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, Collaboration

*Word count:* X

Replication in Social Psychology

In the 17th century Robert Boyle 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 complexity and expence. Another scientistChristiaan 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 was 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 to come to similar conclusions. Yet, these might only be read and criticized without following through with tests. 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”.

# What is the “replication crisis”?

Sanja Srivastava maintains a blog, “The Hardest Science” which on psychology with insites 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 as sciences 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 clearly answer. As such, we may want to discontinue the description of psychology as a “soft science” and adopt “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. These problems are not just unique to these fields but may span across the whole of science (Ioannidis, 2005; Sterne & Smith, 2001).

Arguably the easiest identifier of concern is the rate of replication studies in psychological research. One estimate places the rate of replication in scientific journals at 1.07% of publications (Makel, Plucker, & Hegarty, 2012). (Cohen, 1962)

Pashler and Harris (2012) examined three arguments concerning the magnitude of the replication crisis. The first argument centers around the standard 5% likelihood of acceptive a false null hypothesis and 80% likelihood of rejecting a false null hypothesis. These may seem like safe thresholds but the false positive rate is higher than at first glance. We can calculate the probability that a positive result is false () 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 . A rate of .36 for false positives findings, given these criteria, should amplify the concern.

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 articles. His formualae include variations for researcher bias and number of teams involved. 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, more the majority of research – across all fields – are likely to be false findings.

The second argument Pashler and Harris (2012) tackle is suggestiong that despite a lack of direct replication, conceptual relications are more frequent and test not just the validity of the original research but the generability, too. The authors echo concerns from Ioannidis (2005) that published conceptual replications represent “interesting” findings that publications favor. The rate of published replication in psychology is estimated at 1.07% of publications; direct replications constituting only 14% of these (Makel et al., 2012). The publication rate for failed direct replications is nearly twice that of failed conceptual replications (14.6% and 7.5%, respectively).

Cohen (1994) takes another approach to 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).

## False positive-psychology

## Estimating the reproducibility…

The Open Science Collaboration (OSC) of the Center for Open Science (COS) set out on an endeavor to promote the interest in replication studies and set standards for replications (Open Science Collaboration, 2012). The OSC selected 100 contemporary studies across three journals in psychology and and reported the results of their replications. 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 experiences was greater than social (21/42 and 14/55, respectively[[2]](#footnote-26)) Possibly coincidentally, this is concordant with the rough estimate that 36% of positive findings in research are false (Ioannidis, 2005; Pashler & Harris, 2012). 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.

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.

# Why replications don’t occur

## Why is this even a crisis?

Publishers tend to value in experiments and studies over replications. Confirmatory replications may not caught the attention of a journal because ‘*we already know that*’. Face value, repeated someone else’s work does not seem to contribute much to the scientific field. Yet, replicable “serves as a kind of demarcation criterion between science and non-science” (Braude, 2002, p. 33). Science is science because it is replicable.

## Data quality and ethics

No one is possibly more infamous for fabricating data than Andrew Wakefield. This well known example of a researcher fabricating data related to autism spectrum distorder and the MRR vaccine. Yet, incidences like these may be more frequent but more benign.

# Case studies

## Roasting the “Marshmallow test”

In the mid 1960’s, Walter Mischel and

## 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 by Brad Bushman and Craig Anderson – two reserachers known for their findings on the influence of violent media on violent behaviors and attitudes. 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.[[3]](#footnote-33) 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). An investigation into data quality was initiated by a data request due to concerns over appropriateness of statistical analyses. 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 the study (Bushman, 2018). However, not all effects were replicated: the main findings of controller type were not found to be significant, but others were. Further, the effects reported here are small, the significance levels are – in most cases – just under .05, and the number of participants have been increased.

# Concluding remarks

In oppositional contrast to the equating of social sciences as “soft sciences” Sanjay Srivastava at the University of Oregon likes to refer to Psychology as the “hardest science” has put together a semi-satirical class syllabus on Psychology research methods titled “Everything is F\*\*\*\*ed” (Srivastava, n.d.). Here Dr. Srivastava has curated a selection of articles to highlight shortcomings in psychological research that range from uninterpretable results, failures to replicate, and poor statistical analyses. Yet he still offers hope the everything is okay. Direct replication is an issue in psychology - especially in the social realm, but that that does not mean that all past work is questionable. 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. 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. There may exists some concepts and theories that are downright wrong and invalid but shotty research continues to push these to publishers. In the other sense, 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 because we have some concerns about reproducibility. Early findings have been a little grim, but these have tended towards testing recent research rather than anything foundational.

Not matter the results, reproducibility, openness, and transparancy will continue to grow from these expeditions. Already the ability to directly share research is made easy with online storage and sharing. Journals are requiring data in order to publish – some even with requirements that data must be publically available (“Academic journals data sharing requirements,” n.d.). Often it may seem, when reading the interpreations of researcher’s work, that hypotheses are strongly confirmed when the data suggest a weaker magnitude with more exceptions.

# References

Academic journals data sharing requirements. (n.d.). Johns Hopkins Libraries, urldate=2018-11-24. Retrieved from <http://dms.data.jhu.edu/data-management-resources/plan-research/funders-data-sharing-requirement/academic-journals-and-data-sharing-requirements/>

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.

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. (1962). The statistical power of abnormal-social psychological research: A review. *The Journal of Abnormal and Social Psychology*, *65*(3), 145.

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

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.

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

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.

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

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/>

Srivastava, S. (n.d.). Everything is fucked: The syllabus. Retrieved from <https://thehardestscience.com/2016/08/11/everything-is-fucked-the-syllabus/>

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

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. 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-33)