Replication in Social Psychology

|  |
| --- |
| Jordan Mark Barbone1 |
| 1 West Chester University of Pennsylvania University |
|  |

# Author note

Correspondence concerning this article should be addressed to Jordan Mark Barbone, Postal address. E-mail: [jb769904@wcupa.edu](mailto:jb769904@wcupa.edu)

Abstract

One of the most important issues for those in Psychology, especially Social Psychology, is to reason with the continuing pressures stemming from issues of replication. More research is showing a lack of reliability in research findings, challenging long-held studies, and exposing concerns with adequately conducted research and possible fraud. Direct replication is rare in Psychology, not sought after by publishers, and often lack other incentives. An overview of the “replication crisis” will be presented and two specific areas in social psychology will be presented and discussed in relation to these themes.

*Keywords:* social psychology, replication, false-positives, collaboration

*Word count:* X

Replication in Social Psychology

In the 17th century Robert Boyle’s reported success in observing anomalous suspension of water with his sophisticated air pump was put to question. These findings were difficult to replicate due to the instruments cost and complexity. Another scientist, Christiaan Huygens was able to demonstrate a similiar effect with his own air pump but was met with the same reaction. Thus a debate over validating claims and differentiating between “thought experiments” and actual experiments was birthed. The Royal Society was not satisfied until in 1663 when Huygens was invited to England and able to successfully replicate this phenomenon (Shapin, 1984).

What came from this skeptic event was a growing sense of the necessity of replicability and establishing facts that could be observed by others. Today, to publish in well-respected journals, researchers must describe the methods of their study or experiment well-enough for the reader to be able to successfully implement the same methodologies and assumingly reach similar conclusions. Yet, these might only be read and criticized with no follow-through of actual testing. In this way, the review of an article more akin to the thought experiments Boyle discussed than to the effort of replicating the events. Replicability is easy to stress through the words of the others, such as Braude (Braude, 2002) who tells us that “only experiments whose results can be repeated are considered genuine and reliable” and that this can be used as a “demarcation criterion between science and non-science”. Hype has been generated around a “replication crisis” with reports of null findings and failures to replicatea – thus asserting a serious threat the the integretity and character of psychological science.

# What is the “replication crisis”?

Sanja Srivastava maintains a blog, “The Hardest Science” on psychology with insights he holds as a professor and researcher in social psychology. His blog is named in part to retaliate against the insistance that science exists on a continuum from the soft sciences, such as psychology and sociology, to the hard sciences, like physics (Srivastava, 2009). He argues that each field is equal to the other scientifically as they all seek to answer inquries through applying logic and reasoning to evidence. However, the issues tackled are clearly different. Psychology tends to focus on understanding complex systems and tries to find patterns and reason to behavior. In this way, the issues psychology faces are more difficult to answer clearly. As such, we may want to discontinue the description of psychology as a “soft science” and adopt the name of his blog, “the hardest science”. This will be an important idea to maintain as issues with psychological replication are put forth and when attention is focused on psychology; or specically social psychology here; and while understanding that these problems are not just unique to us but may span across the whole of science (Ioannidis, 2005; Sterne & Smith, 2001).

Researchers understand the scientific importance of conducting replications in providing to provide contuing support for a hypothesis. This is cheifly served by direct replications but other functionalities ought to be noted, such that demonstrating the same effect can be found with virtually the same conditions provides further evidence against random chance and sampling errors. Conceptual replications, which should logically occur afterwards, would verify the generalizability of a hypothesis with other possible measures, means, or populations. And in what is hoped to be a smaller issue but one none-the-less, replications can help control for fraud – better accomplished when completed by those independent from the original research team. Although these functionalities seem critical to science, the rate of replications in psychology speaks a different story.

One estimate places the rate of replication in 100 psychology journals at 1.07% of publications (Makel, Plucker, & Hegarty, 2012), with rates of reported replication steadily increasing towards 2.5% in the 2010s. It is easy to consider why this is the case. Conducting a replication can be a tedious, time consuming and resource draining process. The replicating researchers may have to go to additional lengths to understand the methodology of the original researchers to ensure a set-up as similiar to the original as possible. Even if the replication is a success, there is a general consensus that journals favor novelty and positive finding.

Makel and colleagues (2012) also note that the median number of citations for replication articles was estimated at 17 which, unsurprisingly, was lower than the median rate for the original study at 64.5. But 17 citations for an article is still a decent accomplishment; only 3 of the 100 journals examined had 5 year impact factors above 17. This suggests that replications articles, although they may not make frequent appearances, are generally well-recieved by researchers and possibly sought out. Publishing companies may want to focus on providing new research findings but metrics like these may providing a convincing argument that that accepting replication articles might contribute to higher impact factors and attention to journals. If the journal published the original article, this might almost represent a sort of “double-dipping” of references.

Replication in psychology is a rare yet rewarded occurence. However, it is worthwhile to really delve into the necessity of replications. Afterall, calls and requests for greater replication may require large, consuming endeavors. Psychology has been functioning with a small amount of replications for quite some time, so the question can be posed, “do we really need more replications?” or “is this replication ‘crisis’ even that severe?”.

Pashler and Harris (2012) examined three arguments concerning the magnitude of the replication crisis. The first argument centers around the standard 5% likelihood of acceptance of a false null hypothesis and 80% likelihood of rejecting a false null hypothesis. These may seem like safe thresholds but the false positive rate of published articles is likely much higher than face expectancy. We can calculate the probability that a positive result is false finding () by dividing the proportion of false positives () by the sum of the proportion of false positives and the proportion of correct rejections (); i.e., .

With an assumption that the prior probability of a true effect is 10%, and given an level of .05, . Given a power level of .80, we would find that . We can then conclude that given the standard statistical assumptions, and a prior probability of 10%, that 36% of findings are likely to be false. This is a generous assumption that does not take into account other factors that could lead to biased false positives.

In his brazenly titled article, “Why most published research findings are false”, Ioannidis (2005) provides a more expansive review of the probability of false positive or exaggerated articles. His formualae include variations for researcher bias and the number of teams involved in a particular field. Using his methods, Iaonnidis identifies six correllaries to estimated false positive rates: studies with smaller sizes; studies with smaller effect sizes; the greater number of tests and less pre-selection of tests in a study; the greater the design or methodology flexibility; the greater the financial or interest in publishing the finding; and the more teams involved in a particular area. His last correllary seems at first counterintuitive without explanation. Ioannidis proposes that the “hotter” the area of research the more pressure researchers have to disseminate their “impressive” findings. Ioannidis comes to the conclusion that, given all these factors, the majority of research – across all fields – are likely to be false or exaggerated reports of effects.

The second argument Pashler and Harris (2012) tackle is the thought that despite a lack of *direct* replication, *conceptual* relications are more frequent and test not just the validity of the original research but the generalizability. The authors echo concerns from Ioannidis (2005) that published conceptual replications represent the favored “interesting” findings. If conceptual replications did provide a more rigourous testing of a hypothesis it might reason that the rate of failure here would be higher than it is. Of the estimate 1.07% of publications; direct replications constituting only 14% of these (Makel et al., 2012). The publication rate for failed direct replications (14.6%) is nearly twice that of failed conceptual replications (7.5%). This could be result of the difficulty of estimating the likelihood that a false finding is due to the original hypothesis being incorrect, that the theory does not generalize to the specific difference, or that the replication was not designed or executed well enough (Earp & Trafimow, 2015).

Cohen (1994) provides insight that can be applied n parallel to this the conceptual replication through criticizing thresholds of null hypothesis significance testing (NHST) and outlines a few misconceptions and misuses of statistical analyses. Often the *null hypothesis* is interpreted as a definition of rather than the hypothesis which is to be nulled. These tests are not just to compare whether or not there exists an effect but that levels of effect ought ot be compared as well. A quick note Cohen makes is the absurdity of testing rater reliability: that is, testing whichever computeted statistic against the *null hypothesis* that there exists no reliability between or amongst raters – even with a small sample these results would look significant. This fundemental missuse of NHST may then lead researchers to believe that because they have successfully reject a null hypothesis – rather than nullified a previous hypothesis – that there theory must be true (Meehl, 1990). Greater understanding of null hypothesis testing may aid in the design terpretation of direct and coneptual replications.

Science, we like to think, is self-correcting. Pashler and Harris (2012) challenge this arguement. They note their quick Google Scholar search for replication failures showed that the median replication attempt delay was 4 years [^replication\_delay] with 10% of replications occuring longer than 10 years. This is could be taken as a good sign that researchers are quick to scrutinize research with their own attempts. Howeveer, Pasher and Harris contend that this may simply represent the “faddish” nature of psychological research and that research which has failed to replicate may have also failed to maintain the herd’s interest. Older, possibly less contemporarily interesting research, should also be provided the service of scrutiny and rexamination.

[^replication\_delay] I attempted a similiar search by looking at the first 15 articles to which I had immediate access. The median and approximate mean was 6 years – although 11 of the replication articles were published in the 80s or 90s.

When 1,500 scientists were surveyed on comments and rates of replications, rouhgly half of those in psychology (approximately 54 respondents) reported having failed to reproduce their own work or someon else’s work (Baker, 2016). Compared to the whole, approximately 55% of respondents failed to replicate their own work and more than 70% failed to replicate anothers, and only 16% of these respondents successfully published a failure to reproduce. Self-correction, through publication of failed replications, does not seem to be that great of an argument if these results are all lost to the dreaded file drawer – or, increasingly appropriate, forgotten flash drive.

Replications, although well-recieved by researchers, are a bit of a rarity in the literature. However, there also appears to be growing interest as the rates of replications are increasing and as groups are becoming interested in taking on this problem.

# Large-scale replication attempts

The task of reproducibility testing can be performed with single instances and thorough examination of specific results as they pertain to the original article. Possibly a more effective means of garnering more attention is large-scale replications attempts which either test a great many hypotheses or select hypotheses multiple times. With such a great endeavor, researchers are leveraging cooperation across multiple reseaerch sites and labs. Two examples of large-scale replication attempts are reported below.

## The Open Science Collaboration

The Open Science Collaboration (OSC) of the Center for Open Science (COS) has promoted interest in replication studies and called for setting standards for replication attempts (Open Science Collaboration, 2012). The OSC selected 100 contemporary studies across three journals in psychology and reported the results of their replications in what would become a publicized and hihgly cited article. Of these 100, 93 were originally reported with significant findings[[1]](#footnote-25), but only 36 of the replication findings reached the same conclusion. When separting by discipline, the rate of success for cognitive studies was greater than social (21/42 and 14/55, respectively[[2]](#footnote-26)). This demonstrates a much lower proportion of false positives than 36% of positive findings (Pashler & Harris, 2012) and provides some support for models suggesting that the bulk of published research is false (Ioannidis, 2005). Original studies that reported greater effect sizes and more significant findings were more likely to be replicated than those with smaller effects and less significance, in line with suggestions from Ioannidis (Ioannidis, 2005) and presumed correlates of false positives.

The COS has made clear possible issues with their findings. They acknowledge that there is no single standard for determining the success of a replication and thus report on several measures that may be taken into account (Open Science Collaboration, 2012, 2015). Their selection of articles was not entirely random either, and they acknowledge that there are inherently greater challenges in replicating some psychological research that may rely on a specific population or dependent on an historical event. Thee are also some criticisms regarding some changes made in the replications which might invalidate their status as direct replications (Gilbert, King, Pettigrew, & Wilson, 2016).

## The “Many Labs” Project

In a different method, the “Many Labs” project attempted 36 replications of 13 studies, with the main purpose to understand the variability of replication findings (Klein et al., 2014). This method, although more demanding than the OSC’s attempts, provides much a more comprehensive evaluation of findings reported by studies. Variations in effect sizes compared against original reports and *p* values from each replication were taken into consideration of replication success. Of the 13 articles tested, 10 of them showed clear indications of successful replications. The articles tested were chosen for their simplicity, ability to be complete both in person or online, and the general level of procesdural replicability – similiar to the criteria from the OSC.

Klein and colleagues (2014) also tested the variation of effecet sizes found across each lab by computing an intra-class correlation and using an ANOVA across each lab and between online or in-person testing. These showed an acceptable level of agreement (ICC = .75) across study sites and little impact of study location (all < .023). The only impact which may have been noted was from a better knowledge in the anchoring tests[[3]](#footnote-28).

# Case studies

Although these may represent singular instances of less than ideal research, attributing the concepts and ideas that have been mentioned may help provide an example of their effects. Two case studies have been provided below. The first of this will examine a well-known study and some of the limitations in understanding confounding variables. The second takes a look at a divisive field and the possibly of bias and poor data handling.

## Roasting the “Marshmallow test”

Walter Mischel’s now famous Stanford Marshmallow studies have generally found correlates between the delay of gratification to better life outcomes (Ayduk et al., 2000; Mischel, Shoda, & Rodriguez, 1989). A conceptual replication identified some criticism of the original studies (Watts, Duncan, & Quan, 2018). Namely, the original children tested were from a highly selected sample in the Stanford University community. The studies also failed to account for possible confounds such as mother’s education and home environment. The sample retained for longitudinal studies were also much lower than their original experiment (35-89 and over 600, respectively). In their replication, Tyler, Duncan, and Quan utilized data from the National Institude of Child Health and Human Development (NICHD) Study of Early Child Care and Youth Development (SECCYD). Data included information on a delayed gratification test and behavioral outcomes at age 15. An important note is that the children were all born of mothers who did not have have or complete a college education. Only this group was examined due to concerns with truncation of gratification delay measures for children born to mothers who completed college and because the population used is more appropriate and of interest to policy-makers of developmental interventions.

In their analysis, Watts and colleages were able to show a replication of achievement at age 15, although smaller than in the original studies. They were not able to find significant results with delayed gratification and behavioral measures. The significant results they found were also moderated by variables such as child background, home environment, and early cognitive skills – to the point of the interaction of delayed gratification being insignificant. Albeit this is one study that has dampened the relationship of self-control to later life outcomes, it is an important one. Displayed here is the concern of controlling for variables that may not have been of interested to the original authors. Although the initial and follow-up findings have show significance and the track record appears sound, revisiting the study itself highlights concerns. Replications like these are important for understanding how even ubiquitous findings that seem unanimous in the literature may require additional scrutiny.

## Violence in video-games

There exists a contentious forum between researchers in the field of media whether or not violent media – specifically video games – contributes to violent acts by individuals. A meta-analysis (Anderson & Bushman, 2001) of violent videogames concluded that violent videogames increase aggression, physiological arousal, and aggression-releated thoughts and feelings – with small but positive lationships. Yet, years later, another meta-analysis (Ferguson, 2007) found that videogames were not linked to aggression after controlling for publication bias. This second meta-analysis also reported on improvement on visuospatial cognition. Here, too, there was an issue with publication bias – but even when adjusting for this, the improvement effect was still significant.

Now, with conflicting meta-analyses, researchers may be stuck *inter canem et lupum* when reasoning out how to interpret these findings. There also exists another bias from either side: the first publication was presented in part by Brad Bushman – a researcher known in this particular niche niche for finding ways media and video games influence aggressive behaviors. The second was published by Christopher John Ferguson, a researcher who quite frequently challenges these views.

Along with these claims, this area of social psychology has been subjected to a complicated scandal centering around Brad Bushman and a doctoral student of his, Jodie Whitaker. In April 2013, Whitaker and Bushman published their article, *“Boom, Headshot!”: Effect of video game play and controller type on firing aim and accuracy* online in the journal *Communication Research* (2012). A data request made by concerns over appropriateness of statistical analyses initiated a long process of correspondences and a research investigation[[4]](#footnote-32). Four years later, the article was retracted (“Retraction notice,” 2017) and Jodie Whitaker’s PhD was revoked by Ohio State University. The senior author was dismissed of all allegations and was able to publish a replication of this study shortly after (Bushman, 2018). Not all original effects were replicated, includiing the main findings of controller type. Further, reported in the replication were small effects, significance levels – in most cases – just under .05, and an increase in sample size.

# Future directions in social psychology

It is clear that replication is an issue to be resolved. The COS has considered ways in which to examine the results of replications (Open Science Collaboration, 2012, 2015). In a similar vein, there need to be a further discussion on how to go about replications. Designing a replication study may not be as easy as exactly reproducing the same methods from the investigative study (Maxwell, Lau, & Howard, 2015), particularly around sample size and estimated powers, to the point where we may have to be just as critical of replication articles as the original investigative article.

Replicability is also not well addressed in psychology education and training. We are taught to think about how our study, our experiment is unique and what it offers. What’s our twist on the topic? How are we differentiating this work from the work of others? All in attempts to find your work in a reputable academic journal. The “publish or perish” attitudes that may be conveyed at some instutitions would lead to prioritization of new ideas than the confirmation of new and current ones.

Begeley and Iannidis (2015) have currated a list of ways in which we can help correctfor the lack of replications and the overall goodness of science. Among these are emphasizing greater statistical and experimental methodology knowledge; providing open access to data for examination and analytic replication; using more meta-analytic techniques for establishing general findings in areas of research; lobbying journals to solict replication bids; and pushing institutional responsibiliy, possibly to the point of requiring some level of open access or replicability. A suggested cultural change is to consider, more greatly, the quality of research and judge academics and other researchers on their reproducibily, openness and sharing.

# Concluding remarks

The lack of direct replication should undoubtly be an issue in psychology – especially in the social realm – but that that does not mean that all past work is questionable. Making the assumption that older published works are wrong would create even more issues with meta-analytic methods (Makel et al., 2012). As mentioned earlier, *indirect* and *conceptual* replications occur often. For now, these will have to do to demonstrate that phenomonon of past research is sound and replicable.

We should leave with a few notes of caution and tips. Research reproduced outside the original laboratory or collaborators aids more in demonstrating reproducibiliy than continuous articles from the same institution. Although most published replications have at least one of the original authors (Makel et al., 2012) Failure to replicate does not mean that the original effect is invalid. As the COS has suggested, the issue may even be with the attempt the replicate and methodological differences that were overlooked or unnoticed (Open Science Collaboration, 2012). There may exists some concepts and theories that are downright wrong and invalid but bad research continues to push these to publishers. Inveresely, a good idea may have gone unnoticed because the first experiment was not successful. But we ought not to disqualify any large portion of psychology simply due to concerns about reproducibility. Early findings may be a little grim, but these have tended towards testing recent research rather than anything foundational.

Two schools of thought, extreme in opposition, can be called upon when reacting to the overall threat of the “replication crisis”. To take the pessimistic meta-inductionism approach would be to begin throwing out research on the grounds that the previous works have been falsified, disproved, or *un-proven*. It would take that if studies and theories, which were one or still are held as truth are incorrect, why should be not dismiss the bulk of what we know? The epistemic optimist would asser that through rigourous scientific process, what we know now is true, or the approximately true. Yes, some long-held concepts will be dismantled and others held in a possibly state of limbo of acceptance, but we cannot simply disregard everything and start at square one again.

# References

Anderson, C. A., & Bushman, B. J. (2001). Effects of violent video games on aggressive behavior, aggressive cognition, aggressive affect, physiological arousal, and prosocial behavior: A meta-analytic review of the scientific literature. *Psychological Science*, *12*(5), 353–359.

Ayduk, O., Mendoza-Denton, R., Mischel, W., Downey, G., Peake, P. K., & Rodriguez, M. (2000). Regulating the interpersonal self: Strategic self-regulation for coping with rejection sensitivity. *Journal of Personality and Social Psychology*, *79*(5), 776.

Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. *Nature News*, *533*(7604), 452.

Begley, C. G., & Ioannidis, J. P. (2015). Reproducibility in science: Improving the standard for basic and preclinical research. *Circulation Research*, *116*(1), 116–126.

Braude, S. E. (2002). *ESP and psychokinesis: A philosophical examination*. Universal-Publishers.

Bushman, B. J. (2018). “Boom, headshot!”: Violent first-person shooter (fps) video games that reward headshots train individuals to aim for the head when shooting a realistic firearm. *Aggressive Behavior*.

Cohen, J. (1994). The earth is round (p<. 05), *49*(12), 997–1003.

Earp, B. D., & Trafimow, D. (2015). Replication, falsification, and the crisis of confidence in social psychology. *Frontiers in Psychology*, *6*, 621.

Ferguson, C. J. (2007). The good, the bad and the ugly: A meta-analytic review of positive and negative effects of violent video games. *Psychiatric Quarterly*, *78*(4), 309–316.

Gilbert, D. T., King, G., Pettigrew, S., & Wilson, T. D. (2016). Comment on “estimating the reproducibility of psychological science”. *Science*, *351*(6277), 1037–1037.

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS Medicine*, *2*(8), e124.

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Bahník, Š., Bernstein, M. J., … Nosek, B. A. (2014). Investigating variation in replicability a "many labs" replication project. *Social Psychology*, *45*(3), 142–152. <https://doi.org/10.1027/1864-9335/a000178>

Makel, M. C., Plucker, J. A., & Hegarty, B. (2012). Replications in psychology research: How often do they really occur? *Perspectives on Psychological Science*, *7*(6), 537–542.

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.

Meehl, P. E. (1990). Why summaries of research on psychological theories are often uninterpretable. *Psychological Reports*, *66*(1), 195–244.

Mischel, W., Shoda, Y., & Rodriguez, M. I. (1989). Delay of gratification in children. *Science*, *244*(4907), 933–938.

Open Science Collaboration. (2012). An open, large-scale, collaborative effort to estimate the reproducibility of psychological science. *Perspectives on Psychological Science*, *7*(6), 657–660. <https://doi.org/10.1177/1745691612462588>

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251). <https://doi.org/10.1126/science.aac4716>

Pashler, H., & Harris, C. R. (2012). Is the replicability crisis overblown? Three arguments examined. *Perspectives on Psychological Science*, *7*(6), 531–536.

Retraction notice. (2017). *Communication Research*, *44*(1), 144–144. <https://doi.org/10.1177/0093650217690274>

Shapin, S. (1984). Pump and circumstance: Robert boyle’s literary technology. *Social Studies of Science*, *14*(4), 481–520.

Srivastava, S. (2009). Making progress in the hardest sciecne. Retrieved from <https://thehardestscience.com/2009/03/14/making-progress-in-the-hardest-science/>

Sterne, J. A., & Smith, G. D. (2001). Sifting the evidence—what’s wrong with significance tests? *Physical Therapy*, *81*(8), 1464–1469.

Watts, T. W., Duncan, G. J., & Quan, H. (2018). Revisiting the marshmallow test: A conceptual replication investigating links between early delay of gratification and later outcomes. *Psychological Science*, *29*(7), 1159–1177. <https://doi.org/10.1177/0956797618761661>

Whitaker, J. L., & Bushman, B. J. (2012). RETRACTED: “Boom, headshot!”: Effect of video game play and controller type on firing aim and accuracy. *Communication Research*, *41*(7), 879–891. <https://doi.org/10.1177/0093650212446622>

1. Significant effects here being defing by a reported *p* of less than .05. [↑](#footnote-ref-25)
2. This reported only on the 97 original articles with reported significance and reported on the those with p < .05 in the original direction. [↑](#footnote-ref-26)
3. Included examples were the height of Mt Everest, distanced to New York City, and the population of Chicago. [↑](#footnote-ref-28)
4. A timeline of the events and correspondenced, as well as document of files and reports exchanged, has been curated by Malte Elson here: <http://www.malte-elson.com/headshot> [↑](#footnote-ref-32)