Published in final edited form as:

Cortex. 2019 April ; 113: 347–349. doi:10.1016/j.cortex.2018.10.030.

Less “Story” and more “Reliability” in cognitive neuroscience

David E. Huber, **Kevin W. Potter**, **Lucas D. Huszar**

University of Massachusetts, Amherst

In contrast to the data from physics or chemistry experiments, cognitive neuroscience data are noisy, requiring careful analysis of statistical reliability. In this case, noisy data are expected: Every person is different, every brain is different, and every experimental trial is different (e.g., fluctuations in attention, etc.). To handle this variability, social scientists developed inferential statistics to assess reliability (e.g., a p-value less than .05). However, statistical tests require an idealized situation in which the statistical model is correct (i.e., no violations of the statistical assumptions, such as failing to respect the measurement scale) and statistical guidelines have been followed (e.g., no peeking at the data before collecting the required sample size).

Because these statistical guidelines are often violated, and because there is a preference for publishing positive results (the so-called ‘file drawer problem’), the average reliability in the literature is lower than expected (e.g., Aarts et al., 2015; Simonsohn, Nelson, & Simmons, 2014). Even when guidelines are followed, null-hypothesis significance testing with a ‘bright line’ p-value of .05 indicates that 5% of actual null effects will be false positives, *regardless of statistical power*. For instance, if 20% of investigated effects are actually null effects, 1,000 of the 100,000 Neuroscience publications each year will be false positives. In the last few years, discussion has focused on statistical practices to address low reliability – the so-called ‘replication crisis’ (e.g., Benjamin et al., 2018; Gelman, 2018). Our goal in this commentary is not to debate statistical practices. Instead, we focus on the role played by journals, editors, and post-publication replications.

Most cognitive neuroscientists are aware of the replication crisis, but anecdotal evidence suggests this is viewed as a problem for the social sciences. To examine whether the replication crisis has had an effect in cognitive neuroscience, we ran an analysis of the publication literature using the Web of Science database. As seen in Figure 1, prior to 2014, a Psychology publication (which includes *Cognition* and *Cognitive Science)* was twice as likely to concern a replication as compared to a Neuroscience publication (which includes *Cortex* and *Cognitive Neuroscience)*. Furthermore, since 2013, there has been a 50% increase in the proportion of Psychology publications concerning a replication, and yet the situation for Neuroscience publications is unchanged.

In light of low reliability, we suggest that the field of cognitive neuroscience has implicitly adopted a different, non-statistical evaluation process – does the manuscript tell a good story? If a reported result fits into the literature, reads well, and imparts understanding, that result is deemed to be scientifically important. However, this criterion often reflects the literary and scholarly prowess of the authors as much as it does the strength of the results. In brief, the words used to report data may matter more than the data being reported.

Furthermore, once a compelling story is accepted, it is nearly impossible to excise in the event that the foundation of that story proves to be unreliable.

To be sure, storytelling will always play an important role in science – if the story is not compelling, it will not spread through the scientific community. However, we argue that science will progress at a greater clip by adopting a stronger filter for reliability before a good story is unleashed on the scientific community. The most effective way to enact this change is through the publication process. We advocate publication procedures that: 1) check reliability before publication (e.g., pre-publication replication or preregistration); and 2) allow for easy correction in the event that a result proves to be unreliable (e.g., dissemination of post-publication replications).

To make our proposal more compelling, we tell a story of our own attempt to unravel a particular finding in the literature. In the Spring of 2015, we were contacted by the New York Times to comment on an article forthcoming in *Nature Neuroscience*. We provided an alternative account of the results and ran our own replication of the study's behavioral findings to test this alternative account. To our surprise, our replication produced a null result. Subsequently, we ran a more direct replication, and when that also produced null results, we submitted a critique. The critique process at *Nature Neuroscience* sends the critique to the original authors, allowing them to write a response, and then the critique and the response are both sent **only** to the original reviewers. These are the same reviewers who saw merit in the story told in the original publication, but a wealth of data from social psychology indicates that these reviewers are likely to hold a biased opinion of any critique (e.g., Zentall, 2010). In light of this editorial process, we were not surprised when our critique failed to gain acceptance at *Nature Neuroscience*.

The action editor at *Nature Neuroscience* saw merit in our work, and said they would reconsider the decision if we ran a third, highly powered, pre-registered direct replication attempt. It took nearly a year to do so, and by the time that third attempt was complete (once again producing null results), the editor at *Nature Neuroscience* had moved on to a new position. Unexpectedly, the new action editor assigned to our case rejected our submission without review. We appealed our case to the editorial board to no avail. Subsequently, we published our replication failures in this journal (Potter, Huszar, & Huber, 2018), where our submission was reviewed by different scientists than those who reviewed the original publication.

Throughout this process, which took several years and considerable effort, we received polarized responses from our colleagues. Some were shocked by the editorial process that hindered report of our replication failures. Others questioned whether our efforts constituted the best use of limited resources – rather than amending the literature regarding this one story, we could have produced three new experiments to add to the literature. This view is rational given the incentive structure of tenure/promotion, which suggests that something needs to change to alter the cost/benefit calculation when deciding to run a replication study. In terms of maximizing personal productivity as a scientist, one should constantly strike out in new directions, rather than check prior work. However, in terms of maximizing the steady progression of science, these missteps are akin to making an initial wrong turn in the course

of a maze – they may spawn decades of subsequent research that build upon an unreliable result, producing a house of cards that can topple with little advance to scientific understanding.

The key is to prevent placement of that first (unreliable) card. Broadly speaking, scientific studies and analyses can be divided into those that are ‘exploratory’, investigating a new paradigm or novel analysis technique, versus those that are ‘confirmatory’, replicating a prior result, or running a pre-registered study. Exploratory work is critical for the advancement of science and it should not be discouraged. However, exploratory work involves a great deal of flexibility (e.g., a finding that is not apparent with one analysis might be revealed with a different analysis), and this flexibility is not acknowledged by statistical tests of reliability. Thus, exploratory components of a study should be clearly labeled as such, or if the entire study is exploratory, this could be indicated with an ‘Exploratory Reports’ article type (McIntosh, 2017). There are many situations in which it is unreasonable or infeasible to mandate a subsequent confirmatory study of an exploratory finding, such as when reporting a new life-saving drug treatment, or when running a multi-year longitudinal study. However, in situations where it is reasonable/feasible, we suggest that journals and editors request confirmation. This could consist of an initial finding followed by replication before publication. Alternatively, this could be achieved within a single study by preregistering that study, such as with the ‘Registered Reports’, article type adopted by *Nature Human Behaviour* and *Cortex* (Chambers, 2013). Critically, preregistration needs to include **all** of the proposed data analysis methods, which in the field of cognitive neuroscience afford a great deal of flexibility in light of the large number of options for pre-processing and analyzing neuroimaging data (e.g., Caballero-Gaudes & Reynolds, 2017).

Because confirmation is infeasible/unreasonable in many situations, post-publication replications are needed to correct the literature (i.e., identify the unreliable card before too many cards are placed on top). Based on our experience, peer-review of replication studies should include some reviewers beyond those who reviewed the original study, and possibly involve a different editor. Furthermore, there should be an accounting method for tracking the reliability of prior publications. Similar to ‘errata’ or ‘retraction’, there should be forward-going links that reference subsequent replication attempts (these could be maintained by third-party databases such as PubMed or GoogleScholar, but ideally they would be maintained by the original journal of publication). This way, a reader encountering an article in an online database could assess the reliability of the study before deciding whether to invest time on the study. These forward-going links to the outcomes of subsequent direct (not conceptual) replications could be sorted into ‘contraria’ (replication failures) and ‘confirmata’ (successful replications), with a stipulation of sufficient statistical power for inclusion (journals could state their criteria for these references). Alternatively, these could be grouped together as ‘replications’, with report of effect sizes and confidence intervals, avoiding any ambiguity regarding the designation of success versus failure. The authors of these replication attempts (including any attempts made by the original authors) would be responsible for contacting the original journal or relevant database with these references. These replications could be of any archival form (e.g., bioRxiv), lowering the bar for dissemination.

In closing, we advocate for policies that place greater emphasis on reliability over storytelling. This is better for science in the long run, even if it imposes additional hurdles for publishers. Our case highlights the need for independent evaluation of replication failures, but it will take more than this one change. The establishment of reliability should be a collaborative enterprise rather than an adversarial process; even when statistical guidelines are followed, false positives will occur. Researchers should be motivated to replicate their own work and publish results that contradict previous studies that they authored. When feasible, top journals should require pre-registered reports or direct replications even for initial publication. This last suggestion is easy for journals to implement and should heighten the marketability of the top journals – not only will such a publication tell a good story, but it will come with a guarantee of reliability.

Acknowledgments

This research was supported by NIMH RF1MH114277 and NSF BCS-1431147

References

Aarts AA, Anderson JE, Anderson CJ, Attridge PR, Attwood A, Axt J, ... Collaboration S (2015). Estimating the reproducibility of psychological science. Science, 349(6251).

Benjamin DJ, Berger JO, Johannesson M, Nosek BA, Wagenmakers EJ, Berk R, ... Johnson VE (2018). Redefine statistical significance. Nature Human Behaviour, 2(1), 6–10. doi: 10.1038/s41562-017-0189-z

Caballero-Gaudes C, & Reynolds RC (2017). Methods for cleaning the BOLD fMRI signal. Neuroimage, 154, 128–149. doi: 10.1016/j.neuroimage.2016.12.018 [PubMed: 27956209]

Chambers CD (2013). Registered Reports: A new publishing initiative at Cortex. Cortex, 49*(3)*, 609–610. doi: 10.1016/j.cortex.2012.12.016 [PubMed: 23347556]

Gelman A (2018). The Failure of Null Hypothesis Significance Testing When Studying Incremental Changes, and What to Do About It. Personality and Social Psychology Bulletin, 44*(1)*, 16–23. doi: 10.1177/0146167217729162 [PubMed: 28914154]

McIntosh RD (2017). Exploratory reports: A new article type for Cortex. Cortex, 96, A1–A4. doi: 10.1016/j.cortex.2017.07.014 [PubMed: 29110814]

Potter KW, Huszar LD, & Huber DE (2018). Does inhibition cause forgetting after selective retrieval? A reanalysis and failure to replicate. Cortex, 104, 26–45. [PubMed: 29715583]

Simonsohn U, Nelson LD, & Simmons JP (2014). P-Curve: A Key to the File-Drawer. Journal of Experimental Psychology-General, 143*(2)*, 534–547. doi: 10.1037/a0033242 [PubMed: 23855496]

Zentall TR (2010). Justification of Effort by Humans and Pigeons: Cognitive Dissonance or Contrast? Current Directions in Psychological Science, 19*(5)*, 296–300. doi: 10.1177/0963721410383381

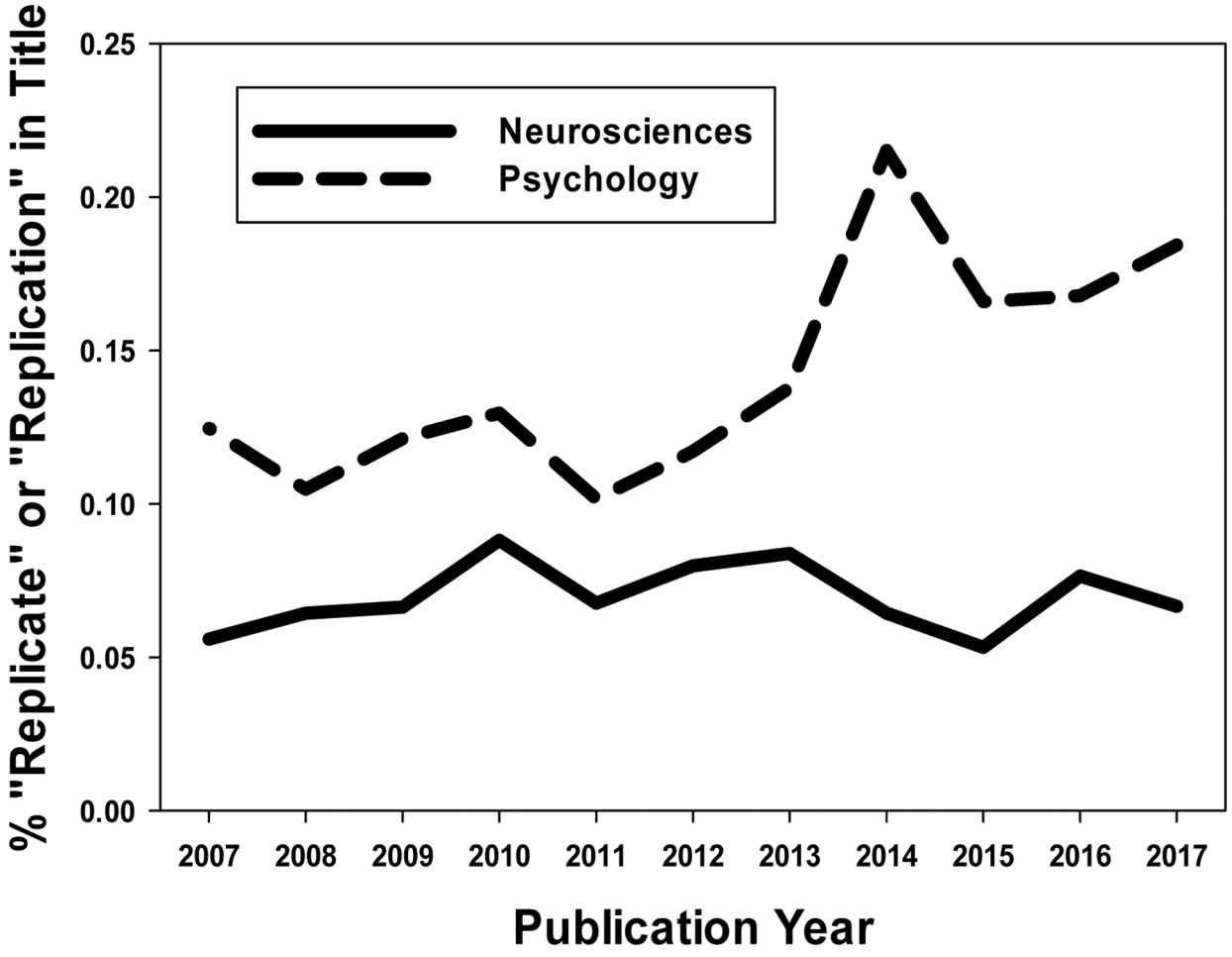

Figure 1.

The percent of publications with “replicate” or “replication” in the title as a function of publication year for publications within the research area of ‘neurosciences’ as compared to the research area of ‘psychology’, as determined through the Web of Science database (retrieved July 19, 2018). The number of neuroscience publications in a given year ranged from 76,975 in 2007 to a peak of 108,552 in 2016. The number of psychology publications in a given year ranged from 39,299 in 2007 to a peak of 67,896 in 2016.