Galetzka, C. (2019). A Short Introduction to the Reproducibility Debate in Psychology. *Journal of European Psychology Students, 10*(1), 16–25. Doi:10.5334/jeps.469

LITERATURE REVIEW

A Short Introduction to the Reproducibility Debate in Psychology

Cedric Galetzka[1]

Reproducibility is considered a defining feature of science: Trust in scientific discovery and progress are argued to depend on the ability to reproduce previous results. However, recent large-scale replication studies have spurred debate on the reproducibility of scientific findings and suggested that psychology is facing a crisis. The reproducibility of results has been related to current publication practices, which favor sensational and statistically significant results over replication studies. In turn, this skewed incentive system may encourage researchers to engage in questionable research practices, thereby distorting the psychological literature. Important findings and criticisms, as well as potential measures to improve the reproducibility of results, such as preregistered reports, replication studies, and open science, are discussed.

Keywords: reproducibility, replication, publication bias, questionable research practices, preregistration

Introduction

More than 50 years ago, Rosenthal and Jacobson (1968) published the first study on the Pygmalion effect, showing that teachers' positive expectations of students' performance result in an increase in students' IQ points compared to students for whom no additional expectations were formed. The Pygmalion effect inspired a wide range of research and has been frequently cited (5469 times according to Google Scholar on February 8, 2019; Calin-Jageman, 2018; Rosenthal & Jacobson, 1968). Yet, subsequent replication studies did not provide consistent confirmatory evidence, giving rise to debates on methodological design, flexibility in data analysis, and the value of replication studies in science (Calin-Jageman, 2018; Spitz, 1999). Indeed, a review of 35 years of research on the impact of teachers' expectations on intelligence indicated that the original effect is rather negligible, but that context is a powerful mediator (Jussim & Harber, 2005). Nowadays, large-scale replication studies have raised concerns regarding the reproducibility of research in psychology, and issues resembling the controversy over the Pygmalion effect have been widely discussed (Baker, 2016; Calin-Jageman, 2018; Fanelli, 2018; Open Science Collaboration, 2015). This review first introduces the concept of reproducibility and highlights findings that have led to a debate on reproducibility in psychology. In addition, likely causes of non-reproducible research and proposed solutions are discussed.

What Does Reproducibility Imply?

Reproducibility was first emphasized by Boyle (1660) who conducted elaborate air-pump experiments on how to generate a vacuum. Contemporary

[1] Department of Psychology, University of Potsdam, DEU

Corresponding author: Cedric Galetzka (cedric.galetzka@gmail.com)

philosophers contested the idea of empty space, leading Boyle (1660) to formulate detailed descriptions of his methods for others to replicate (Shapin & Schaffer, 1985): "I thought it necessary to deliver things circumstantially, that the Person I addressed them to … be able to repeat such unusual Experiments" (p.5). Following the same procedures ensures yielding similar results if relationships are deterministic, but is not guaranteed for systems involving randomness or unknown variables (Goodman, Fanelli, & Ioannidis, 2016; Zwaan, Etz, Lucas, & Donnellan, 2018). For instance, the number of action potentials in response to a certain stimulus is observed to vary from trial to trial due to uncontrollable cognitive and physiological processes and is therefore expressed probabilistically as a firing rate (Dayan & Abbott, 2005). In such cases, closely reproducing the procedures and design of an experiment (i.e., conducting a direct replication) may not reproduce the original findings, which led Goodman and colleagues (2016) to distinguish between methods and results reproducibility. In addition, the degree to which researchers agree on interpretations about research findings is referred to by Goodman and colleagues (2016) as inferential reproducibility; varied conclusions might arise after re-analyzing existing results using a different statistical approach or conducting replication studies that employ different methodological designs (i.e., conceptual replications; Crandall & Sherman, 2016; Goodman et al., 2016). Throughout the remainder of this review, these distinctions of reproducibility will be adopted.

Reporting the design of a study and the data analysis enables independent researchers to verify experimental findings. However, the degree of transparency needed to achieve methods reproducibility depends on the effect in question (Goodman et al., 2016); documenting word frequency, for instance, may be relevant for a study's manipulation but not for participant instructions. Further, reproducing a result via direct replications is argued to be necessary for a finding to be considered reliable (Popper, 1959). However, does a single replication that does not indicate an effect refute previous findings? Replications necessarily vary along some dimension (e.g., time), which makes it possible for discrepant results to be attributed to random error or unknown variables (Crandall & Sherman, 2016). For example, some effects may be mediated by experience in implementing a certain experiment (Collins, 1975; Earp & Trafimow, 2015). In this case, direct replications may only falsify a finding under specific circumstances (e.g., no expert knowledge available), and these insights can be used to gain a better understanding of an effect's implicit assumptions (Earp & Trafimow, 2015). Additionally, conceptual replications may uncover a theory's boundary conditions by answering whether similar inferences can be drawn across contexts and experimental designs (Crandall & Sherman, 2016). In an ideal scenario, conducting replication studies ensures science to remain self-correcting and theories to become increasingly comprehensive and is therefore considered a hallmark of scientific method (Braude, 1979; Ioannidis, 2012; Schmidt, 2009).

The Current Debate in Psychology

The reproducibility debate in psychology originated from reports that results cannot be replicated (Baker, 2016). For example, independent research teams of the Open Science Collaboration (2015) conducted direct replications of 100 studies in the areas of social and cognitive psychology. The authors reported statistically significant findings (p-values smaller than .05) in the same direction as the original studies in 36 percent of replications (Open Science Collaboration, 2015). Likewise, Camerer and colleagues (2018) examined psychological studies that were published in *Nature* and *Science* and found converging evidence for 61 percent of replicated studies while effect sizes were on average half as large as previously reported. Furthermore, a multilab effort by Wagenmakers and collaborators (2016) found null results for the facial feedback hypothesis by Strack, Martin, and Stepper (1988) across 17 independent replications. Similar findings were obtained for the ego depletion effect (Hagger et al., 2016) and the blocking effect in conditioning (Maes et al., 2016). These findings suggest that psychology faces a crisis since many effects are unreliable and may be due to random error (i.e., false-positive results; Ioannidis, 2005). In addition, these effects may have inspired new studies that rest on faulty assumptions, thereby leading to a profusion of resources and undermining trust in psychological research (Begley & Ioannidis, 2015).

This view is complicated by the fact that there is no generally accepted statistical definition for results reproducibility (Zwaan et al., 2018). For example, the Open Science Collaboration (2015) employed null hypothesis significance testing, effect size comparisons, and investigators' subjective estimates to examine whether a finding was replicated. Moreover, these methods might provide biased estimates of results reproducibility; a replication's effect size may be similarly strong as in the original study, but the effect might be considered statistically non-significant because of differences in sample size (Goodman et al., 2016). In addition, it is unclear how many statistically non-significant replications are needed to falsify a result (Earp & Trafimow, 2015), which is why Goodman and colleagues (2016) argued to pool the results of replication studies to analyze their cumulative strength of evidence. Relevant to this argument, the "Many Labs" project by Klein and colleagues (2014) reported the pooled results for 36 replications of 13 effects from the psychological literature and showed that 11 out of 13 effects were replicated. Besides, Gilbert, King, Pettigrew, and Wilson (2016) noted that if Klein and colleagues (2014) had conducted significance tests for each replication and reported their results like the Open Science Collaboration (2015), the rate of "successful" replications would have decreased from 85% to 34%. Therefore, critics contend that psychology faces a crisis and argue that certain decision criteria may underestimate results reproducibility (Fanelli, 2018).

Gilbert and colleagues (2016) also pointed toward methodological differences between direct replication studies of the Open Science Collaboration (2015) and the original experiments. For example, one study that examined attitudes toward African Americans was replicated with Italian and not American participants (Gilbert et al., 2016; Payne, Burkley, & Stokes, 2008). Comparing the replication studies that were endorsed by the original authors to those that were not, Gilbert and colleagues (2016) found endorsed replications to be more likely to reproduce an effect (59.7% vs. 15.4%). Conversely, a direct replication may not disprove a finding across contexts (Earp & Trafimow, 2015). For example, Iso-Ahola (2017) argued that several identical replication studies on the ego-depletion effect by Hagger and colleagues (2016) may not falsify the phenomenon, since ego-depletion has been observed across a variety of experimental designs. For these reasons, opponents of the crisis movement argue to shift the focus from results to inferential reproducibility in order to develop a more nuanced understanding of psychological effects (Drummond, 2018; Goodman et al., 2016; Zwaan et al., 2018).

Causes of Non-Reproducible Results

While the degree to which findings in psychology can be reproduced is debated, a large number of studies have investigated possible causes of non-reproducible results. The following section outlines mechanisms by which current publication practices may affect related scientific practice.

Publication Bias. In a recent *Nature* survey, more than 60% of researchers mentioned the burden to publish as one of the main reasons for non-reproducible research (Baker, 2016). The number of publications and citations are commonly used to assess the productivity of researchers, institutions, and countries, and impact career development, such as prospects of promotion or tenure (Garfield, 1955; Moher et al., 2018). Crucially, statistically significant findings are more likely to be published than null results (Begg & Berlin, 1988; Rosenthal, 1979). For example, Fanelli (2012) observed that the frequency of statistically significant findings in psychology increased from around 70 to more than 90 percent between 1990 and 2007. Conversely, this pattern may discourage researchers from attempting to publish statistically non-significant results, which has been termed the "file-drawer effect" (Rosenthal, 1979). In addition, publication bias is accompanied by a focus on novelty, which makes replication studies increasingly irrelevant for career advancement (Franco, Malhotra & Simonovits, 2014). This skewed incentive structure (i.e., rewarding novel and positive results over negative findings and replications) may be an obstacle to self-correction in psychology (Ioannidis, 2012; Young, Ioannidis, & Al-Ubaydli, 2008).

Study Design and Power. Incentivizing publication of statistically significant findings has repercussions on results reproducibility from a statistical perspective (Gelman & Carlin, 2014). Psychology experiments are often underpowered, that is, the probability of discovering true effects is decreased due to small effects and sample sizes (Bakker, van Dijk, & Wicherts, 2012; Cohen, 1990; Smaldino & McElreath, 2016). Low power means that effects that are discovered are less likely to be true (Ioannidis, 2005) while true effects are more likely to be inflated or found in the wrong direction (type M[agnitude] and type S[ign] error; Gelman & Carlin, 2014). The latter occurs with small-sample studies when a significance threshold is only passed by an exaggerated estimate of the true effect, which will be less likely to replicate in the future (Button et al., 2013). Goodman (2018) hypothesized that the study by the Open Science Collaboration (2015) hinted at the presence of type M errors in the psychological literature: Results of the original studies clustered below the .05 significance threshold, whereas the p-values of the high-powered replications were more broadly distributed and therefore potentially regressed to the true effect estimate (Goodman, 2018; Open Science Collaboration, 2015). Significantly, Smaldino and McElreath (2016) argued that running small-sample studies may be an adaptive response to publication bias since it allows to obtain statistically significant results at low costs.

Questionable Research Practices. Flexibility in data analysis can be utilized to increase the probability of detecting statistically significant results (Ioannidis, 2005). For example, Simmons, Nelson, and Simonsohn (2011) demonstrated that *researcher degrees of freedom*, or "p-hacking", during statistical analysis (e.g., excluding experimental conditions, adding participants, or including covariates) can raise the likelihood of finding p-values below .05 by more than 60%. In addition, selective reporting of results may have further implications for conclusions drawn from meta-analyses (Hutton & Williamson, 2000). These questionable research practices (QRPs) may inflate the number of false-positives and diminish results reproducibility, but are currently a gray area in scientific conduct (John, Loewenstein, & Prelec, 2012; Wicherts et al., 2016). Importantly, methods that produce statistically significant and non-reproducible results may appear to be appropriate choices rather than intentional attempts of distorting research findings in order to achieve publication (e.g., hindsight and confirmation bias; Munafò et al., 2017).

HARKing. Current publication practices emphasize consistency between results and hypotheses, a requirement that has been communicated to aspiring and early-career social scientists (Bem, 1987; Bishop, 2017; Kerr, 1998). This rule of thumb may lead researchers to hypothesize after results are known (HARKing), that is, presenting post hoc hypotheses as a priori to confirm findings from statistical analyses (Kerr, 1998). HARKing may introduce bias since post hoc hypotheses might be implausible; thorough literature research should narrow down on the most plausible hypotheses prior to data collection (Kerr, 1998). Therefore, HARKing promotes the development of narrow theories that are potentially based on false-positives (Kerr, 1998). Murphy and Aguinis (2017) observed that selective reporting of hypotheses can lead to considerable bias in the literature, especially when population parameters are subtle. Importantly, this discussion is not intended to discount exploratory hypothesis testing; progress in science is also based on hypotheses that gathered support after experimentation but seemed implausible at first (Franklin, 2005). However, if post hoc insights are not distinguished from a priori hypotheses, HARKing may lead to hypotheses that are less likely to be reproduced (Hollenbeck & Wright, 2017; Kerr, 1998).

Prevalence of Questionable Research Practices. Publication bias, the file-drawer effect, and low power are longstanding issues in psychology, raising questions about the prevalence of QRPs (Cohen, 1990; Greenwald, 1975; Rosenthal, 1979). John and colleagues (2012) questioned more than 2000 psychologists and estimated that over half had selectively reported dependent variables or results, rounded off p-values, collected or excluded data after testing for significance, and HARKed hypotheses at least once during their career. However, Fiedler and Schwarz (2016) reported that the actual frequencies of QRPs are much lower. In addition, the prevalence of QRPs as well as the number of article retractions per journal have not been observed to increase (Fanelli, 2018). Furthermore, Head, Holman, Lanfear, Kahn, and Jennions (2015) examined distributions of p-values (i.e., p-curves) across research fields and observed that psychology showed the highest amount of p-values that clustered around the .05 significance threshold. This pattern may indicate p-hacking since p-values are more likely to be close to .05 when, for example, data was collected until results reached statistical significance (Simonsohn, Nelson, & Simmons, 2014). However, the authors also obtained significant evidential value, indicating that the effects in question were likely to be true positives (Head et al., 2015). Thus, while QRPs appear to be present in psychological research, it is unclear to which degree the literature is biased (Fanelli, 2018; Murphy & Aguinis, 2017).

Measures to Improve the Reproducibility of Results

Publication bias and its possible repercussions on research practices are well documented, yet their precise impact on results reproducibility remain controversial (Ioannidis, 2005; Fanelli, 2018). The final section addresses proposed actions to enhance results reproducibility.

Preregistration. Researchers have argued for the adoption of preregistered reports in the publication process, which entails the inclusion of an additional stage of peer-review prior to data collection (Nosek & Lakens, 2014). This step is argued to ensure that provisional acceptance of a manuscript will depend on the overall quality and relevance of a study (Ioannidis et al., 2014). This would eliminate incentives for publishing statistically significant results and engaging in QRPs and p-hacking (Ioannidis et al., 2014; Nosek & Lakens, 2014). In addition, methodological flaws may be addressed before studies are conducted (Nosek & Lakens, 2014). Distinguishing post hoc insights from the initial predictions also reduces hindsight bias, overconfidence in one's own hypotheses, and HARKing (Nosek, Ebersole, DeHaven, & Mellor, 2018). Importantly, preregistration does not prohibit exploratory hypothesis testing but requires those results to be labeled as exploratory in the final publication (Nosek & Lakens, 2014). Publishing preregistered studies is increasingly being incentivized: Examples include the "Registered Replication Reports" program by the Association for Psychological Science and the introduction of preregistration badges by the Center of Open Science (Association for Psychological Science, 2018; Kidwell et al., 2016). A range of journals, including *Cortex* and all journals by The British Psychological Society, have opened up the possibility for researchers to preregister studies (Chambers, 2013; The British Psychological Society, 2018).

Replication Studies. To improve results and inferential reproducibility, Zwaan and colleagues (2018) made the case for replication studies to become a common part of the research process. For instance, Gernsbacher (2018) proposed the inclusion of incremental replications in each study, that is, additional experiments that replicate the main findings and incrementally test the conditions in which they occur. However, if replication studies are to become frequent tools, it has to become a norm that inconclusive results from publication studies are not perceived as damaging to researchers' careers (Pennycook, 2018). Pennycook (2018) further notes that failures to replicate findings do not necessarily imply the use of QRPs or flawed methodological designs. To correct this perception, replication studies could be implemented in the training of social scientists, requiring, for instance, the implementation of replication studies during PhD programs (Kochari & Ostarek, 2018; Zwaan et al., 2018). However, direct replications may be difficult to implement for certain types of research (e.g., longitudinal studies), and critics have argued replication studies to only provide minimal value as long as publication bias distorts the literature (Coyne, 2016; Schmidt & Oh, 2016; for a detailed discussion, see Zwaan et al., 2018).

Open Science. Scientists are also increasingly endorsing open science practices to enhance methods and results reproducibility (Nosek et al., 2015; Open Science Collaboration, 2015). This includes making hypotheses and data available to the public and other researchers to increase transparency in science (Munafò et al., 2017). Current publication practices have been counteracting transparency by incentivizing "good story scripts", thereby requiring researchers to leave out large amounts of information (Munafò et al., 2017). Open science principles are argued to improve verification of experimental analyses, heighten accountability among scientists, and eliminate the file-drawer effect (Lerner & Tetlock, 1999; Nosek & Lakens, 2014; Nosek et al., 2015). Recently, the Transparency and Openness Promotion (TOP) committee suggested a set of guidelines to transform journals' incentive structure (Nosek et al., 2015). These guidelines target transparency in design, analysis, and data, as well as standards for citations, preregistration, and replication studies (Nosek et al., 2015). The journal *Science*, for instance, is advocating to implement the TOP guidelines (McNutt, 2016). However, Drummond (2018) argued that open science is a laborious process which would require reviewers to spend increasing amounts of time on verifying the accuracy of results rather than focusing on the quality of research. In addition, the frequency of QRPs is still contested, and intentional fraud is not a frequent occurrence nor considered to bias the psychological literature extensively, leaving doubts about the benefits of complete transparency (Fanelli, 2018).

Alpha-Threshold Adjustment. Furthermore, a recent article by Benjamin and colleagues (2018) advocated adjusting the threshold for statistical significance from 0.05 to 0.005. While a lower alpha criterion does not guard against QRPs, p-hacking or HARKing, Benjamin and collaborators (2018) emphasized that a lower alpha threshold may reduce the amount of false-positive results. (Benjamin et al., 2018; Fanelli, Costas, & Ioannidis, 2017). However, McShane, Gal, Gelman, Robert, and Tackett (2017) argued that a more stringent alpha threshold might amplify biases that are associated with a dichotomous decision rule, such as overconfidence in results and disregard for meaningful contributions that do not meet the new threshold. Instead, the p-value should be relegated and treated continuously, shifting attention to factors that are usually viewed as secondary, such as quality and soundness of the design and hypotheses (McShane et al., 2017). McShane and colleagues (2017) also pointed towards a recent statement by the American Statistical Association, which indicated that statistical significance should not be considered synonymous with good scientific practice (Wasserstein & Lazar, 2016).

Alternative Statistics. These considerations have led to suggestions to replace p-values in favor of alternative statistical approaches, such as confidence intervals and effect sizes (Gardner & Altman, 1986; Sullivan & Feinn, 2012; Thompson, 2002). Cumming (2014) argued that confidence intervals, in contract to p-values, express a range of uncertainty and are therefore suited to support a cumulative scientific process. Combined with meta-analyses and replication studies, the use of confidence intervals might be able to yield increasingly precise estimates over time (Cumming, 2014). The journal *Psychological Science*, for example, actively encourages researchers to report confidence intervals and effect sizes in order to avoid dichotomous decision-making (Eich, 2013).

However, interpretations drawn from confidence intervals are contested (Morey, Hoekstra, Rouder, & Wagenmakers, 2016), and critics of frequentist statistics argue to embrace Bayesian methods (Dienes, 2011). For example, Wagenmakers and colleagues (2017) argued that a benefit of the Bayesian framework is that it allows quantifying evidence under the alternative hypothesis. Specifically, the Bayes factor expresses the plausibility of the data under the alternative compared to the null hypothesis as a likelihood ratio, which is argued to provide interpretations that appeal to an intuitive understanding of statistics (Dienes, 2011; Jeffreys, 1961). Moreover, the Bayesian framework encourages accumulation of evidence by taking the prior probability into account and allowing for continuous decision-making (Wagenmakers et al., 2017). Further, sequential testing is not considered a QRP, but researchers are able to observe how strongly the data shifts their beliefs as the sample size grows (Etz & Vandekerckhove, 2016; Rouder, 2014). However, Bayesian statistics does not resolve publication bias nor guard against the use of QRPs to inflate the importance of research findings (Banks et al., 2016; Savalei & Dunn, 2015). In fact, a common criticism is that the specification of a prior probability can introduce additional bias into the statistical analysis (Efron, 2013).

Conclusion

To conclude, the current debate on reproducibility in psychology originated from a variety of reports indicating that results cannot be directly replicated (Baker, 2016; Open Science Collaboration, 2015). Publication bias assumes a central role in the debate on reproducibility and is argued to cause a skewed incentive system that rewards publication over sound science, thereby encouraging QRPs and methodologically flawed studies that worsen results reproducibility (Greenwald, 1975; Simmons et al., 2011). As such, the view that psychology faces a crisis is increasingly common (Baker, 2016; Fanelli, 2018). However, there is no precise definition of results reproducibility, and concerns about methodological aspects of previous replication efforts have been raised (Gilbert et al., 2016; Goodman et al., 2016). Therefore, researchers have argued that the reproducibility crisis is exaggerated and that supporting the perception of a reproducibility crisis may be damaging to psychology (Fanelli, 2018; Fiedler & Schwarz, 2016).

Several possibilities to enhance methods, results, and inferential reproducibility have been proposed, including preregistered reports, replication studies, open science, alpha-threshold adjustment, and the use of alternative statistical approaches (Benjamin et al., 2018; Cumming, 2014; Dienes, 2011; Nosek & Lakens, 2014; Nosek et al., 2015; Zwaan et al., 2018). These options are widely discussed and while they may potentially resolve longstanding issues (e.g., publication bias), adopting some of these guidelines may be costly and shift the focus from conducting innovative research (Crandall & Sherman, 2016; Drummond, 2018; Nosek & Lakens, 2014). Choosing and implementing the appropriate methods will require a joint effort from all members of the psychological scientific community (Nosek et al., 2015).

Acknowledgements

I am grateful to Oliver Lindemann for providing valuable comments on an earlier version of this manuscript. In addition, I thank Mhairi Gador-Whyte for proofreading.

Conflicts of Interest

The author has no conflicts of interest to declare.

References

Association for Psychological Science. (2018). *Registered Replication Reports*. Retrieved from https://www.psychologicalscience.org/publications/replication

Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. *Nature News*, *533*(7604), 452–454. https://doi.org/10.1038/533452a

Bakker, M., van Dijk, A., & Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science*, *7*(6), 543–554. https://doi.org/10.1177/1745691612459060

Banks, G. C., O'Boyle Jr, E. H., Pollack, J. M., White, C. D., Batchelor, J. H., Whelpley, C. E., … Adkins, C. L. (2016). Questions about questionable research practices in the field of management: A guest commentary. *Journal of Management*, *42*(1), 5–20. https://doi.org/10.1177/0149206315619011

Begg, C. B., & Berlin, J. A. (1988). Publication bias: A problem in interpreting medical data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *151*(3), 419–463. https://doi.org/10.2307/2982993

Begley, C. G., & Ioannidis, J. P. (2015). Reproducibility in science: Improving the standard for basic and preclinical research. *Circulation Research*, *116*(1), 116–126. https://doi.org/10.1161/CIRCRESAHA.114.303819

Bem, D. J. (1987). Writing the empirical journal article. In Zanna, M. P. & Darley, J. M. (Eds.). *The compleat academic: A practical guide for the beginning social scientist*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E. J., Berk, R., … Cesarini, D. (2018). Redefine statistical significance. *Nature Human Behaviour*, *2*(1), 6–10. https:/doi.org/10.1038/s41562-017-0189-z

Bishop, D. V. M. (2017). *The why and how of reproducible science* [Video file]. Retrieved from http://neuroanatody.com/2017/11/oxford-reproducibility-lectures-dorothy-bishop/

Boyle, R. (1660). *New experiments physico-mechanicall, touching the spring of the air, and its effects*. Oxford: Printed by Henry Hall.

Braude, S. E. (1979). *ESP and psychokinesis: A philosophical examination*. Philadelphia, PA: Temple University Press.

Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*(5), 365–376. https://doi.org/10.1038/nrn3475

Calin-Jageman, R. J. (2018). *We've been here before: The replication crisis over the pygmalion effect* [Blog post]. Retrieved from https://thenewstatistics.com/itns/2018/04/03/weve-been-here-before-the-replication-crisis-over-the-pygmalion-effect/

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T. H., Huber, J., Johannesson, M., ... Altmejd, A. (2018). Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour*, *2*(9), 637–644. https://doi.org/10.1038/s41562-018-0399-z

Chambers, C. D. (2013). *Registered Reports: A new publishing initiative at Cortex* [Editorial]. *Cortex*, *49*(3), 609–610. https://doi.org/10.1016/j.cortex.2012.12.016

Cohen, J. (1990). Things I have learned (so far). *American Psychologist*, *45*(12), 1304–1312. https://doi.org/10.1037/0003-066X.45.12.1304

Collins, H. M. (1975). The seven sexes: A study in the sociology of a phenomenon, or the replication of experiments in physics. *Sociology*, *9*(2), 205–224. https://doi.org/10.1177/003803857500900202

Coyne, J. C. (2016). Replication initiatives will not salvage the trustworthiness of psychology. *BMC Psychology*, *4*(1), 28. https://doi.org/10.1186/s40359-016-0134-3

Crandall, C. S., & Sherman, J. W. (2016). On the scientific superiority of conceptual replications for scientific progress. Journal of Experimental *Social Psychology*, *66*, 93–99. https://doi.org/10.1016/j.jesp.2015.10.002

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*, *25*(1), 7–29. https://doi.org/10.1177/0956797613504966

Dayan, P., & Abbott, L. F. (2005). *Theoretical neuroscience: Computational and mathematical modeling of neural systems*. Cambridge, MA: MIT Press.

Dienes, Z. (2011). Bayesian versus orthodox statistics: Which side are you on? *Perspectives on Psychological Science*, *6*(3), 274–290. https://doi.org/10.1177/1745691611406920

Drummond, C. (2018). Reproducible research: A minority opinion. *Journal of Experimental & Theoretical Artificial Intelligence*, *30*(1), 1–11. https://doi.org/10.1080/0952813X.2017.1413140

Earp, B. D., & Trafimow, D. (2015). Replication, falsification, and the crisis of confidence in social psychology. *Frontiers in Psychology*, *6*, 621. https://doi.org/10.3389/fpsyg.2015.00621

Efron, B. (2013). Bayes' theorem in the 21st century. *Science*, *340*(6137), 1177–1178. https://doi.org/10.1126/science.1236536

Eich, E. (2013). Business not as usual [Editorial]. *Psychological Science*, *25*(1), 3–6. https://doi.org/10.1177/0956797613512465

Etz, A., & Vandekerckhove, J. (2016). A Bayesian perspective on the reproducibility project: Psychology. *PLOS ONE*, *11*(2), e0149794. https://doi.org/10.1371/journal.pone.0149794

Fanelli, D. (2012). Negative results are disappearing from most disciplines and countries. *Scientometrics*, *90*(3), 891–904. https://doi.org/10.1007/s11192-011-0494-7

Fanelli, D. (2018). Opinion: Is science really facing a reproducibility crisis, and do we need it to? *Proceedings of the National Academy of Sciences*, *115*(11), 2628–2631. https://doi.org/10.1073/pnas.1708272114

Fanelli, D., Costas, R., & Ioannidis, J. P. (2017). Meta-assessment of bias in science. *Proceedings of the National Academy of Sciences*, *114*(14), 3714–3719. https://doi.org/10.1073/pnas.1618569114

Fiedler, K. & Schwarz, N. (2016). Questionable research practices revisited. *Social Psychological and Personality Science*, *7*(1), 45–52. https://doi.org/10.1177/1948550615612150

Franco, A., Malhotra, N., & Simonovits, G. (2014). Publication bias in the social sciences: Unlocking the file drawer. *Science*, *345*(6203), 1502–1505. https://doi.org/10.1126/science.1255484

Franklin, L. R. (2005). Exploratory experiments. *Philosophy of Science*, *72*(5), 888–899. https://doi.org/10.1086/508117

Gardner, M. J., & Altman, D. G. (1986). Confidence intervals rather than p values: Estimation rather than hypothesis testing. *The BMJ*, *292*(6522), 746–750. https://doi.org/10.1136/bmj.292.6522.746

Garfield, E. (1955). Citation indexes for science. *Science*, *122*(3159), 108–111. https://doi.org/10.1126/science.122.3159.108

Gelman, A., & Carlin, J. (2014). Beyond power calculations: Assessing type S (sign) and type M (magnitude) errors. *Perspectives on Psychological Science*, *9*(6), 641–651. https://doi.org/10.1177/1745691614551642

Gernsbacher, M. A. (2018). Three ways to make replication mainstream. *Behavioral and Brain Sciences*, *41*, e129. https://doi.org/10.1017/S0140525X1800064X

Gilbert, D. T., King, G., Pettigrew, S., & Wilson, T. D. (2016). Comment on "Estimating the reproducibility of psychological science". *Science*, *351*(6277), 1037. https://doi.org/10.1126/science.aad7243

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

Goodman, S. N. (2018). *Science on trial* [Video file]. Retrieved from https://www.youtube.com/watch?v=YO9Cqn5gSVo

Goodman, S. N., Fanelli, D., & Ioannidis, J. P. (2016). What does research reproducibility mean? *Science Translational Medicine*, *8*(341). https://doi.org/10.1126/scitranslmed.aaf5027

Hagger, M. S., Chatzisarantis, N. L., Alberts, H., Anggono, C. O., Batailler, C., Birt, A. R., ... Calvillo, D. P. (2016). A multilab preregistered replication of the ego-depletion effect. *Perspectives on Psychological Science*, *11*(4), 546–573. https://doi.org/10.1177/1745691616652873

Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., & Jennions, M. D. (2015). The extent and consequences of p-hacking in science. *PLOS Biology*, *13*(3). https://doi.org/10.1371/journal.pbio.1002106

Hollenbeck, J. R., & Wright, P. M. (2017). Harking, sharking, and tharking: Making the case for post hoc analysis of scientific data. *Journal of Management*, *43*(1), 5–18. https://doi.org/10.1177/0149206316679487

Hutton, J. L., & Williamson P. R. (2000). Bias in meta-analysis due to outcome variable selection within studies. *Journal of the Royal Statistical Society*, *49*(3), 359–370. https://doi.org/10.1111/1467-9876.00197

Ioannidis, J. P. (2005). Why most published research findings are false. *PLOS Medicine*, *2*(8). https://doi.org/10.1371/journal.pmed.0020124

Ioannidis, J. P. (2012). Why science is not necessarily self-correcting. *Perspectives on Psychological Science*, *7*(6), 645–654. https://doi.org/10.1177/1745691612464056

Ioannidis, J. P., Munafò, M. R., Fusar-Poli, P., Nosek, B. A., & David, S. P. (2014). Publication and other reporting biases in cognitive sciences: Detection, prevalence, and prevention. *Trends in Cognitive Sciences*, *18*(5), 235–241. https://doi.org/10.1016/j.tics.2014.02.010

Iso-Ahola, S. E. (2017). Reproducibility in psychological science: When do psychological phenomena exist? *Frontiers in Psychology*, *8*, 879. https://doi.org/10.3389/fpsyg.2017.00879

Jeffreys, H. (1961). *Theory of probability* (3rd ed.). Oxford: Oxford University Press.

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, 524–532. https://doi.org/10.1177/0956797611430953

Jussim, L., & Harber, K. D. (2005). Teacher expectations and self-fulfilling prophecies: Knowns and unknowns, resolved and unresolved controversies. *Personality and Social Psychology Review*, *9*(2), 131–155. https://doi.org/10.1207/s15327957pspr0902_3

Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review*, *2*(3), 196–217. https://doi.org/10.1207/s15327957pspr0203_4

Kidwell, M. C., Lazarević, L. B., Baranski, E., Hardwicke, T. E., Piechowski, S., Falkenberg, L. S., ... Errington, T. M. (2016). Badges to acknowledge open practices: A simple, low-cost, effective method for increasing transparency. *PLOS Biology*, *14*(5). https://doi.org/10.1371/journal.pbio.1002456

Klein, R. A., Ratliff, K. A., Vianello, M., Adams Jr, R. B., Bahník, Š., Bernstein, M. J., ... Cemalcilar, Z. (2014). Investigating variation in replicability. *Social Psychology*, *45*, 142–152. https://doi.org/10.1027/1864-9335/a000178

Kochari, A. R., & Ostarek, M. (2018). Introducing a replication-first rule for Ph.D. projects. *Behavioral and Brain Sciences*, *41*. https://doi.org/10.31234/osf.io/6yv45

Lerner, J. S., & Tetlock, P. E. (1999). Accounting for the effects of accountability. *Psychological Bulletin*, *125*(2), 255–275. http://dx.doi.org/10.1037/0033-2909.125.2.255

Maes, E., Boddez, Y., Alfei, J. M., Krypotos, A. M., D'hooge, R., De Houwer, J., & Beckers, T. (2016). The elusive nature of the blocking effect: 15 failures to replicate. *Journal of Experimental Psychology: General*, *145*(9), e49–e71. http://dx.doi.org/10.1037/xge0000200

McNutt, M. (2016). Taking up TOP [Editorial]. *Science*, *352*(6290), 1147. https://doi.org/10.1126/science.aag2359

McShane, B. B., Gal, D., Gelman, A., Robert, C., & Tackett, J. L. (2017, April 10). *Abandon statistical significance* [arXiv preprint]. Retrieved from https://arxiv.org/abs/1709.07588

Moher, D., Naudet, F., Christea, I. A., Miedema, F., Ioannidis, J. P. A., & Goodman, S. N. (2018). Assessing scientists for hiring, promotion, and tenure. *PLOS Biology*, *16*(3). https://doi.org/10.1371/journal.pbio.2004089

Morey, R. D., Hoekstra, R., Rouder, J. N., Lee, M. D., & Wagenmakers, E. J. (2016). The fallacy of placing confidence in confidence intervals. *Psychonomic Bulletin & Review*, *23*(1), 103–123. https://doi.org/10.3758/s13423-015-0947-8

Munafò, M. R., Nosek, B. A., Bishop, D. V., Button, K. S., Chambers, C. D., du Sert, N. P., ... Ioannidis, J. P. (2017). A manifesto for reproducible science. *Nature Human Behaviour*, *1*. https://doi.org/10.1038/s41562-016-0021

Murphy, K. R., & Aguinis, H. (2017). HARKing: How badly can cherry-picking and question trolling produce bias in published results? *Journal of Business and Psychology*, 1–17. https://doi.org/10.1007/s10869-017-9524-7

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., ... Contestabile, M. (2015). Promoting an open research culture. *Science*, *348*(6242), 1422–1425. https://doi.org/10.1126/science.aab2374

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences*, *115*(11), 2600–2606. https:/doi.org/10.1073/pnas.1708274114

Nosek, B. A., & Lakens, D. (2014). Registered reports. *Social Psychology*, *45*(3), 137–141. http://dx.doi.org/10.1027/1864-9335/a000192

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

Payne, B. K., Burkley, M. A., & Stokes, M. B. (2008). Why do implicit and explicit attitude tests diverge? The role of structural fit. *Journal of Personality and Social Psychology*, *94*(1), 16–31. http://dx.doi.org/10.1037/0022-35-14.94.1.16

Pennycook, G. (2018). You are not your data. *Behavioral and Brain Sciences*, *41*. https://doi.org/10.31234/osf.io/92hmr

Popper, K. (1959). *The logic of scientific discovery*. Oxford, England: Basic Books.

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, *86*(3), 638–641. http://dx.doi.org/10.1037/0033-2909.86.3.638

Rosenthal, R., & Jacobson, L. (1968). Pygmalion in the classroom. *The Urban Review*, *3*(1), 16–20. https://doi.org/10.1007/BF02322211

Rouder, J. N. (2014). Optional stopping: No problem for Bayesians. *Psychonomic Bulletin & Review*, *21*(2), 301–308. https://doi.org/10.3758/s13423-014-0595-4

Savalei, V., & Dunn, E. (2015). Is the call to abandon p-values the red herring of the replicability crisis? *Frontiers in Psychology*, *6*, 245. https://doi.org/10.3389/fpsyg.2015.00245

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. http://dx.doi.org/10.1037/a0015108

Schmidt, F. L., & Oh, I. S. (2016). The crisis of confidence in research findings in psychology: Is lack of replication the real problem? Or is it something else? *Archives of Scientific Psychology*, *4*(1), 31–37. http://dx.doi.org/10.1037/arc0000029.supp

Shapin, S., & Schaffer, S. (1985). *Leviathan and the air-pump*. Princeton, NJ: Princeton University Press.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-curve: A key to the filedrawer. *Journal of Experimental Psychology: General*, *143*(2), 534–547. http://dx.doi.org/10.1037/a0033242

Smaldino, P. E., & McElreath, R. (2016). The natural selection of bad science. *Royal Society Open Science*, *3*(9). https://doi.org/10.1098/rsos.160384

Spitz, H. H. (1999). Beleaguered pygmalion: A history of the controversy over claims that teacher expectancy raises intelligence. *Intelligence*, *27*(3), 199–234.

https://doi.org/10.1016/S0160-2896(99)00026-4

Strack, F., Martin, L. L., & Stepper, S. (1988). Inhibiting and facilitating conditions of the human smile: A nonobtrusive test of the facial feedback hypothesis. *Journal of Personality and Social Psychology*, *54*(5), 768–777. http://dx.doi.org/10.1037/0022-3514.54.5.768

Sullivan, G. M., & Feinn, R. (2012). Using effect size—or why the p value is not enough. *Journal of Graduate Medical Education*, *4*(3), 279–282. https://doi.org/10.4300/JGME-D-12-00156.1

The British Psychological Society (2018). *We're offering registered reports across all eleven of our academic journals*. Retrieved from https://www.bps.org.uk/news-and-policy/were-offering-registered-reports-across-all-eleven-our-academic-journals

Thompson, B. (2002). What future quantitative social science research could look like: Confidence intervals for effect sizes. *Educational Researcher*, *31*(3), 25–32. https://doi.org/10.3102/0013189X031003025

Wagenmakers, E. J., Beek, T., Dijkhoff, L., Gronau, Q. F., Acosta, A., Adams Jr, R. B., ... Bulnes, L. C. (2016). Registered replication report: Strack, Martin, & Stepper (1988). *Perspectives on Psychological Science*, *11*(6), 917–928. https://doi.org/10.1177/1745691616674458

Wagenmakers, E. J., Verhagen, J., Ly, A., Matzke, D., Steingroever, H., Rouder, J. N., & Morey, R. D. (2017). The need for Bayesian hypothesis testing in psychological science. In S. O. Lilienfeld & I. D. Waldman (Eds.), *Psychological science under scrutiny: Recent challenges and proposed solutions* (pp.123–138). New York, NY: John Wiley & Sons Limited.

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's statement on p-values: Context, process and purpose. *The American Statistician*, *70*(2), 129–133. https://doi.org/10.1080/00031305.2016.1154108

Wicherts, J. M., Veldkamp, C. L., Augusteijn, H. E., Bakker, M., van Aert, R. C., & van Assen, M. A. (2016). Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid p-hacking. *Frontiers in Psychology*, *7*(1832). https://doi.org/10.3389/fpsyg.2016.01832

Young, N. S., Ioannidis, J. P. A., & Al-Ubaydli, O. (2008). Why current publication practices may distort science. *PLOS Medicine*, *5*(10), 1418–1422. https://doi.org/10.1371/journal.pmed.0050201

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