Replication Research

David M. Condon, Northwestern University, david-condon@northwestern.edu
Eileen K. Graham, Northwestern University, eileen.graham@northwestern.edu
Daniel K. Mroczek, Northwestern University, daniel.mroczek@northwestern.edu

---

> "Non-reproducible single occurrences are of no significance to science."
> — Popper, *The Logic of Scientific Discovery*, p. 86

When he first published this observation in 1934, Karl Popper's statement would not have met with much controversy among the scientific community. The ability to reproduce an effect — whether through natural observation or a carefully controlled experiment — is generally viewed by scientists as a prerequisite for declaring the effect's existence.[1] Replication research focuses on the extent to which previously observed effects can be reproduced. This type of research ranges in scope from identical reproduction of reported results using the exact same methods, equipment, conditions, or data to more nuanced explorations of the ways an effect is altered by the use of different methods (sensitivity analyses) or is conditional upon the presence of specific circumstances (generalizability).

In psychology (and several other scientific disciplines), the fundamental importance of replicability has been overshadowed in the last few decades for a variety of reasons. Though there have been some prominent and well-publicized examples of outright fraud, most scholars attribute the so-called replicability crisis to an increasingly competitive "publish or perish" culture in scientific research. There is good evidence that frequent publication of false-positive findings is a direct result of the use of publication metrics as the principal criterion for career advancement (Smaldino and McElreath, 2016). It seems that questionable research practices (e.g., poor research methods, low power and inadequate analytical rigor) are common consequences of the selection pressure for high research output.

The emphasis on quantity over quality has also increased the motivation, especially among junior researchers, to pursue novel lines of inquiry rather than conduct replication research. Replication research is often slower, more complicated, and less easily published than work that is perceived as novel by reviewers. This perception is not only driven by the belief that groundbreaking research is more interesting it is an implied feature of the peer-review model that reviewers are often senior researchers whose work may be threatened by failures to replicate. Early-career researchers who conduct replication research put themselves in the crosshairs of those with more credibility, respect, and power to derail their careers (Baumeister, 2016). While this phenomenon is rarely acknowledged in a public forum (see Fiske [2016] and Gelman [2016] for an exception), the results have been increasingly evident in recent years: there has been a disproportionate focus on novelty in psychology research at the expense of replicability and rigor.

It should be noted that the terms *reproducibility* and *replicability* are often conflated. There is some debate about the origin (Liberman, 2015) and importance of their distinction (Open Science Collaboration, 2012). Until recently, the need to distinguish between these two ideas was not particularly relevant as reanalyses of existing data were less common prior to the widespread adoption of computers and electronic communication. The relative ease of sharing data and statistical scripts (i.e., code or syntax) has made reproducibility analyses much more common. In fact, there are now software programs — for example, *StatCheck* (Epskamp and Nuijten, 2016) — that exist for the sole purpose of confirming the statistical integrity of published findings.

The distinction between these terms and related concepts is demonstrated in Figure 1. An effect is said to be reproducible if it can be re-created by an independent researcher (or research team) using the same data and analytic methods described with the original finding — this is the upper left quadrant of Figure 1. Evaluations of reproducibility are, in some sense, tests for accuracy in the reporting of results and can therefore be characterized dichotomously. Either the finding can be reproduced exactly as reported or not. While reproducibility may seem pedantic for findings that require only rudimentary statistics, the reproduction of findings that are based on complicated statistical procedures is an integral first step towards

[1] This is not strictly true across *all* scientific disciplines, but there are very few exceptions (particle physics is one example).

replicability, not only because it may identify errors in the reported results but also because it can guide the next steps in replication procedures.

Replicability, by contrast, relates to the *extent* to which an effect can be re-created under various other circumstances, as shown in the remaining three quadrants of Figure 1. The replicability of a finding tends to be dimensional rather than dichotomous. *Sensitivity analyses* are evaluations of replicability within the same dataset using different analytical techniques. Evaluations of *generalizability* explore the robustness of an effect across different samples using the same design and analytical methods. When different methods and data are used, it is possible to consider the extent to which an effect may be *fully generalizable* or *conceptually replicable* or related to other known effects within some broader nomological network.

To clarify the distinctions, consider an example with two research teams. The first team posits a hypothesis (whether formally stated or informal/exploratory); designs a procedure to evaluate this hypothesis; collects data (whether primary or secondary); makes numerous decisions about the most appropriate methods for cleaning, analyzing and interpreting the data; and then publishes the results. The second team wants to evaluate these findings further, perhaps because they seem counter-intuitive or because the researchers want to see if the findings can be extended. A logical first step would be to reproduce the original results, if possible. They would pursue this by obtaining the original data and following the procedures described in the publication — this may require direct input (and, hopefully, cooperation) from the first research team. If the results are reproducible, the second team may then pursue replication of the effect when using different design procedures and/or different analytical methods and/or different samples. Ideally, the second research team will publish the results of their attempts at reproducibility and replicability regardless of the outcome, for this will serve to inform the rest of the research community about the robustness of the effect under consideration.

Unfortunately, replication research has often failed to proceed smoothly. This is partially due to the reality that it has historically been difficult to find outlets for the results of replication research, especially for effects that fail to replicate. These circumstances have been highlighted by several events in recent years. As noted by Gelman (2016), concerns about null hypothesis significance testing (NHST), especially in the context of low power and convenience samples, have been voiced for some time, going back to critiques by Meehl (1978). Despite warnings that many results should not be trusted across many fields (e.g., Ioannidis, [2005]) and that replication studies should play a larger role in peer-reviewed outlets, certain articles were published that epitomized the problem of over-reliance on NHST combined with low power. These examples involved now discredited effects regarding extra-sensory perception, "power poses," and airplane rage, among many others. The most prominent examples were flashy findings that garnered many peer citations and widespread media attention. The problems noted above, such as quantity over quality and an overemphasis on novelty, combined with career-advancement pressures, brought the issues of replication to a boiling point in many fields (but especially psychology) by the mid 2010s.

In response, a small but growing cadre of psychologists have responded with increased efforts to bolster the transparency of research practices, encourage open collaboration, and create incentives for replication research. The general goal is to promote replication research as both a viable path for career advancement and a welcome component in scientific dialogue. These efforts have led to slow but noticeable changes in protocols among many researchers, publishers, and educational institutions. New organizations like the Open Science Framework (OSF, https://osf.io/) and the Society for the Improvement of Psychological Science (SIPS, http://improvingpsych.org/) have helped to facilitate and encourage these changes on several fronts.

There are also a number of new tools and resources to address circumscribed components of replication research. For data sharing, these include online data repositories like Dataverse (https://dataverse.harvard.edu/) and Figshare (https://figshare.com/). These tools are much more flexible than the traditional approach of one to one data sharing by electronic mail or delivery of physical copies of the data. Data repositories allow researchers to control the extent of access they wish to give other researchers — from fully open public domain access to much more restrictive alternatives — and other utilities like free, uninterrupted hosting of the data with a permanents digital object identifier (doi) and meta-data tracking of download activity.

Similarly, resources like the Open Science Framework allow researchers to register their data collection protocols and collaborate with greater transparency than is typically possible in the traditional publication model (Tullett, 2015). Transparency at each stage of the scientific process — not just in published findings —

is particularly critical for increasing replicability. Pre-registration of one's research aims and protocols before the data are analyzed, or even better, before the data are collected, reduces the incidence of "p-hacking" and positive findings that are mainly driven by researcher degrees of freedom (Luck and Gaspelin, 2016). Following the example of other disciplines (e.g., mathematics), OSF and SIPS have jointly developed a free pre-print service, PsyArXiv (https://osf.io/preprints/psyarxiv), where researchers can post unpublished drafts of manuscripts for reference and critique by others before they are submitted for peer-review. In a similar vein, both existing and new journals have migrated towards publishing models that are fully or partially "open-access." These approaches reduce the barriers to publication, which helps to alleviate several concerns about replicability, especially transparency and biases in the peer-review process.

Collectively, the availability of these relatively new resources for replication research bodes well for a more promising future than the recent past, and this is cause for considerable optimism. After all, the utility of science depends on replicable results.

Bibliography

Baumeister, R. F. (2016). Charting the future of social psychology on stormy seas: Winners, losers, and recommendations. *Journal of Experimental Social Psychology*, pages 1–6.

Epskamp, S. and Nuijten, M. B. (2016). *statcheck: Extract statistics from articles and recompute p values. Retrieved from http://CRAN.R-project.org/package=statcheck. (R package version 1.2.2)*.

Fiske, S. T. (2016). A call to change science's culture of shaming. *APS Observer*, 29(9):5.

Gelman, A. (2016). *What has happened down here is the winds have changed*. Retrieved February 18, 2017, http://andrewgelman.com/2016/09/21/what-has-happened-down-here-is-the-winds-have- changed/.

Ioannidis, J. P. A. (2005). Why Most Published Research Findings Are False. *PLoS Medicine*, 2(8):e124.

Liberman, M. (2015). *Replicability vs. reproducibility — or is it the other way around?* . Retrieved February 16, 2017, http://languagelog.ldc.upenn.edu/nll/?p=21956.

Luck, S. J. and Gaspelin, N. (2016). How to get statistically significant effects in any ERP experiment (and why you shouldn't). *Psychophysiology*, 54(1):146–157.

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology*, 46(4):806–834.

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.

Popper, K. R. (1959). *The logic of scientific discovery*. Basic Books, New York. (Original work published 1935).

Smaldino, P. E. and McElreath, R. (2016). The natural selection of bad science. *Royal Society Open Science*, 3(160384).

Tullett, A. M. (2015). In search of true things worth knowing: Considerations for a new article prototype. *Social and Personality Psychology Compass*, 9(4):188–201.

Further reading

Vazire, S. (2017). Quality uncertainty erodes trust in science. *Collabra: Psychology*, 3(1):1-5.

Abstract

The ability to reproduce an effect — whether through natural observation or a carefully controlled experiment — is generally viewed by scientists as a prerequisite for declaring the effect's existence. Replication research focuses on the extent to which previously observed effects can be reproduced. This type of research ranges in scope from identical reproduction of reported results using the exact same methods, equipment, conditions, or data to more nuanced explorations of the ways an effect is altered by the use of different methods (sensitivity analyses) or is conditional upon the presence of specific circumstances (generalizability). Though replication research has often failed to proceed smoothly in psychology for several important reasons, the availability of many new resources for replication research bodes well for a more promising future than the recent past.

Keywords

Reproducibility, Replicability, Open Science, Transparency

Citation

Condon, D. M., Graham, E., & Mroczek, D. K. (in press). Replication Research. For inclusion in the *Wiley- Blackwell Encyclopedia of Personality and Individual Differences.*

	Same Data	Different Data
Same Methods	Reproducibility	Generalizability
Different Methods	Sensitivity	Full Generalizability / Conceptual Replicability / Other Research

Figure 1: Distinguishing among important terms in replication research