-Analysis on the Effects of Experimental Condition on the Desirability of Light-Emitting Products

Next, we conducted a meta-analysis to estimate the overall mean effect size of experimental condition on the desirability of light-emitting products (N = 15). The random effects meta-analysis produced a mean effect size of d = 0.13 ([CL$_{95}$] −0.04 to 0.29). There was no significant effect of experimental condition across all studies on brightness judgments (z = 1.53, p = .13). Figure 2 provides a forest plot of the effect sizes of the desirability of light-emitting products across studies. The effect of experimental condition on the desirability of light-emitting products did not differ for participants from the US (M effect size = 0.25, SE = 0.15) versus participants from the Netherlands (M effect size = 0.01, SE = 0.18), QM(2) = 2.81, p = .25. The effect was larger for studies conducted in the laboratory (M effect size = 0.33, SE = 0.13, p = .01) than for studies conducted online (M effect size = −0.09, SE = 0.15, p = .56), QM(2) = 6.37, p = .04.

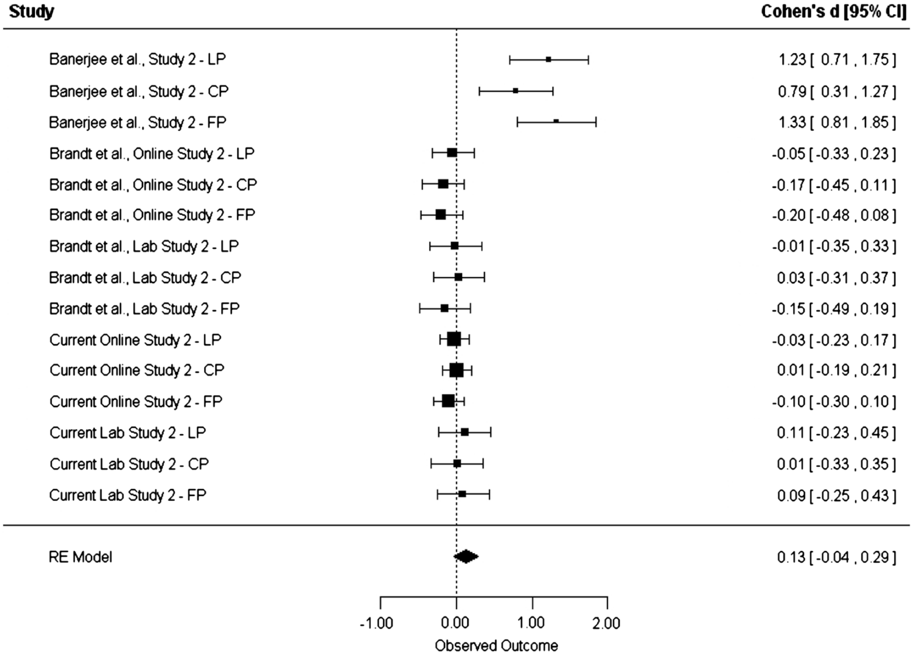

Figure 2. Cohen's *d*, 95% confidence intervals, and estimate of overall effect size from a random effects meta-analytic model for the preferences for light-emitting products. LP = lamp preference, CP = candle preference, FP = flashlight preference.

Our exploratory analysis on research laboratory where the study was conducted yielded an overall significant effect, $QM(3) = 56.22$, $p < .001$, with studies conducted in the Banerjee laboratory (M effect size = 1.10, $SE = 0.15$) showing significant effects in the positive direction ($p < .001$) and studies conducted in the Brandt laboratory (M effect size = −0.03, $SE = 0.06$) and online (M effect size = −0.06, $SE = 0.05$) showing no significant effects ($p = .64$ and $p = .25$, respectively).

Discussion

Despite conducting high-powered replication studies of BCS, we were unable to replicate the original effects in our own replication studies. Recalling ethical or unethical behavior did not have an effect on the estimated brightness of the room, the estimated watts of light in the room, or the preference for light-emitting products. A meta-analysis of available effect sizes of moral recall on perceptions of brightness indicated that on average there is a marginally significant effect that tends to be larger when tested in a laboratory setting. The meta-analysis on preferences for light-emitting products did not reveal any effect of moral recall on product preferences, suggesting that this effect may be less robust. This effect was also moderated by whether the study was in the laboratory or online, with a significant effect on average when conducted in the laboratory. Overall, we believe that there is still much to be learned about the robustness of the effect of moral recall on the perception of light. The replications and meta-analysis reported here suggest that the effect is not robust; however, two independent laboratories have observed the effect.

At this stage we think it is important to try and understand why BCS and (Firestone & Scholl, 2014) were able to detect the effect and we were not. It may be that subtle aspects of the procedure, whether in the formatting of the study, the wording of the consent form, or some other feature is essential for the effect and was different between the available studies. This is a clear possibility because Firestone and Scholl (2014) found the effect in the same online population (i.e., MTurk) where we collected our online data, even though a moderator analysis suggested that online studies produced weaker effects on average. Similarly, it seems unlikely that our Dutch laboratory studies are a cause for concern because others have detected links between immorality/morality and darkness/lightness in Dutch samples (Lakens, Semin, & Foroni, 2012), classic social psychological effects have been replicated in our Tilburg laboratories (Klein et al., 2014), and we also detected a similar null effect with online American samples. This led us to consider the ''laboratory group'' that conducted the study as a potential moderator in the meta-analyses. These moderation analyses suggest that something about the particular laboratory that conducted the study may be driving the effect. This could be something about the precise display of the stimuli within the experimental program or other aspects of the experimental setting and presentation.

One specific direction for future work is to explore the differences between our online replication attempts and the two attempts reported by Firestone and Scholl (2014). We both collected data from the MTurk population; however, subsequently we have learned that whereas we used an 80% approval rating for MTurk workers (an indicator of worker quality), Firestone and Scholl used a more stringent 95% approval rating for MTurk workers. The type of samples drawn from the MTurk population may significantly differ between these two approval rate levels and this may explain the differences between our online studies and the Firestone and Scholl online studies. It should be noted, however, that this does not explain the discrepancy between our laboratory replications and the laboratory replications of BCS.

A second direction researchers could explore is the fact that both in our laboratory studies and our online replication studies participants estimated a large range of watts as lighting the room. For example, the standard deviations for our laboratory study, where all of the participants were in individual cubicles illuminated by the same 66 Watt fluorescent light, were larger than 150 Watts. BCS, on the other hand, estimated the standard deviation to be about one sixth of the size. This may indicate more knowledge or attention to the light in the room by BCS's participants compared to our participants in the laboratory studies. In light of the issues and potential causes for our null results discussed above, future investigations into the nature of the effect of moral recall on perceptions of brightness should keep careful records of the differences between the original and replication study on more basic issues in regard to stimulus presentation, experimental context, and attention to one's surroundings to potentially find the key to the effect (cf. Brandt et al., 2014; Cesario, 2014).

In conclusion, we are hesitant to proclaim the effect a false positive based on our null findings, or a true success based on the marginally significant meta-analytic effect. Instead we think that scholars interested in how morality is grounded should be hesitant to incorporate the studies reported by BCS into their theories until the effect is further replicated and a possible explanation for the discrepancy between our findings and the original findings is identified *and* tested. Until answers to these questions are available it appears that the possibility of an abstract concept (morality) changing people's perception of something more concrete (perception of light) will remain just that, a possibility.

Acknowledgments

We thank Ellen Evers for helping collect data for the Study 2 replication in the laboratory. We especially thank Banerjee, Chatterjee, and Sinha for providing their original materials and to both the Banerjee group and Firestone and Scholl for providing additional study details for the meta-analysis. We report all data exclusions, manipulations, and measures, and how we determined our sample sizes. The authors declare no conflict-of-interest with the content of this article. Designed research: M. B., H. I., Performed research: M. B. Analyzed data: M. B. Conducted and Reported Meta-analysis: I. B. Wrote paper: M. B., H. I. The replications reported in this article were supported by a grant from the Center for Open Science awarded to the three authors (Grant # 2013-002). All materials, data, syntax, supplemental information, supplemental analyses, and the preregistered design are available here: http://osf.io/seqgd/.

PREREGISTERED OPEN DATA OPEN MATERIALS

References

*Banerjee, P., Chatterjee, P., & Sinha, J. (2012). Is it light or dark? Recalling moral behavior changes perception of brightness. *Psychological Science, 23*, 407–409.

Barsalou, L. W. (1999). Perceptual symbols system. *Behavioral and Brain Sciences, 22*, 577–660.

*Brandt, M. J., IJzerman, I., & Blanken, I. (2013). *Replication of Banerjee, Chatterjee, and Sinha (2012)* Unpublished replication report and data. https://osf.io/jpfsi/.

Brandt, M. J., IJzerman, H., Dijksterhuis, A., Farach, F. J., Geller, J., Giner-Sorolla, R., . . . Van't Veer, A. (2014). The replication recipe: What makes for a convincing replication? *Journal of Experimental Social Psychology, 50*, 217–224.

Cesario, J. (2014). Priming, replication, and the hardest science. *Perspectives on Psychological Science, 9*, 40–48.

*Firestone, C., & Scholl, B. J. (2014). "Top-down" effects where none should be found: The El Greco fallacy in perception research. *Psychological Science, 25*, 38–46.

Frank, M. G., & Gilovich, T. (1988). The dark side of self-and social perception: Black uniforms and aggression in professional sports. *Journal of Personality and Social Psychology, 54*, 74–85.

Hamlin, J. K., Wynn, K., & Bloom, P. (2007). Social evaluation by preverbal infants. *Nature, 450*, 557–559.

IJzerman, H., & Koole, S. L. (2010). From perceptual rags to metaphoric riches – Bodily, social, and cultural constraints on sociocognitive metaphors: Comment on Landau, Meier, and Keefer (2010). *Psychological Bulletin, 137*, 355–361.

IJzerman, H., & Semin, G. R. (2010). Temperature perceptions as a ground for social proximity. *Journal of Experimental Social Psychology, 46*, 867–873.

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B. Jr., Bahník, Š., Bernstein, M. J., . . . Nosek, B. A. (2014). Investigating variation in replicability: A "many labs" replication project. *Social Psychology, 45*, 142–152.

Lakens, D., Semin, G. R., & Foroni, F. (2012). But for the bad, there would not be good: Grounding valence in brightness through shared relational structures. *Journal of Experimental Psychology: General, 141*, 584–594.

Lakoff, G., & Johnson, M. (1999). *Philosophy in the flesh*. New York, NY: Basic Books.

Lee, S. W., & Schwarz, N. (2012). Bidirectionality, mediation, and moderation of metaphorical effects: The embodiment of social suspicion and fishy smells. *Journal of Personality and Social Psychology, 103*, 737–749.

Sherman, G. D., & Clore, G. L. (2009). The color of sin: White and black are perceptual symbols of moral purity and pollution. *Psychological Science, 20*, 1019–1025.

Slepian, M. L., & Ambady, N. (2014). Simulating sensorimotor metaphors: Novel metaphors influence embodied cognition. *Cognition, 130*, 309–314.

[*] Indicates studies included in the meta-analysis.

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software, 36*, 1–48.

Webster, G. D., Urland, G. R., & Correll, J. (2012). Can uniform color color aggression? Quasi-experimental evidence from professional ice hockey. *Social Psychological and Personality Science, 3*, 274–281.

Zhong, C. B., Bohns, V. K., & Gino, F. (2010). Good lamps are the best police: Darkness increases dishonesty and self-interested behavior. *Psychological Science, 21*, 311–314.

Received February 28, 2013
Accepted December 17, 2013
Published online May 19, 2014

Mark J. Brandt

Department of Social Psychology
Room P06
P.O. Box 90153
Tilburg University
5000 Tilburg
The Netherlands
E-mail m.j.brandt@tilburguniversity.edu

Instructions to Authors – *Social Psychology*

Social Psychology is a publication dedicated to international research in social psychology as well as a forum for scientific discussion and debate. The sole publishing language is English, and there are six issues per year.

Aims and Scope: *Social Psychology* publishes innovative and methodologically sound research and serves as an international forum for scientific discussion and debate in the field of social psychology. Topics include all basic social psychological research themes, methodological advances in social psychology, as well as research in applied fields of social psychology. The journal focuses on original empirical contributions to social psychological research, but is open to theoretical articles, critical reviews, and replications of published research.

The journal welcomes original empirical and theoretical contributions to basic research in social psychology, to social psychological methods, as well as contributions covering research in applied fields of social psychology, such as economics, marketing, politics, law, sports, the environment, the community, or health. Preference will be given to original empirical and experimental manuscripts, but theoretical contributions, critical reviews, and replications of published research are welcome as well.

Experience and Innovation: The journal was published until volume 38 (2007) as the *Zeitschrift für Sozialpsychologie* (ISSN 0044-3514). Drawing on over 30 years of experience and tradition in publishing high-quality, innovative science as the *Zeitschrift für Sozialpsychologie*, *Social Psychology* has an internationally renowned team of editors and consulting editors from all areas of basic and applied social psychology, thus ensuring that the highest international standards are maintained.

Rapid Turnaround: *Social Psychology* offers a rapid and transparent peer-review process and a short time-lag between acceptance of papers and publication. The time between manuscript submission and editorial decision is usually less than eight weeks. Mean time from submission to first decision (2012): 60 days.

Please read the following information carefully *before* submitting a document to *Social Psychology*:

All manuscripts should be submitted online at http://www.editorialmanager.com/sopsy. Please follow the online instructions for submission. Should you have any technical queries regarding this process, please contact Juliane Munson, Hogrefe Publishing (E-mail journalsproduction@hogrefe.com, Tel. +49 551 99950-422, fax +49 551 99950-425). Please direct any editorial questions to the editorial office, E-mail Social-Psychology@uni-koeln.de.

Types and Length of Manuscripts. *Original Articles* report empirical and/or theoretical contributions to social psychological research; they should not exceed 8,000 words excluding tables, figures, and references. *Research Reports* present concise descriptions of innovative empirical findings; they should not exceed 2,500 words excluding tables, figures, and references. *Replications* offer the opportunity to report successful or failed replications of existing research; they should not exceed 2,500 words excluding tables, figures, and references.

Blind Reviewing is Mandatory. Authors should therefore remove all potentially identifying information from the manuscript, replacing names and any indication of the university where a study was conducted by neutral placeholders.

To facilitate blind reviewing, the **Title Page** of the submitted manuscript should include only the paper's title and running head. A second title page including all author information should be submitted as a separate document. This should include the title, author name(s) (preceded by first names, but with no academic titles given); name of institute (if there is more than one author or institution, affiliations should be indicated, using superscript Arabic numerals); and an address for correspondence (including the name of the corresponding author with e-mail and phone numbers).

An **Abstract** (maximum length 120 words) should be printed on a separate sheet for original papers, reviews, and reports. A maximum of 5 keywords should be given after the abstract.

Reference Citations in the text and in the reference list proper should follow conventions listed in the *Publication Manual of the American Psychological Association* 6th ed. (*APA Manual*).

Tables should be numbered using Arabic numerals. Tables must be cited in the text (e.g., "As shown in Table 1, . . ."). Each table should be printed on a separate sheet. Below the table number, a brief descriptive title should be given; this should then be followed by the body of the table. It is recommended that each table should also include a brief explanatory legend.

Figures should be numbered using Arabic numerals. Each figure must be cited in the text (e.g., "As illustrated in Figure 1, . . .") and should be accompanied by a legend on a separate sheet. As online submission requires papers to be submitted as one file, figures and tables etc should be embedded or appended to the paper and not be sent as separate files. However, upon acceptance of an article, it may be necessary for figures to be supplied separately in a form suitable for better reproduction: preferably high-resolution (300 dpi) or vector graphics files. Where this is necessary, the corresponding author will be notified by the publishers. Figures will normally be reproduced in black and white only. While it is possible to reproduce color illustrations, authors are reminded that they will be invoiced for the extra costs involved.

Scientific Nomenclature and Style: Authors should follow the guidelines of the *APA Manual* regarding style and nomenclature. Authors should avoid using masculine generic forms in their manuscripts. General statements about groups of people should be written in gender-neutral form (see *APA Manual*, pp. 73–74); when presenting examples, authors may alternate between female and male forms throughout their text.

Language: It is recommended that authors who are not native speakers of English have their papers checked and corrected by a native-speaker colleague before submission. Standard US American spelling and punctuation as given in *Webster's New Collegiate Dictionary* should be followed.

Proofs: PDF proofs will be sent to the corresponding author. Changes of content or stylistic changes may only be made in exceptional cases in the proofs. Corrections that exceed 5% of the typesetting costs may be invoiced to the authors.

Offprints: Hogrefe will send the corresponding author of each accepted paper free of charge an e-offprint (PDF) of the published version of the paper when it is first released online. This e-offprint is provided for the author's personal use, including for sharing with co-authors.

Copyright Agreement: By submitting an article, the author confirms and guarantees on behalf of him-/herself and any co-authors that the manuscript has not been submitted or published elsewhere, and that he or she holds all copyright in and titles to the submitted contribution, including any figures, photographs, line drawings, plans, maps, sketches, and tables, and that the article and its contents do not infringe in any way on the rights of third parties. The author indemnifies and holds harmless the publisher from any third-party claims.

The author agrees, upon acceptance of the article for publication, to transfer to the publisher the exclusive right to reproduce and distribute the article and its contents, both physically and in nonphysical, electronic, or other form, in the journal to which it has been submitted and in other independent publications, with no limitations on the number of copies or on the form or the extent of distribution. These rights are transferred for the duration of copyright as defined by international law. Furthermore, the author transfers to the publisher the following exclusive rights to the article and its contents:

1. The rights to produce advance copies, reprints, or offprints of the article, in full or in part, to undertake or allow translations into other languages, to distribute other forms or modified versions of the article, and to produce and distribute summaries or abstracts.
2. The rights to microfilm and microfiche editions or similar, to the use of the article and its contents in videotext, teletext, and similar systems, to recordings or reproduction using other media, digital or analog, including electronic, magnetic, and optical media, and in multimedia form, as well as for public broadcasting in radio, television, or other forms of broadcast.
3. The rights to store the article and its content in machine-readable or electronic form on all media (such as computer disks, compact disks, magnetic tape), to store the article and its contents in online databases belonging to the publisher or third parties for viewing or downloading by third parties, and to present or reproduce the article or its contents on visual display screens, monitors, and similar devices, either directly or via data transmission.
4. The rights to reproduce and distribute the article and its contents by all other means, including photomechanical and similar processes (such as photocopying or facsimile), and as part of so-called document delivery services.
5. The right to transfer any or all rights mentioned in this agreement, as well as rights retained by the relevant copyright clearing centers, including royalty rights to third parties.

Online Rights for Journal Articles: Guidelines on authors' rights to archive electronic versions of their manuscripts online are given in the Advice for Authors on the journal's web page at www.hogrefe.com.