Check for updates

Commentary

Science Communication
1–13
© The Author(s) 2024
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/10755470241239947
journals.sagepub.com/home/scx
Sage

Forcing a Deterministic Frame on Probabilistic Phenomena: A Communication Blind Spot in Media Coverage of the "Replication Crisis"

Carol Ting[1] and Sander Greenland[2]

Abstract

The current controversy surrounding research replication in biomedical and psychosocial sciences often overlooks the uncertainties surrounding both the original and replication studies. Overemphasizing single attempts as definitive replication successes or failures, as exemplified by media coverage of the landmark Reproducibility Project: Psychology, fosters misleading dichotomies and erodes public trust. To avoid such unintended consequences, science communicators should more clearly articulate statistical variation and other uncertainty sources in replication, while emphasizing the cumulative nature of science in general and replication in particular.

Keywords

bias, confidence intervals, hypothesis tests, *p*-values, replication, reproducibility, statistical significance, trust in science, uncertainty

[1] University of Macau, Taipa, Macau SAR
[2] University of California, Los Angeles, USA

Corresponding Author:
Carol Ting, The Department of Communication, University of Macau, Rm. 2058, E21, Avenida da Universidade, Taipa, Macau SAR.
Emails: tingyf@gmail.com; tingyf@um.edu.com

Introduction

The issue of replicability or repeatability[1] is not new in biomedical and psychosocial sciences (Altman, 1994; Goodman, 1992; Mulkay, 1984), although in the past decade it has gained attention across scientific fields in hitherto unseen ways. Large-scale replication projects (Camerer et al., 2018; Errington et al., 2021; Open Science Collaboration, 2015a) have created widespread controversies over whether science is broken, raising concerns about undermining public confidence in science (Jamieson, 2018; Shiffrin et al., 2018).

In the Open Science Collaboration's (OSC) Replication Project: Psychology (RPP), 270 researchers across five continents attempted to replicate 100 studies published in three prominent psychology journals in 2008. Using five different criteria, less than half of the studies were classified as successfully replicated (Open Science Collaboration, 2015b). Among those five criteria, the 36% replication rate (based on whether the replications produced statistically significant results as the originals did) received the most attention, solidifying a crisis narrative that drove follow-up reform efforts across journals and disciplines (Peterson & Panofsky, 2023).

The OSC placed the RPP in the context of a distorted incentive system of academic research and publishing, which results in sub-optimal research practices (especially in statistical analyses) and a biased literature in need of major correction. While we largely agree with this diagnosis and appreciate the emphasis OSC put on cumulative evidence rather than single replication projects, we are also concerned with some unintended consequences emerging from the general discussion and media coverage in the wake of RPP. Despite the OSC's precautionary efforts and the extensive literature on common statistical misconceptions and misperceptions (reviewed for example in Amrhein et al., 2019; Greenland, 2017, 2019; Greenland et al., 2016; McShane et al., 2019, 2024; Wasserstein et al., 2019; Wasserstein & Lazar, 2016), discussions of the RPP still exhibit a lack of conceptual clarity. An example is the common error of interpreting the RPP results as suggesting that all or most of the original studies were false positives, without allowing for false-negative replications. To mitigate such misunderstanding and distrust it is crucial to distinguish two types of uncertainty in replication studies: method uncertainty and statistical uncertainty.

Method Uncertainty: The Duhem-Quine Problem

It can be helpful to think of replication along the lines of an analogy with cooking: if we follow a recipe precisely, we should be able to make the same dish over and over again and it should taste the same every time. The reality is more complex, however: we have all encountered situations where, despite our best efforts to follow a recipe meticulously, we still fall short of achieving the desired outcome. Maybe the ingredients were not stored in the right condition, maybe they were added in the wrong order, or maybe the oven temperature was a little too high . . . The point is that it is not always straightforward to tell whether we did something wrong, or if the recipe failed to mention something important.

Replicating scientific studies is, of course, more complex. As research in Science and Technology Studies (STS) has pointed out, in the early stage of research, researchers are often unaware of certain factors that are causally relevant to the phenomenon of interest (Collins, 1985). At this stage a researcher can demonstrate her claim with experiments; and if other experimenters follow her methods and obtain the same results, the case for her claim is strengthened.

When other researchers cannot achieve the same results by following the same methods, another possibility is often raised: the original finding may be contingent on some unspecified details of the original experiment. For example, the original finding might be an experimental artifact caused by an uncontrolled, causally relevant factor (a confounder). However, similar charges can be leveled at the replication study; in particular, replication failure may be due to an uncontrolled confounder. Furthermore, the actual effect may vary with study details that differ between the original and the replication study, such as the composition of the population of study participants.

When replication "fails" a controversy often ensues, generating a long list of potential explanations for the conflict that cannot be exhaustively tested. This is the well-known Duhem-Quine problem (Sismondo, 2010): Because all tests are simultaneous tests of the target hypothesis and numerous implementation details, no single test can conclusively prove or refute a hypothesis. We refer to the resulting uncertainty as *method uncertainty* since it arises from differences in research and analysis methods; these include the criteria for selection into the study, the extent of control of potentially confounding factors, and the implementation of statistical procedures, including choices that vary with or are unspecified by ordinary textbooks (such as how to code variables in a regression).

Statistical Uncertainty and the False-Negative Problem

Biomedical, psychological, and social sciences are "soft" insofar as they focus on phenomena whose regularities are amorphous and situational (in contrast to the universal, exact laws which dominate physical sciences). In doing so, they must confront another major source of research uncertainty: Living organisms are characterized by natural variation and complex feedback within and across organisms (Mayr, 1985; Nelson, 2016), which introduces sampling variation or "noise" that we model as *statistical uncertainty*. The tremendous variability among and within living beings and systems makes it crucial to account for this uncertainty when designing and analyzing studies of those entities.

Unfortunately, our desire for certainty leads to a tendency to approach statistics as if it could eliminate all uncertainty (Gelman, 2016; Gelman & Loken, 2014). Caught in this mindset, researchers identify findings with dichotomous labels (statistically significant vs. statistically insignificant) and treat hypothesis testing as if it could turn a continuous probabilistic phenomenon into a deterministic dichotomy, or at least one with high signal-to-noise ratio in well-controlled experiments (Goodman et al., 2016). This cognitive illusion is a major problem for soft sciences, where effect sizes are typically small, random variability is high, and nonrandom sources of variability—uncontrolled biases—must often be considered (Greenland, 2017). The result is a low ratio of true signal (effect) to random noise and bias, hence the low reliability of study results outside of a relatively few exceptionally large and expensive experimental studies.

Since tests with low statistical precision are less capable of detecting effects, extra caution is necessary when interpreting seemingly conflicting results. Let us illustrate the danger of ignoring low precision with the common example of underpowered studies. Even with 80% power for the actual effect (i.e., a test that correctly rejects the null hypothesis 80% of the time), two perfect studies will both produce "statistically significant" results only 64% of the time, and 32% of the time one will be statistically significant while the other will be statistically nonsignificant. The high probability of spurious conflict in terms of statistical significance even with acceptably powered designs has been called the false-negative problem of replication (Amrhein et al., 2019).

The false-negative problem worsens as power decreases, and as noted below it appears that most experiments on humans fail to achieve as much as 50% power for actual effects. With 50% power for the actual effect in both the initial and the replication study, the two will appear to conflict half the time, and there is only a 25% chance for both to correctly detect the effect, while the chance of both failing to detect the effect is also 25%. Thus, dichotomizing studies according to their "statistical significance" will result in gross exaggeration of perceived replication failure when the signal-to-noise ratio is low. At a minimum, what is needed instead is a direct estimation of the difference between the studies, with an understanding that any difference of importance may be due to failings of either study, as well as a possible difference in the actual effects in the two studies (Greenland & O'Rourke, 2008).

Underpowered Studies and Publication Bias Exacerbate the Gap Between Original and Replication Results

Due to smaller effects than expected and unanticipated problems in study execution, statistical power in actual studies is usually far lower than the 80% commonly calculated in the design stage. For example, it has been estimated that, for medical trials in the Cochrane Database of Systematic Reviews, the power at the actual effect size is typically only 13% (van Zwet et al., 2024). Similarly, one survey in psychology estimated typical power in psychology to be around 35% (Bakker et al., 2012), while a large survey in neuroscience estimated typical power to be only 21% (Button et al., 2013).

The problem of underpowered studies is often attributed to publication bias, whereby journals privilege statistically significant results for publication (Maxwell et al., 2015; Open Science Collaboration, 2015a; Smaldino & McElreath, 2016). This would be less of a problem if the effect under examination were very large relative to noise, in which case most studies of real effects would achieve "statistical significance" and therefore go on the record. However, when power is low, only the most exaggerated estimates will meet this "significance" criterion (Amrhein et al., 2019; Gelman & Carlin, 2014; van Zwet et al., 2021). Put differently, when the effect being studied is small and data noisy (as in soft sciences), statistically significant results tend to be outliers; and by privileging statistically significant results, journals select for estimates that exaggerate the effect. When only outliers are selected for publication, literature is biased toward inflated effect sizes, and this problem is worsened by lowering the threshold (α) at which a p-value would be declared "significant" (Amrhein et al., 2019).

Because researchers rely on effect sizes published in the literature to conduct power analyses, a literature with inflated effect sizes results in an underestimation of sample sizes needed to provide a given power. Consequently, the studies are underpowered and less likely to achieve "statistical significance" even if an effect is present—a recipe for replication failure. Meanwhile, because publication bias selects for outlying estimates, effect sizes observed in replication studies naturally regress to their mean (i.e., they attenuate toward the weaker true effect). Together, low statistical power and attenuating effect size tend to produce higher *p*-values (and thus “statistically insignificant” results) in replication studies, increasing the chance that initially positive results appear to be refuted by replication studies (Maxwell et al., 2015; van Zwet et al., 2021, 2024).

A re-examination of the RPP results can illustrate this problem. Let us assume a more realistic statistical power of 50%, as opposed to the 92% used in the RPP. Then, even if 70% of the original results were true positives, we can still expect to observe the “low replication rate” of 36% reported by the RPP (Amrhein et al., 2019). This should caution us against hastily concluding that the originally “statistically significant” studies were all or mostly false positives, and warn us that many “replication failures” arise from the statistical variability inherent in replications as well as in the original studies. In particular, due to the extreme variability of *p*-values, any claim about “replication failure” based on whether a *p*-value crossed a “significance” threshold is profoundly unreliable and potentially highly misleading (Gelman & Stern, 2006).

In summary, when dealing with probabilistic phenomena, we should view each individual study estimate as just one data point from a sampling distribution. When examining replications, these estimates will jump around a lot *in both directions* when the measurement is noisy, even if the effect is the same in each study. And the estimates will vary even more across studies when, as expected, the effect itself varies with the study design and setting. Thus, to arrive at trustworthy conclusions, we not only need high-quality data from sound studies, but we also need to be patient for studies (the data points) to accumulate enough to provide a clear picture of actual effects and how they vary across study settings. There is no shortcut, and rushing to conclusions based on one replication study is just as unwise as blindly trusting the original study.

Example: Media Coverage of the RPP

To see how the two types of uncertainty are covered in public discourse, the first author (CT) surveyed media coverage of the RPP. Two news databases (Nexus Uni and Altmetric) were used to identify English-language news reports and opinion pieces on the RPP. The final compilation includes 59 non-duplicated news reports and 15 opinion pieces published in 2015 from a total of 52 sources. Methods, data, and analyses are in the Supplemental Materials.

News coverage of the RPP project largely followed the RPP report (Open Science Collaboration, 2015a) and press release (Open Science Collaboration, 2015b), which emphasized method uncertainty but not statistical uncertainty.

Out of 59 news reports and 15 opinion pieces, only one report (Bower, 2015) and two opinion pieces (Earp & Everett, 2015; Martin, 2015) mentioned statistical uncertainty. Although almost all these reports mentioned method uncertainty, many simply quoted from the press release, which offered only very brief explanations for discrepant results:

> 1) Even though most replication teams worked with the original authors to use the same materials and methods, small differences in when, where, or how the replication was carried out might have influenced the results. 2) The replication might have failed to detect the original result by chance. 3) The original result might have been a false positive. (Open Science Collaboration, 2015b)

Such vague coverage fails to explain the sources of variation crucial for assessing the extent of the replication problem and proposed solutions.

The lack of conceptual clarity in media coverage of RPP was evident in many other ways, including explaining replication in deterministic terms; conflicting statements on whether a replication failure definitively disproves the original study; failing to recognize other criteria for interpreting replication outcomes; presenting contradictory opinions without synthesis; and, finally, noting that, among 13 articles attempting to explain the meaning of *p*-value, 12 did so incorrectly. See the Supplemental Materials for details.

One explanation for these problems may be technical complexity, which makes journalistic translation more challenging. More important, though, is people’s tendency to ignore uncertainty when interpreting statistical outputs. As explained earlier, statistical uncertainty should play a major role in judging replication outcomes, and it is a perfectly legitimate explanation for so-called “failed” replications. Yet, among 10 experts who expressed reservations about RPP outcomes—seven through interviews and three through opinion pieces—only one interviewed expert clearly pointed to this explanation. The fact that even experts tend to overlook statistical uncertainty indicates that this is an important blind spot.

In addition to exaggerating the perception of irreplicability, forcing a deterministic frame on probabilistic phenomena also results in unnecessary discord, as some reports implicitly associate “non-replication” with false positives and, at times, fraud. For example, one paper that gained media attention was titled “Nonreplicable publications are cited more than replicable ones” (Serra-Garcia & Gneezy, 2021). The titular claim was however based on whether RPP replications were statistically significant or not (as were claims in other papers such as Schafmeister, 2021, and von Hippel, 2022). The claim was subsequently reported as “Research findings that are probably wrong cited far more than robust ones, study finds” (Sample, 2021) and “Studies that are exciting but less likely to be true are cited more often in academia” (Luntz, 2021). None of these claims were supported by analyses based on proper direct comparisons of the initial and replication studies accounting for both statistical and method uncertainty.

In December 2021, the OSC released the results of its second large replication project—the Reproducibility Project: Cancer Biology (RPCB). Amid the COVID crisis, this project received scant attention from mainstream media, preventing systematic assessment of whether the quality of reporting had improved over the six intervening years. As far as we know, the only high-profile news outlet that covered the RPCB was The Associated Press (Johnson, 2021). That report did not articulate different sources of uncertainty, but instead fixated on which studies “held up” in this one replication project, illustrating a persistent blind spot in mainstream media coverage.

Discussion

It is sometimes lamented that public confidence in science is undermined by high-profile controversies, such as that over effects of diet on cancer (Schoenfeld & Ioannidis, 2013), side effects of medications (Greenland et al., 2022; Rafi & Greenland, 2020), and effects of global warming (Boykoff, 2013; Boykoff & Boykoff, 2007). Much of this controversy reflects the fact that data *alone* say nothing at all about a topic. Instead, “the data tell us. . .” is a misleading preface for a particular data interpretation. Every interpretation is laden with assumptions that can and often should be questioned, as when there are concerns about violations of experimental protocols, data integrity, or statistical assumptions. Such concerns can vary dramatically across researchers, leading to vastly different interpretations of data, even when those interpretations are restricted by various statistical conventions.

When there is disagreement about the conventions that should be applied—as in modern statistics—the range of interpretations of the same data can vary widely even when there is no concern about research protocols or data integrity. Our criticism of reporting on the “replication crisis” is based in part on the misleading conventions that have been used to claim replication failure (Amrhein et al., 2019). Those conventions can be seen as invalid implementations of statistical methods, based in particular on the fallacy of declaring conflict because an initial study attained “significance” but the replication attempt did not (a version of item 16 in the misinterpretation list of Greenland et al., 2016).

The controversies that flow from conflicting study interpretations can provide interesting material for science journalists, but a focus on dueling scientists and flip-flopping across research reports may foster nihilism. In the words of an anonymous journalist quoted by Schneider (2010), "Uncertainty in science gets translated in the popular media as doubt," where doubt generates general distrust of science rather than uncertainty about what effects are real and what actions are warranted. The resulting distrust is one of the biggest challenges facing science today.

When it comes to replication, statistics calls for us to expect reality to remain unclear until enough studies have been done to assess all sources of variability and uncertainty. Replication projects and follow-up efforts are a partial solution to rebuilding science's credibility, but cannot succeed without tackling the problem of methodological misconceptions—as we have argued, ignoring statistical uncertainty and sources of variation in replication projects exaggerates irreplicability and ultimately undermines trust. As the issue of replicability is still being debated across disciplines, science communicators can play a crucial role in preventing further erosion of public confidence in science by raising awareness of methodological uncertainty and emphasizing the cumulative and often painfully gradual nature of scientific progress.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the University of Macau under Grant MYRG2022-00046-FSS.

ORCID iD

Carol Ting https://orcid.org/0000-0002-1135-5689

Supplemental Material

Supplemental material for this article is available online at http://journals.sagepub.com/doi/suppl/10.1177/10755470241239947.

Note

1. Regarding terminology, there is considerable variation in the usage of "replicability" and "reproducibility," with many authors treating them as synonyms, and others limiting one or the other to being able to reproduce published results from the original data, as opposed to obtaining similar results from new data (Meng, 2020).

References

Altman, D. (1994). The scandal of poor medical research. *British Medical Journal*, *308*, 283–284. https://doi.org/10.1136/bmj.308.6941.1438b

Amrhein, V., Trafimow, D., & Greenland, S. (2019). Inferential statistics as descriptive statistics: There is no replication crisis if we don't expect replication. *The American Statistician*, *73*(Suppl. 1), 262–270. https://doi.org/10.1080/00031305.2018.1543137

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

Bower, B. (2015, August 27). Psychology results evaporate upon further review. *Science News*. https://www.sciencenews.org/article/psychology-results-evaporate-upon-further-review

Boykoff, M. T. (2013). Public enemy no. 1? Understanding media representations of outlier views on climate change. *American Behavioral Scientist*, *57*(6), 796–817. https://doi.org/10.1177/0002764213476846

Boykoff, M. T., & Boykoff, J. M. (2007). Climate change and journalistic norms: A case-study of US mass-media coverage. *Geoforum*, *38*(6), 1190–1204. https://doi.org/10.1016/j.geoforum.2007.01.008

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

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Nave, G., Nosek, B. A., Pfeiffer, T., Altmejd, A., Buttrick, N., Chan, T., Chen, Y., Forsell, E., Gampa, A., Heikensten, E., Hummer, L., Imai, T., & Wu, H. (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

Collins, H. (1985). *Changing order: Replication and induction in scientific practice*. SAGE.

Earp, B. D., & Everett, J. A. C. (2015, October 25). How to fix psychology's replication crisis. *The Chronicle of Higher Education*. https://www.chronicle.com/article/how-to-fix-psychologys-replication-crisis/

Errington, T. M., Mathur, M., Soderberg, C. K., Denis, A., Perfito, N., Iorns, E., & Nosek, B. A. (2021). Investigating the replicability of preclinical cancer biology. *eLife*, *10*, Article e71601. https://doi.org/10.7554/elife.71601

Gelman, A. (2016). The problems with p-values are not just with p-values. *The American Statistician, Supplemental Materials to ASA Statement on P-values and Statistical Significance*, *70*, 1–2. http://www.stat.columbia.edu/~gelman/research/published/asa_pvalues.pdf

Gelman, A., & Carlin, J. (2014). Beyond power calculations. *Perspectives on Psychological Science*, *9*(6), 641–651. https://doi.org/10.1177/1745691614551642

Gelman, A., & Loken, E. (2014). The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time. https://statmodeling.stat.columbia.edu/2021/03/16/the-garden-of-forking-paths-why-multiple-comparisons-can-be-a-problem-even-when-there-is-no-fishing-expedition-or-p-hacking-and-the-research-hypothesis-was-posited-ahead-of-time-2/

Gelman, A., & Stern, H. (2006). The difference between "significant" and "not significant" is not itself statistically significant. *The American Statistician*, *60*(4), 328–331. https://doi.org/10.1198/000313006X152649

Goodman, S. N. (1992). A comment on replication, P-values and evidence. *Statistics in Medicine*, *11*(7), 875–879. https://doi.org/10.1002/sim.4780110705

Goodman, S. N., Fanelli, D., & Ioannidis, J. P. (2016). What does research reproducibility mean? *Science Translational Medicine*, *8*(341), 341ps12. https://doi.org/10.1126/scitranslmed.aaf5027

Greenland, S. (2017). The need for cognitive science in methodology. *American Journal of Epidemiology*, *186*, 639–645. https://academic.oup.com/aje/article/186/6/639/3886035

Greenland, S. (2019). Valid P-values behave exactly as they should: Some misleading criticisms of P-values and their resolution with S-values. *The American Statistician*, *73*(Suppl. 1), 106–114. https://doi.org/10.1080/00031305.2018.1529625

Greenland, S., Mansournia, M. A., & Joffe, M. (2022). To curb research misreporting, replace significance and confidence by compatibility. *Preventive Medicine*, *164*, 107–127. https://doi.org/10.1016/j.ypmed.2022.107127

Greenland, S., & O'Rourke, K. (2008). Ch. 33 Meta-analysis. In K. J. Rothman, S. Greenland & T. L. Lash (Eds.), *Modern epidemiology* (3rd ed.). (pp. 652–682). Lippincott-Wolters-Kluwer.

Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., & Altman, D. G. (2016). Statistical tests, P values, confidence intervals, and power: A guide to misinterpretations. *The American Statistician*, *70*. https://amstat.tandfonline.com/doi/suppl/10.1080/00031305.2016.1154108/suppl_file/utas_a_1154108_sm5368.pdf

Jamieson, K. H. (2018). Crisis or self-correction: Rethinking media narratives about the well-being of science. *Proceedings of the National Academy of Sciences*, *115*(11), 2620–2627. https://doi.org/10.1073/pnas.1708276114

Johnson, C. K. (2021, December 7). Study can't confirm lab results for many cancer experiments. *The Associated Press*. https://apnews.com/article/science-business-health-cancer-marcia-mcnutt-93219170405e3de753651b89d4308461

Luntz, S. (2021, May 22). Studies that are exciting but less likely to be true are cited more often in academia. *Iflscience*. https://www.iflscience.com/studies-that-are-exciting-but-less-likely-to-be-true-are-cited-more-often-in-academia-59798

Martin, P. (2015, December 11). Wee problem with credibility of studies. *The Sydney Morning Herald*. https://www.smh.com.au/opinion/wee-problem-with-credibility-of-studies-20151211-gll6ic.html

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), 487–498. https://doi.org/10.1037/a0039400

Mayr, E. (1985). How biology differs from the physical sciences. In D. Depew & B. Weber (Eds.), *Evolution at a crossroads: The new biology and the new philosophy of science* (pp. 43–63). A Bradford Book.

McShane, B. B., Bradlow, E. T., Lynch, J. G., & Meyer, R. J. (2024). "Statistical significance" and statistical reporting: moving beyond binary. *Journal of Marketing*. Advance online publication. https://doi.org/10.1177/00222429231216910

McShane, B. B., Gal, D., Gelman, A., Robert, C., & Tackett, J. L. (2019). Abandon statistical significance. *The American Statistician*, *73*(Suppl. 1), 235–245.

Meng, X. (2020). Reproducibility, replicability, and reliability. *Harvard Data Science Review*, *2*(4). https://doi.org/10.1162/99608f92.dbfce7f9

Mulkay, M. (1984). The scientist talks back: A one-act play, with a moral, about replication in science and reflexivity in sociology. *Social Studies of Science*, *14*(2), 265–282. https://doi.org/10.1177/030631284014002008

Nelson, R. R. (2016). The sciences are different and the differences matter. *Research Policy*, *45*(9), 1692–1701. https://doi.org/10.1016/j.respol.2015.05.014

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

Open Science Collaboration. (2015b, August 27). *Massive collaboration testing reproducibility of psychology studies publishes findings* [Press release]. https://www.cos.io/about/news/massive-collaboration-testing-reproducibility-psychology-studies-publishes-findings

Peterson, D., & Panofsky, A. (2023). Metascience as a scientific social movement. *Minerva*, *61*(2), 147–174. https://doi.org/10.1007/s11024-023-09490-3

Rafi, Z., & Greenland, S. (2020). Semantic and cognitive tools to aid statistical science: Replace confidence and significance by compatibility and surprise. *BMC Medical Research Methodology*, *20*, 244. https://doi.org/10.1186/s12874-020-01105-9

Sample, I. (2021, May 21). Research findings that are probably wrong cited far more than robust ones, study finds. *The Guardian*. https://www.theguardian.com/science/2021/may/21/research-findings-that-are-probably-wrong-cited-far-more-than-robust-ones-study-finds#:~:text=2%20years%20old-,Research%20findings%20that%20are%20probably%20wrong%20cited,than%20robust%20ones%2C%20study%20finds&text=Scientific%20research%20findings%20that%20are,for%20papers%20with%20grabbier%20conclusions

Schafmeister, F. (2021). The effect of replications on citation patterns: Evidence from a large-scale reproducibility project. *Psychological Science*, *32*(10), 1537–1548. https://doi.org/10.1177/09567976211005767

Schneider, J. (2010). Making space for the "nuances of truth": Communication and uncertainty at an environmental journalists' workshop. *Science Communication*, *32*(2), 171–201. https://doi.org/10.1177/1075547009340344

Schoenfeld, J. D., & Ioannidis, J. P. (2013). Is everything we eat associated with cancer? A systematic cookbook review. *The American Journal of Clinical Nutrition*, *97*(1), 127–134. https://doi.org/10.3945/ajcn.112.047142

Serra-Garcia, M., & Gneezy, U. (2021). Nonreplicable publications are cited more than replicable ones. *Science Advances*, *7*(21), eabd1705. https://doi.org/10.1126/sciadv.abd1705

Shiffrin, R. M., Börner, K., & Stigler, S. M. (2018). Scientific progress despite irreproducibility: A seeming paradox. *Proceedings of the National Academy of Sciences*, *115*(11), 2632–2639. https://doi.org/10.1073/pnas.1711786114

Sismondo, S. (2010). *An introduction to science and technology studies*. Wiley-Blackwell.

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

van Zwet, E., Gelman, A., Greenland, S., Imbens, G., Schwab, S., & Goodman, S. N. (2024). A new look at P values for randomized clinical trials. *NEJM Evidence*. https://doi.org/10.1056/evidoa2300003

van Zwet, E., Schwab, S., & Greenland, S. (2021). Addressing exaggeration of effects from single RCTs. *Significance*, *18*(6), 16–21. https://doi.org/10.1111/1740-9713.01587

von Hippel, P. T. (2022). Is Psychological science self-correcting? Citations before and after successful and failed replications. *Perspectives on Psychological Science*, *17*(6), 1556–1565. https://doi.org/10.1177/17456916211072525

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

Wasserstein, R. L., Schirm, A. L., & Lazar, N. A. (2019). Moving to a world beyond "$p < 0.05$." *The American Statistician*, *73*(Suppl. 1), 1–19. https://doi.org/10.1080/00031305.2019.1583913

Author Biographies

Carol Ting (corresponding author) is an assistant professor at the Department of Communication, University of Macau. Her research interest focuses on social science methodology and, in particular, factors complicating research replication/reproducibility. Her recent work was published in *Social Epistemology, International Journal of Social Research Methodology*, and the *Social Science Journal*.

Sander Greenland is an Emeritus Professor of Epidemiology and Statistics at the University of California, Los Angeles. A Fellow of the American Statistical Association and the Royal Statistical Society, he is a leading contributor to epidemiologic methodology, with a focus on delineating and preventing misuse of statistical methods in observational studies. He has published over 400 articles and book chapters in epidemiology, statistics, and medicine, and has given several hundred invited lectures, seminars, and courses worldwide in epidemiologic and statistical methodology.