Saving Science Through Replication Studies

John E. Edlund, PhD, Rochester Institute of Technology

Kelly Cuccolo, M.A., University of North Dakota

Megan S. Irgens, M.A., University of Arizona

Jordan R. Wagge, PhD, Avila University

Martha S. Zlokovich, PhD, Psi Chi, the International Honor Society in Psychology

Abstract

The scientific enterprise has long been based on the presumption of replication, although recently scientists have become aware of various corruptions of the enterprise which have hurt replicability. In this paper, we begin by considering three illustrations of research paradigms that have all been subject to intense scrutiny through replications and theoretical concerns. The three paradigms are one for which the corpus of research points to a real finding, one for which the corpus of research points to a significantly attenuated effect, and one for which the debate is ongoing. We then discuss what scientists can learn -- and how science can be improved -- through replications more generally. From there, we discuss what we believe needs to be done to improve scientific inquiry with regard to replication moving forward. Finally, we conclude by providing readers with several different approaches to replication and how these approaches progress science. The approaches discussed include multilab replications of many effects, multilab replications of specific effects, adversarial collaborations, and stand-alone applications.

Keywords: Replication, Science, Collaboration, Metascience

Saving Science Through Replication Studies

Is the reputation of science in a crisis moment? Numerous recent studies have highlighted examples of corruption of the scientific enterprise, ranging from questionable statistical analyses and reporting (Servick, 2018), to the falsification of sources (Bartlett, 2019), data (Levelt Committee, Noort Committee, & Drenth Committee, 2012), and peer review (Retraction Notice, 2014), among other problems. This recent focus on scientific practice and reporting has led scientists, philosophers, and the public to question how much of the scientific enterprise can be trusted. Such issues lead the general public and media to mistakenly believe that most science is not conducted using rigorous scientific methodology, seeing it as being unreplicable and failing to make precise predictions (Lilienfeld, 2012). Such public skepticism is particularly troubling as it may contribute to individuals' decisions to avoid mental health services, not support federal funding of psychological research, believe misrepresentation of psychological research in "pop" culture media outlets, or be unable to accurately apply findings to political policy (Lilienfeld, 2012). Given such, some authors have argued that the scientific enterprise is truly corrupt and needs wholesale changes (Hubbard, 2015). Although we agree that serious problems have recently been highlighted in both public and scientific discourse, we believe that because these issues are being examined and addressed, the scientific enterprise is working well. In this paper, we will use examples to highlight how replications contribute to the scientific enterprise.

Overview of Replication

Replications, varying in their forms, aims, and breadth, all hold scientific value (Soderberg & Errington, 2019; Zwaan et al., 2018). One may undertake a replication for various reasons, ranging from investigating suspected fraud, measurement and sampling error, to considering

moderators, mediators, and boundary effects. Replications may be especially needed when the original study contains flashy findings that are hard to believe, counterintuitive to existing data, or controversial; however, replications should ideally be done either prior or post-publication of any data. Importantly, replications represent a systematic investigation of previously published findings which, over time, can identify constraints on the original effect (e.g., an idiosyncrasy in the original sample or research design). Different types of replications and scales on which they are conducted can provide unique pieces of evidence towards explaining an original effect. For example, direct replications are critical for identifying systematic and unsystematic errors in findings because they are strict repetitions of experimental procedures (Simons, 2014). Through engagement in direct replication by multiple laboratories, Type I and Type II errors, boundary conditions, or idiosyncrasies in the original study can be identified, leading to refinement of the original hypothesis. Conceptual replications are typically employed when there are theoretical reasons for assuming the effect could vary in different groups or settings (e.g., moderating effect of culture; Zwaan et al., 2018).

Further, the strength of evidence and conclusions drawn about the replication will vary with its scale. A replication by one lab might identify idiosyncrasies in the original sample that impacted the effect, but with coordinated replications from many labs in various locations, the impact of idiosyncrasies (e.g., local demographic factors) on an effect can be explicitly examined. Below, we use examples to highlight various types of replications and underscore replications as constituting a continual process that ultimately enhances the accumulation of scientific knowledge.

Illustrations of Replication in Action

Illustration One: Sex Difference in Jealousy

Our first illustration of how replication supported the original effect comes from the evolutionary behavioral sciences. In 1992, Buss et al. demonstrated that men (relative to women) were more upset about a partner's imagined sexual infidelity, whereas women (relative to men) were more upset about a partner's imagined emotional infidelity. Using both physiological and self-report measures across three studies, Buss et al. (1992) demonstrated this effect. This was one of the first widely read and cited (nearly 2000 times) papers from evolutionary behavioral science.

Given the provocative findings (suggesting an innate difference between men and women in what makes them jealous), this study inspired many replications attempting to either support or refute its findings. The first published replications (e.g., DeSteno & Salovey, 1996; DeSteno et al., 2002; Harris, 2002) demonstrated some perceived limitations to the effect (e.g., the effect was replicated when looking at hypothetical but not actual infidelity; Harris, 2002) whereas follow-up studies addressed those concerns (e.g., Edlund et al., 2006). Importantly, these follow-up studies allowed for important moderators to be identified and systematically tested (e.g., sexual orientation; Scherer et al., 2013; the emotional term being used; Shackelford et al., 2000). With a robust field having emerged, a recent meta-analysis found that the sex difference in jealousy is a real effect that occurs consistently with a moderate effect size (Sagarin et al., 2012).

We believe that the sex difference in jealousy findings nicely illustrates how science is supposed to work, and in this case, suggests that the original finding was a real effect. An original study was published making a notable claim, which was then supported using everything

from reproducibility to direct replications to conceptual replications and indeed beyond as researchers have explored related phenomena and explanations, such as studies that have suggested that the original theoretical explanation was too narrow (Edlund et al., 2019). Some of the replications suggested alternative explanations, whereas others further clarified the effect. Importantly, no *single* study that attempted to replicate the effect was the arbiter of whether the original result was a true finding -- it was the accumulation of many studies that replicated the original finding, as well as the failed replications pointing to important moderators that contributed to our comprehensive understanding of the effect. Furthermore, simply because an effect has an important moderator does not mean that it is not real. Rather, moderators refine the original idea so that it can be more precisely understood.

Illustration Two: Power Posing

In our second illustration scientific evidence produced by the replication attempts conflicted with an original finding. In 2010, Carney, Cuddy, and Yap published research suggesting that posing for two minutes in a manner associated with power (e.g., standing tall with one's arms crossed) would cause hormonal and behavioral changes in participants (such that they would actually be more powerful). Cuddy subsequently described this research in a TED Talk that has been viewed millions of times.

Despite the widespread popularity of the concept, research that followed the initial study did not find similar results. For example, a study that explored the hormonal measures used by Carney et al. (2010) with a much larger sample size did not find similar results (Ranehill et al., 2015). Another early Carney et al. (2010) replication found the opposite pattern of results on the "feelings" measures: power poses were associated with lower feelings of power (Garrison et al.,

2016). Importantly, this replication attempt was publicly pre-registered and had a sufficient sample size to detect the effect based on the original research.

Interest in power posing was so strong in the mid-2010s that a peer-reviewed journal sought out teams to study power posing comprehensively for a special issue (Cesario, Jonas, & Carney, 2017). Eight separate teams published research exploring power posing. Across the submissions, the editors concluded that there is limited to no evidence to suggest that power posing impacts any measure. However, a meta-analysis (Gronau et al., 2017) of the projects showed that the participants felt somewhat more powerful even though they did not behave differently nor show any physiological change.

Other researchers have explored this line of research from a purely statistical approach. Simmons and Simonsohn (2017) applied a p-curve statistical analysis to data presented in an early review article (Carney et al., 2015), and found that the reported results were indistinguishable from the pattern one would expect from a null effect with publication bias. Subsequently, Cuddy, Schultz, and Fosse (2018) argued that the p-curve analysis revealed evidentiary value of power posing across studies. However, Crede (2018) determined -- by looking at the studies' methods rather than the statistics -- that any effect of power posing was likely a result of comparing expansive (power) poses to contractive (shrinking) poses rather than neutral (control) poses, and it was likely the contractive poses that negatively affected comparison group outcomes rather than power poses positively affecting outcomes.

This example illustrates how replication is supposed to work when the original effect (hormonal and behavioral changes based on power posing) is a spurious finding (Type I error). What matters more for our purposes is that no single study or approach proved to be the

definitive replication. The consensus that power posing does not affect hormones or behaviors has been derived by the accumulation of many studies, all looking at the research question somewhat differently.

Illustration Three: Cleanliness Primes and Moral Judgments

Priming is an umbrella term that refers to a phenomenon in which exposure to a stimulus will affect subsequent responses to related stimuli. In 2005, Holland, Hendriks, and Aarts demonstrated that the concept of "cleanliness" can be primed (i.e., with a citrus cleanser scent), and that this cleanliness prime affected subsequent behaviors (i.e., cleaning a table while eating). This finding led to a series of experiments (and subsequent replications) on the relationship between cleanliness and moral judgments (Schnall et al., 2008). In two experiments, Schnall and colleagues (2008) observed that individuals make less severe moral judgments of actions taken by characters facing moral dilemmas (i.e., judged the actions taken as less "wrong") when primed with cleanliness. In experiment one, participants completed sentence scramble tasks that contained either neutral words or cleanliness related words, and then subsequently rated actions taken by six characters facing moral dilemmas. After the cleanliness prime, participants rated the moral dilemmas less severely than they did after the neutral prime, which suggested that certain concepts can impact moral judgments. A second study demonstrated a similar effect on moral judgments using handwashing (versus no handwashing) as a manipulation following a short film that elicited disgust. Here, two different experimental protocols were used to explore the same concept: the impact of cleanliness primes on moral judgments. Notably, the authors interpreted these results as evidence that primes about cleanliness can reduce the severity of moral judgments. A fast approach to interpreting results was used -- generalizability was assumed, and would be, until contradictory evidence accumulated (Yarkoni, 2019).

Given the surprising nature of these findings, replication attempts were soon underway. Johnson et al. (2014) undertook a direct replication designed to obtain statistical power of at least .99. Johnson and colleagues (2014) used a similar protocol to Schnall et al.; nevertheless, they could not replicate the original study's findings. Effect sizes in the replication attempts were smaller than in the original studies and were not statistically significant. Perceptions about this "failed" replication also sparked debate and led to subsequent work seeking to clarify this 'cleanliness hypothesis.' Some undertook investigations to identify potential sources of error that may have contributed to Johnson et al.'s results. For example, Huang (2014) conducted a replication and focused on identifying potential moderator variables using Johnson et al.'s protocol with participants from the United States on Amazon Mechanical Turk (MTurk). In study one, MTurk participants completed the sentence scramble and survey measures as per previous studies, but no priming effect emerged (i.e., moral ratings did not differ). However, when examining response effort, an effect did emerge; the cleanliness hypothesis was supported in the low response effort subsample.

A second study further examined this effect. In study two, Huang (2014) induced two levels of response effort; participants received instructions that emphasized either working quickly or being accurate. While no overall effect of prime emerged, there was an interaction effect such that the cleanliness conditions rated specific vignettes less harshly in the low response effort condition. As such, Huang (2014) concludes that there is a boundary condition on Schnall's original finding -- one of response effort. Other replications (e.g., Besman et al., 2013) have found contradictory support for the original study. These studies highlight how a "failed" replication may have valid underlying causes (i.e., a moderator), and serves as a reminder that each replication is a piece of evidence in the emerging effect puzzle. Indeed, Johnson et al.

(2014) note that these replications are pieces of evidence contributing to an explanation of the relationship between cleanliness and perceived severity of moral dilemmas. Indeed, the rich accumulation of empirical literature stemming from efforts to clarify the original effect reflects our view of science, in which results are repeatedly verified and updated as new evidence emerges.

The Role of Replications in Understanding Effects - Moving Beyond Our Illustrations

When a researcher initially finds an effect, and a follow-up replication does not, one possibility is that one of the studies was "wrong" in its conclusion (Type I error). However, other possibilities exist. For instance, subtle changes in the method employed by the follow-up study may fundamentally change the study, which then may point to an important moderator to the underlying effect. *As such, both studies may be "right" at the same time, despite reaching different conclusions.*

One demonstration of this was seen in our first replication illustration. Several studies using terms like "happiness" or "focus" (e.g., Harris, 2002; Shackelford et al., 2000) failed to find a sex difference, whereas other studies using the term "jealousy" found an effect (Edlund et al., 2006). Sagarin et al. (2012) meta-analyzed these studies and determined that the specific emotion label used was an important moderator of the effect. As such, all of these studies are correct -- there are no sex differences in happiness in response to a partner's infidelity, whereas there are sex differences in jealousy in response to a partner's infidelity.

Replications can demonstrate the limits of a particular effect, but these limits do not mean the original effect is not real. For example, persuasion research has shown that the lowball effect is real (Cialdini et al., 1978), but it only manifests when the same person makes both requests

(Burger & Petty, 1981). Research in psychiatry has suggested that Naloxone is effective in treating opioid dependence (Pfeiffer et al., 1986), but it works best in conjunction with other therapy approaches (Carroll et al., 2001). With regard to cancer medicine, research has suggested that only certain types of breast cancers will respond to hormone therapies (Ek et al., 2008). Importantly, it is uncontroversial to suggest that moderators exist and provide important boundary conditions to effects.

The previous illustrations highlight that replication is a *process*: one replication by one lab cannot "prove" a finding true false. There are numerous reasons why an effect may not replicate when being examined by investigators outside of the original team, such as measurement or sampling error, misconstrual of operationalization of key variables, misrepresentation of the original effect size, or the existence of moderators (Lynch, et al., 2015; Stroebe & Strack, 2014). Further, there is no single, agreed-upon, marker of replication "success" (Maxwell et al., 2015). Individuals conducting replications should use *p*-values, effect sizes, and meta-analyses to provide information about the strength of the replication's effect. As such, it may be useful for researchers to conceptualize replications beyond a binary (failure/success) because -- as seen in the examples provided -- both of these outcomes spark rich debates, spur further research, and identify qualifiers for the original effect. All in all, the summation of replications is necessary to identify error and any extraneous factors that mediate the emergence of the effect, and for statements to be made about the strength of (or level of evidence for) a particular effect.

These examples may also highlight a concern regarding the nexus of replicability and generalizability, a concern that is thoroughly explored by Yarkoni (2019). When undertaking replication efforts, one needs to contextualize the original findings. This contextualization

includes understanding specifications made in statistical models used by original authors, specification of group comparisons and operationalizations, and acknowledgments pertaining to constraints on generalizability. For instance, as Crede (2019) highlighted, a plethora of power posing studies either a) failed to obtain baseline measures for feelings of power or b) failed to use a natural/neutral control group. In these studies, power posing was compared to a contractive (e.g., slouching) postural position, and as such, answer questions more aligned with "does power posing increase feelings of power relative to contractive poses?" rather than "does power posing increase feelings of power?" as results have been interpreted. As pointed out by Crede (2019) these results suggest a recommendation to not slouch but do not support a recommendation to power pose. This also illustrates how it can be difficult to achieve a clear picture of an effect or its moderators until a set of replications can be considered, in context, by independent researchers and labs.

Increasing Trust in Science Through Replication

Through replications, we can increase our confidence that a particular effect can be found under specified conditions. Successful replications by the original lab (i.e., 'self-check') can increase trust by flushing out Type I error (Simons, 2014). Replications by other labs allow for increased credibility, generalizability of results, room for identification of moderators, and trust that the finding was not a Type I error (Schmidt, 2009; Simons, 2014). Relatedly, both peers and the public perceive researchers who admit shortcomings (i.e. limitations/errors) of their findings to be more ethical and able (Ebersole et al., 2016), and fewer negative reputational outcomes are ascribed to those who admit their "wrongness" after a failed replication (Fetterman & Sassenberg, 2015). Along a similar vein, a 2018 survey on public attitudes towards science indicated that funding is a major reason for distrusting science (Wissenschaft im Dialog/Kantar

Emnid, 2018). This suggests that replications from independent labs removed from pro-original-effect funding sources may be important for increasing public trust in science. As such, engaging in replications and having discourse surrounding the outcomes of replication attempts may signal to others in the scientific community, and the public, that science can monitor and correct findings.

Indeed, replications can be strong indicators to other scientists, the public, and policymakers that things are working as they should. As mentioned above, because replications can ensure sound results and spark conversations about the research findings, they can propel science forward. However, it is important that scientists come to a broad agreement that replications are an important tenet of science and that conducting them is worth their time and effort. By valuing replications in our scientific communities, we normalize replications as part of the scientific process, allowing for beliefs to be modified as evidence emerges (Ferguson, 2015).

Moving Forward: Best Steps for Science

As we hope our illustrations demonstrate, we believe that science is doing well with replications. In conjunction with the weight that replications may hold pertaining to trust in science, we believe that there is room for improvement, as evidenced by the continued vigorous debates on improving scientific inquiry and reporting. We suggest that all scientific fields adopt better policies, incentives, and procedures which more highly value replications as a method of correcting - or confirming - scientific findings.

One of the consistent themes that emerges from the illustrations above is that the success or failure of a single replication does not alone define whether or not an original effect is supported. *Replication is a process.* The single study is ambiguous by its very nature -- it should

be considered the beginning of a conversation, not the end. It is only when we collect many data points across many studies that we begin to understand the nature of the phenomenon we are studying. We need to move beyond thinking that a particular replication is a "success" or a "failure" -- rather, each study represents a piece of evidence in favor of or against a particular effect. Therefore, science (through scientific societies, journals, research centers, and academic institutions) needs to take steps that will encourage this kind of research. We believe all of these groups have a role to play in incentivizing replications. For brevity, we briefly detail incentivizations in Table 1. Additionally, we note that there are several approaches to replications. Although fully detailing these approaches is beyond this paper's scope, we briefly summarize four generalized approaches that can yield very fruitful results in Table 2.

Another factor complicating the understanding of replications is the reaction of some scientists to replications of their work. When a scientist publishes work that becomes commonly cited or famous, part of that researcher's identity becomes wrapped up in that work. As such, any work that questions those findings can easily be interpreted as an attack on the self, and the researcher then may react in kind. This was illustrated by Baumeister (2019), who details much of the replication debate surrounding ego-depletion. Part of what Baumeister notes is how personal and vitriolic the attacks have been on the initial research and researchers in that scientific debate, as illustrated by this excerpt: "Unfortunately, perhaps, many scholars now rely increasingly on social media to get their information about the field, and social media can be dominated by bullies and ad hominem attacks, thereby intimidating others from speaking out" (p., 3). As if to prove his point, a preprint of Baumeister's chapter was posted in a Facebook group dedicated to research methods, sparking a sustained, personal, and vitriolic attack on the points raised in the chapter. Other examples have been seen in the literature where the researcher

whose work is being questioned launches personal and vitriolic attacks against the replicators. Indeed, some researchers have even threatened to sue reporters and researchers who question their work (Yong, 2012).

In our view, the vitriol shown in such debates hurts science generally by discouraging good-faith replication attempts. Furthermore, we contend that the sooner replications are conducted and published, the less likely it is that the original researcher will have become so ego-invested in the original findings as to resist any null findings by future researchers. The sorts of vitriolic exchanges described above do not benefit the field in any way. Rather than helping advance the science, they lead the various sides of the debate to entrench, not to progress the field. Research has shown that the public's view of science is being hurt by the ugliness of the debates and other unethical activities by a minority of scientists (Anvari & Lakens, 2019). We believe that rather than engaging in personal and vitriolic attacks, scientists should continue collegial exchanges about research, including a recognition that *replication is part of the normal thrust and parry of science*. This is the way science is supposed to work: science needs replication.

It is also critical for scientists to recognize that short-term uncertainty is a part of the field. As humans, we struggle with ambiguity and uncertainty (Hofstede, 1984) and as scientists we are fundamentally and foremost humans. However, science is rarely unambiguous, particularly in the short term. As scientists, we need to better understand this for ourselves, but especially mindful when we communicate about science to our students and the wider world (Fischhoff, 2013). Effectively communicating about science can help the general public make informed, sound decisions across many areas of their lives. We also need to recognize that for

many of us, our training has biased us against replications, even though valuing and encouraging replications can better science.

“The King is dead. Long live the King (Circa 15th Century)”

Overall, we believe that science as a whole is doing well. The debate about replication is part of what is supposed to occur in science. Although we believe that incentives should be enhanced to further encourage replications, science is in good shape and we have more tools than ever to support it. As such, (some of) the ways things were done in the 20th and early 21st centuries are (and should be) dead; yet the current focus on replications did not arise out of a vacuum. Scientific inquiry and reporting processes will continue to grow stronger as the culture of science more readily accepts replication science. Long live science!

References

Anvari, F., & Lakens, D. (2019). The replicability crisis and public trust in psychological science. *Comprehensive Results in Social Psychology*, 1-21. https://doi.org/10.1080/23743603.2019.1684822

Bartlett, T. (2019). Trump's 'China Muse' Has an Imaginary Friend. *Chronicle of Higher Education*, https://www.chronicle.com/interactives/20191015-navarro

Baumeister, R. (September 4, 2019). Self-Control, Ego Depletion, and Social Psychology's Replication Crisis. Retrieved from: https://psyarxiv.com/uf3cn/

Besman, M., Dubensky, C., Dunsmore, L., & Daubman, K. (2013). Cleanliness primes less severe moral judgments. Retrieved October 01, 2020 from http://www.PsychFileDrawer.org/replication.php?attempt=MTQ5

Burger, J. M., & Petty, R. E. (1981). The low-ball compliance technique: Task or person commitment?. *Journal of Personality and Social Psychology*, 40(3), 492. https://doi.org/10.1037/0022-3514.40.3.492

Burns, D.M., Fox, E.L., Greenstein, M., Olbright, G., & Montgomery, D. (2019). An old task in new clothes: A preregistered direct replication attempt of enclothed cognition effects on Stroop performance. *Journal of Experimental Social Psychology, 83,* 150-156. https://doi.org/10.1016/j.jesp.2018.10.001.

Buss, D. M., Larsen, R. J., Westen, D., & Semmelroth, J. (1992). Sex differences in jealousy: Evolution, physiology, and psychology. *Psychological Science*, 3(4), 251-255. https://doi.org/10.1111/j.1467-9280.1992.tb00038.x

Carney, D. R., Cuddy, A. J., & Yap, A. J. (2010). Power posing: Brief nonverbal displays affect neuroendocrine levels and risk tolerance. *Psychological Science*, *21*(10), 1363-1368. https://doi.org/10.1177/0956797610383437

Carroll, K. M., Ball, S. A., Nich, C., O'Connor, P. G., Eagan, D. A., Frankforter, T. L., ... & Rounsaville, B. J. (2001). Targeting behavioral therapies to enhance naltrexone treatment of opioid dependence: efficacy of contingency management and significant other involvement. *Archives of General psychiatry*, 58(8), 755-761. doi:10.1001/archpsyc.58.8.755

Cesario, J. (2014). Priming, replication, and the hardest science. *Perspectives on Psychological Science*, 9(1), 40-48. https://doi.org/10.1177/1745691613513470

Cesario, J., Jonas, K.J., & Carney, D.R. (2017) CRSP special issue on power poses: what was the point and what did we learn?, *Comprehensive Results in Social Psychology, 2:1,* 1-5, DOI: 10.1080/23743603.2017.1309876

Chambers, C. (October 15th, 2018). Reproducibility meets accountability: introducing the replications initiative at Royal Society Open Science Retrieved from: https://blogs.royalsociety.org/publishing/reproducibility-meets-accountability/

Cialdini, R. B., Cacioppo, J. T., Bassett, R., & Miller, J. A. (1978). Low-ball procedure for producing compliance: commitment then cost. *Journal of personality and Social Psychology*, 36(5), 463. https://doi.org/10.1037/0022-3514.36.5.463

Cuccolo, K., Irgens, M. S., Zlokovich, M. S., Grahe, J., & Edlund, J. E. (2020). What Crowdsourcing Can Offer to Cross-Cultural Psychological Science. *Cross-Cultural Research*. https://doi.org/10.1177/1069397120950628

Cuddy, A. J., Schultz, S. J., & Fosse, N. E. (2018). P-curving a more comprehensive body of research on postural feedback reveals clear evidential value for power-posing effects: Reply to Simmons and Simonsohn (2017). *Psychological Science*, *29*(4), 656-666. https://doi.org/10.1177/0956797617746749

DeSteno, D. A., Bartlett, M. Y., Bravermann, J., & Salovey, P. (2002). Sex differences in jealousy: Evolutionary mechanism or artifact of measurement? *Journal of Personality and Social Psychology*, 83, 1103-1116. https://doi.org/10.1037/0022-3514.83.5.1103

DeSteno, D. A. & Salovey, P. (1996). Evolutionary origins of sex differences in jealousy? Questioning the "fitness" of the model. *Psychological Science*, 7, 367-372. https://doi.org/10.1111/j.1467-9280.1996.tb00391.x

Drotar, D. (2010). Editorial: A Call for Replications of Research in Pediatric Psychology and Guidance for Authors. Journal of Pediatric Psychology, (35:8), 801–805, https://doi.org/10.1093/jpepsy/jsq049

Ebersole, C. (May 28, 2019). A Critique of the Many Labs Projects. Retrieved from: https://cos.io/blog/critique-many-labs-projects/

Ebersole, C. R., Atherton, O. E., Belanger, A. L., Skulborstad, H. M., Allen, J. M., Banks, J. B., & Nosek, B. A. (2016). Many Labs 3: Evaluating participant pool quality across the academic semester via replication. *Journal of Experimental Social Psychology, 67,* 68–82. https://doi.org/10.1016/j.jesp.2015.10.012

Edlund, J.E., Buller, D.J., Heider, J.D., Scherer, C.R., Farc, M.M., Sagarin, B.J., & Ojedokun, O. (2019). The sex difference in jealousy: Lost certainty or lost opportunities? *Psychological Reports*, 122 (2), 575-592. https://doi.org/10.1177/0033294118806556

Edlund, J.E. Heider, J. D., Nichols, A. L., McCarthy, R.J., Wood, S.E., Scherer, C. R., Hartnett, J.H., & Walker, R. (2018). Sex Differences in Jealousy: The (Lack of) Influence of Researcher Theoretical Perspective. Journal of Social Psychology, 158 (5), 515-520.

Edlund, J. E., , Fare, M.-M., & Sagarin, B. J. (2006). Sex differences in jealousy in response to actual infidelity. *Evolutionary Psychology*, 4, 462-470. https://doi.org/10.1177/147470490600400137

Ek, R. O., Yildiz, Y., Cecen, S., Yenisey, C., & Kavak, T. (2008). Effects of tamoxifen on myocardial ischemia-reperfusion injury model in ovariectomized rats. *Molecular and Cellular Biochemistry*, 308(1-2), 227-235. DOI 10.1007/s11010-007-9633-0

Ferguson, C. J. (2015). “Everybody knows psychology is not a real science”: Public perceptions of psychology and how we can improve our relationship with policymakers, the scientific community, and the general public. *American Psychologist*, 70(6), 527. https://doi.org/10.1037/a0039405

Fetterman AK and Sassenberg K (2015) The reputational consequences of failed replications and Wrongness admission among scientists. *PLoS ONE* 10(12): e0143723. https://doi.org/10.1371/journal.pone.0143723

Fischhoff, B. (2013). The sciences of science communication. Proceedings of the National Academy of Sciences, 110 (Supplement 3), 14033-14039. https://doi.org/10.1073/pnas.1213273110

Grahe, J. E., Reifman, A., Hermann, A. D., Walker, M., Oleson, K. C., Nario-Redmond, M., & Wiebe, R. P. (2012). Harnessing the Undiscovered Resource of Student Research Projects. *Perspectives on Psychological Science*, *7*(6), 605–607. https://doi.org/10.1177/1745691612459057

Gronau, Q.F., Erp, S.V., Heck, D.W., Cesario, J., Jonas, K.J., & Wagenmakers, E.J. (2017). A Bayesian model-averaged meta-analysis of the power pose effect with informed and default priors: the case of felt power. *Comprehensive Results in Social Psychology*, 2:1, 123-138, DOI: 10.1080/23743603.2017.1326760

Harris, C. R. (2002). Sexual and romantic jealousy in heterosexual and homosexual adults. *Psychological Science*, 13, 7-12. https://doi.org/10.1111/1467-9280.00402

Hofstede, G. (1984). The cultural relativity of the quality of life concept. *The Academy of Management Review*, 9(3), 389-398. doi:http://dx.doi.org.ezproxy.rit.edu/10.2307/258280

Holland, R. W., Hendriks, M., & Aarts, H. (2005). Smells like clean spirit: Nonconscious effects of scent on cognition and behavior. *Psychological Science*, 16(9), 689-693. https://doi.org/10.1111/j.1467-9280.2005.01597.x

Huang, J. L. (2014). Does cleanliness influence moral judgments? Response effort moderates the effect of cleanliness priming on moral judgments. *Frontiers in Psychology*, *5*, 1276. https://doi.org/10.3389/fpsyg.2014.01276

Hubbard, R. (2015). Corrupt research: The case for reconceptualizing empirical management and social science. Sage Publications.

Johnson, D. J., Cheung, F., and Donnellan, M. B. (2014a). Does cleanliness influence moral judgments? A direct replication of Schnall, Benton, and Harvey (2008). *Social Psychology*. 45, 209–215. doi: 10.1027/a000001

Johnson, B. T., & Eagly, A. H. (1989). Effects of involvement on persuasion: A meta-analysis. *Psychological Bulletin*, 106(2), 290–314. https://doi.org/10.1037/0033-2909.106.2.290

Kerr, N. L., Ao, X., Hogg, M. A., & Zhang, J. (2018). Addressing replicability concerns via adversarial collaboration: Discovering hidden moderators of the minimal intergroup discrimination effect. *Journal of Experimental Social Psychology*, 78, 66–76. https://doi.org/10.1016/j.jesp.2018.05.001

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Bahník, Š., Bernstein, M. J., . . . Nosek, B. A. (2014a). Investigating variation in replicability: A "Many Labs" replication project. *Social Psychology, 45,* 142–152. https://doi.org/10.1027/1864-9335/a000178

Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams, R. B., Alper, S., … Nosek, B. A. (2018). Many Labs 2: Investigating variation in replicability across samples and settings. *Advances in Methods and Practices in Psychological Science*, *1*(4), 443–490. https://doi.org/10.1177/2515245918810225

Levelt Committee, Noort Committee, Drenth Committee (2012). Flawed science: The fraudulent research practices of social psychologist Diederik Stapel. https://poolux.psychopool.tu- dresden.de/mdcfiles/gwp/Reale%20F%C3%A4lle/Stapel%20-%20Final%20Report.pdf

Lilienfeld, S. O. (2012). Public skepticism of psychology: why many people perceive the study of human behavior as unscientific. *American Psychologist*, 67(2), 111.

Lynch Jr, J. G., Bradlow, E. T., Huber, J. C., & Lehmann, D. R. (2015). Reflections on the replication corner: In praise of conceptual replications. *International Journal of Research in Marketing*, 32(4), 333-342. https://doi.org/10.1016/j.ijresmar.2015.09.006

Maxwell, S. E., Lau, M. Y., & Howard, G. S. (2015). Is psychology suffering from a replication crisis? What does "failure to replicate" really mean?. *American Psychologist*, *70*(6), 487. https://doi.org/10.1037/a0039400

Mellers, B., Hertwig, R., & Kahneman, D. (2001). Do Frequency Representations Eliminate Conjunction Effects? An Exercise in Adversarial Collaboration. *Psychological Science*, 12(4), 269–275. https://doi.org/10.1111/1467-9280.00350

Molden, D. C. (2014). Understanding priming effects in social psychology: What is "social priming" and how does it occur?. *Social Cognition*, 32(Supplement), 1-11. https://doi.org/10.1521/soco.2014.32.supp.1

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*(6), 615-631. https://doi.org/10.1177/1745691612459058

Nelson, M. S., Wooditch, A., & Dario, L. M. (2015). Sample size, effect size, and statistical power: A replication study of Weisburd's paradox. *Journal of Experimental Criminology*, 11(1), 141-163. doi:http://dx.doi.org.ezproxy.rit.edu/10.1007/s11292-014-9212-9

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251), aac4716.

Pfeiffer, A, Brantl, V, Herz, A, Emrich, HM (1986). Psychotomimesis mediated by kappa opiate receptors. *Science.* 233, 774–6. DOI: 10.1126/science.3016896

Ranehill, E., Dreber, A., Johannesson, M., Leiberg, S., Sul, S., & Weber, R. A. (2015). Assessing the robustness of power posing: No effect on hormones and risk tolerance in a large sample of men and women. Psychological Science, 26(5), 653-656. doi:http://dx.doi.org.ezproxy.rit.edu/10.1177/0956797614553946

Retraction notice. (2014). *Journal of Vibration and Control*, *20*(10), 1601–1604. https://doi.org/10.1177/1077546314541924

Sagarin, B. J., Martin, A. L., Coutinho, S. A., Edlund, J. E., Patel, L., Skowronski, J. J., & Zengel, B. (2012). Sex differences in jealousy: A meta-analytic examination. *Evolution and Human Behavior*, 33(6), 595-614. https://doi.org/10.1016/j.evolhumbehav.2012.02.006

Scherer, C. R., Akers, E. G., & Kolbe, K. L. (2013). Bisexuals and the sex differences in jealousy hypothesis. *Journal of Social and Personal Relationships*, 30, 1064-1071. https://doi.org/10.1177/0265407513481446

Schnall, S., Benton, J., & Harvey, S. (2008). With a clean conscience: Cleanliness reduces the severity of moral judgments. *Psychological Science*, 19(12), 1219-1222. https://doi.org/10.1111/j.1467-9280.2008.02227.x

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*(2), 90-100. https://doi.org/10.1037/a0015108

Servick, K. (2018). Cornell nutrition scientist resigns after retractions and research misconduct finding. *Science*. https://www.sciencemag.org/news/2018/09/cornell-nutrition-scientist-resigns-after-retractions-and-research-misconduct-finding

Shackelford, T. K., LeBlanc, G. J., & Drass, E. (2000). Emotional reactions to infidelity. *Cognition and Emotion*, 14, 643-659. https://doi.org/10.1080/02699930050117657

Simons, D. J. (2014). The value of direct replication. *Perspectives on Psychological Science*, *9*(1), 76-80. https://doi.org/10.1177/1745691613514755

Simmons, J. P., & Simonsohn, U. (2017). Power posing: P-curving the evidence. *Psychological Science*. 28(5), 687-693. https://doi.org/10.1177/0956797616658563

Soderberg, C. K., & Errington, T. M. (2019). 14 Replications and the Social and. *Advanced Research Methods for the Social and Behavioral Sciences*, 229.

Srivastava, S. (Septmeber 27, 2012). A Pottery Barn rule for scientific journals. Retrieved from: https://thehardestscience.com/2012/09/27/a-pottery-barn-rule-for-scientific-journals/.

Stroebe, W., & Strack, F. (2014). The alleged crisis and the illusion of exact replication. *Perspectives on Psychological Science*, *9*(1), 59-71. https://doi.org/10.1177/1745691613514450

The Systemizing Confidence in Open Research and Evidence Program (2019). Retrieved from: http://cos.io/score/

Wagge, J. R., Brandt, M. J., Lazarevic, L. B., Legate, N., Christopherson, C., Wiggins, B., & Grahe, J. E. (2019). Publishing research with undergraduate students via replication work: The collaborative replications and education project. *Frontiers in Psychology*, *10*, 247. https://doi.org/10.3389/fpsyg.2019.00247

Wissenschaft im Dialog/Kantar Emnid (2018) Science Barometer 2018. Berlin. Available at: https://www.wissenschaft-im-dialog.de/en/our-projects/science-barometer/science-barometer-2018/

Yarkoni, T. (2019, November 22). The Generalizability Crisis. https://doi.org/10.31234/osf.io/jqw35

Yong, E. (September 2012). A failed replication draws a scathing personal attack from psychology professor. Retrieved from https://www.nationalgeographic.com/science/phenomena/2012/03/10/failed-replication-bargh-psychology-study-doyen/

Zwaan, R. A., Etz, A., Lucas, R. E., & Donnellan, M. B. (2018). Making replication mainstream. *Behavioral and Brain Sciences*, *41*. DOI: https://doi.org/10.1017/S0140525X17001972

Table One

Ways to incentivize replications by responsible party

Responsible party	Approach	Details	Relevant citations
Journals	Formally support replications	Journals should formally support and encourage replications as the publication of replications does not adversely impact the impact of the journal or articles.	Drotar, 2010; Edlund, 2016
	Adopt “Pottery Barn Rule” (you break it, you buy it/remake it)	If journals find research impactful enough to publish initially, any replications should automatically be considered impactful in that journal.	Chambers, 2018 Srivastava, 2012
Funder	Adopt replication policies	Funders should require some level of replication in their funding portfolio.	The Systemizing Confidence in Open Research and Evidence Program; http://cos.io/score/
	Require publication of null findings in open repositories	Although significant findings are frequently published, many null effects linger in the file drawer. Funders can require brief reports be submitted to open sourced venues.	NSF 15-52
University	Explicitly value replications	Often, hiring and tenure committees explicitly value novelty in findings; these groups should move to explicitly	Nosek, Spies, & Motyl, 2012

		valuing high quality replications as much as novel findings.	
	Incorporate replications into undergraduate and graduate curricula	A way to demonstrate the value of replications is to teach young scientists the value early in their training. Instructors should incorporate engaging in replications into the curriculum.	Cuccolo et al., 2020; Grahe et al., 2012; Wagge et al 2019

Table Two

Approaches To Replication

Type of Approach	Description	Strengths	Key Examples
Multi-lab Studies on General (or Many) Effects	A series of studies are selected for replication in a integrated fashion (where many studies are investigated at the same time). The rational for study selection can vary based on the project.	This approach tends to be very impactful (e.g., well-cited) and the widely distributed research labs allows for confounds to be ruled out. This approach also tends to be very well well-powered	*ManyLabs* (Klein et al., 2014; Klein et al., 2018; Ebersole et al., 2016). *Collaborative Replications and Education Project* (CREP; Wagge et al., 2019)
Multi-lab Studies on a Specific Effect	This approach focuses on answering a specific question (only one question or a very closely related set of questions) that cannot be answered by one individual laboratory.	This approach tends to be well-powered and allows for increased generalizability due to the increased diversity of samples). This approach also allows for investigation of moderators and/or mediators that could not be investigated by a single laboratory.	*Network for International Collaborative Exchange* (NICE; Cuccolo et al., 2020) *Sex Differences in Jealousy: The (Lack of) Influence of Researcher Theoretical Perspective*. Edlund et al., (2018)
Adversarial Collaborations	In this approach, two teams of researchers who represent different theoretical perspectives agree to work together on research in order to	In this approach, both parties gain knowledge about the theory under study and the field as a whole gains a fuller/richer	*Conjunction effects*. Mellers et al. (2001) *Minimal intergroup discrimination effect.* Kerr et al. (2018)

	try and resolve an open research question.	understanding of the effect due to the different approaches considered	
Stand-Alone Replications	In this approach a single researcher or research lab attempts to replicate a particular effect. This approach has significant variability in how direct the replications are and how large the scope is (additionally this approach can be adapted to a conceptual replication).	This approach is achievable for many labs as it tends not to need significant coordination or resources and as a result it is perceived as the easiest of the listed options.	*Enclothed Cognition and Stroop Effect.* Burns et al., 2019. *Weisburd's paradox.* Nelson et al., 2015