Theoretically Important Failures to Reject the Null Hypothesis: Introduction to a special section of *Memory*

Mark L. Howe and Henry Otgaar

City, University of London and Maastricht University

One vexing problem that many researchers face when attempting to publish their research in prominent, peer-reviewed scientific journals is that data are more likely to be published when they reveal statistically significant than non-significant findings. Indeed, the point to publishing research is often the pitting of theories against each other in order to determine which one provides a better account of phenomena under investigation. If the findings do not differentiate among theoretical accounts, particularly in terms of statistical significance, then, it can be argued, there is no advance in science. Thus, manuscripts in which data analyses have led to failures to reject the null hypothesis are routinely excised from many top scientific publications, perhaps in large measure because of the inherent ambiguity in interpreting such outcomes.

There are a number of consequences associated with this model of publishing. For example, when only statistically significant findings are being published, the phenomenon under study may appear to be a particularly important one given the apparent replicability of the outcome in the scientific literature. That is, as failures to replicate are not being published, the phenomenon in question appears to be quite robust. A corollary to this illusory robustness is what has been termed the "file-drawer" problem or publication bias. Simply put, failures to replicate well-established effects go unnoticed in a scientific discipline because they go unpublished, remaining silent in laboratory file cabinets. The important issue here is that if there are a sufficient number of failures to replicate that do not receive scientific attention, then studies that have found a particular effect become just the "tip-of-the-iceberg," painting a skewed and unbalanced picture of the nature of the true effect in the population. This might even occur when such failures have adhered to the same methodology as previously published studies. Thus, through no fault of our own, as

scientists we come to believe that these effects are real and robust, simply because we are not privy to the extent of replication failures.

Not to put too fine a point on this, but failures to replicate well-known psychological phenomena have drawn considerable attention in recent years. This interest has been part of a wider movement called Open Science that strives for more transparency in scientific research. An important issue of Open Science is that more emphasis and value should be given to conducting replication studies as such studies might clarify the reliability of scientific phenomena. For example, attempts to replicate 100 published effects resulted in only a small minority (39%) of successes (Open Science Collaboration, 2015). Similarly, attempts to replicate 21 experiments that were previously published in *Nature* and *Science*, resulted in more than a third of these attempts failing (Camerer et al., 2018). Although it is well known that replication efforts can fail for any number of reasons unrelated to the phenomena under study (e.g., simply a Type II error), many have focussed on potential methodological and statistical practices in the original research (e.g., low power, p-hacking). Indeed, many researchers have made an urgent plea for methodological and statistical reform to these practices, ones that, it is argued, should make psychological science more replicable (e.g., Bishop, 2020; Chambers, 2017; Nelson, Simmons, & Simonsohn, 2018; Otgaar, Sagana, & Tupper, 2020).

Of course, attempts to replicate previous findings must have an impeccable set of standards for methodological and statistical rigor, perhaps even greater than those original studies that the replications are based on (e.g., McShane, Tackett, Böckenholt, & Gelman, 2019). However, the problem remains that if we do not have access to these failed replications, attempts at meta-analyses are doomed to failure because the research that does get published presents an incomplete picture of the robustness of the phenomenon of

interest. This problem can be circumvented in meta-analyses if researchers are willing to share what is in their "file-drawer." Although this can ameliorate this problem to some extent, perhaps with the resurgence of analyses that improve inferences about null effects (e.g., Bayesian inference, see Wagenmakers et al., 2018; equivalence testing, Lakens et al., 2020), it may now be prudent to invite researchers to contribute well-controlled experiments that fail to replicate well-known effects to the memory literature.

To this end, the editors of *Memory* welcome contributions to the journal that fail to replicate well-known effects in the memory literature. This initiative builds on another recent change in *Memory* where we have encouraged researchers to preregister their work using our Registered Reports facility. To kick off this "celebration" of all research that is memory related, whether the findings are *or* are not statistically significant in the traditional sense (for an interesting analysis of the pitfalls of traditional statistical significance, see Amrhein, Greenland, & McShane, 2019), we present four papers in which attempts to replicate well-known memory phenomena came up with null results. In the article by Howe, Akhtar, Bland, and Hellenthal (2020), the effects of aging on interference and reconsolidation are examined in episodic memory. Although as expected older adults took longer to learn than younger adults when asked to complete an episodic memory task, there were no interference or reconsolidation effects as a consequence of memory reactivation for either age group. In the article by Taylor, Sanson, Burnell, Wade, and Garry (2020) the promising effects of font type (Sans Forgetica) on memory were examined in four experiments. Although participants rated this new font as more difficult to read than Arial (where this increased difficulty was touted as leading to increased memorability), the use of Sans Forgetica did not lead to improvements in memory performance. Aust and Stahl (2020) examined the purported mnemonic effects of caffeine and found that there were no

memory performance advantages as a function of caffeine dosing. Finally, Wessel, Albers, Zandstra, and Heininga (2020) examined memory suppression effects in the think/no-think paradigm. The pattern of outcomes using both published and unpublished research raises serious questions concerning memory inhibition theory and its suggestion that repression represents a viable mechanism that underlies forgetting.

We hope that these articles are not only of interest in and of themselves, but also that it tempts you to examine your backlog of "file drawer" data. If you do have a series of studies that have failed to replicate well-known memory phenomena, and they are rigorously controlled methodologically and statistically analysed using techniques designed to investigate null hypothesis effects, then we welcome such submissions. In this way, we hope that the journal *Memory* will be able to give researchers a more complete understanding of the entirety of memory research.

References

Amrhein, V., Greenland, S., & McShane, B. (2019). Retire statistical significance. *Nature, 567*, 305-307.

Aust, F., & Stahl, C. (2020). The enhancing effect of 200 mg caffeine on mnemonic discrimination is at best small. *Memory, 28*, xxx-xxx.

Bishop, D. V. (2020). The psychology of experimental psychologists: Overcoming cognitive constraints to improve research: The 47th Sir Frederic Bartlett Lecture. *Quarterly Journal of Experimental Psychology, 73*, 1–19.

Camerer, C.F., Dreber, A., Holzmeister, F. et al. (2018). Evaluating the replicability of social science experiments in *Nature* and *Science* between 2010 and 2015. *Nature Human Behaviour, 2*, 637–644.

Chambers, C. (2017). *The seven deadly sins of psychology: A manifesto for reforming the culture of scientific practice*. Princeton University Press.

Howe, M. L., Akhtar, S., Bland, C. E., & Hellenthal, M. (2020). Reconsolidation or interference? Aging effects and the reactivation of novel and familiar episodic memories. *Memory, 28*, xxx-xxx.

Lakens, D., McLachie, N., Isager, P. M., Scheel, A. M., & Dienes, Z. (2020). Improving inferences about null effects with Bayes factors and equivalence tests. *Journal of Gerontology: Psychological Sciences, 75*, 45-57.

McShane, B. B., Tackett, J. L., Böckenholt, U., & Gelman, A. (2019). Large-scale replication projects in contemporary psychological research. *The American Statistician, 73*, 99-105.

Nelson, L. D., Simmons, J., & Simonsohn, U. (2018). Psychology's renaissance. *Annual Review of Psychology, 69*, 511-534.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), 1–8.

Otgaar, H., Sagana, A., & Tupper, N. (2020). A brief reflection on Open Science. *Psychological Research on Urban Society, 3*, 1-5.

Taylor, A., Sanson, M., Burnell, R., Wade, K. A., & Garry, M. (2020). Disfluent difficulties are not desirable difficulties: The (lack of) effects of Sans Forgetica on memory. *Memory, 28*, xxx-xxx.

Wagenmakers, E.-J. et al. (2018). Bayesian inference for psychology. Part I: Theoretical advantages and practical ramifications. *Psychonomic Bulletin & Review, 25*, 35-57.

Wessel, I., Albers, C. J., Zandstra, A. R. E., & Heinenga, V. E. (2020). A multiverse of early attempts to replicate memory suppression with the think/no-think task. *Memory, 28*, xxx-xxx.