Editorial
Replicability in Brain Imaging

Robert E. Kelly, Jr. [1,*] and Matthew J. Hoptman [2,3]

1   Department of Psychiatry, Weill Cornell Medicine, White Plains, NY 10605, USA
2   Clinical Research Division, Nathan S. Kline Institute for Psychiatric Research, Orangeburg, NY 10962, USA; matthew.hoptman@nki.rfmh.org
3   Department of Psychiatry, New York University Grossman School of Medicine, New York, NY 10016, USA
*   Correspondence: rek2005@med.cornell.edu

In the early 2010s, the "replication crisis" and synonymous terms ("replicability crisis" and "reproducibility crisis") were coined to describe growing concerns regarding published research results too often not being replicable, potentially undermining scientific progress [1]. Many psychologists have studied this problem, contributing groundbreaking work resulting in numerous articles and several Special Issues in journals, with titles such as "Replicability in Psychological Science: A Crisis of Confidence?", "Reliability and replication in cognitive and affective neuroscience research", "Replications of Important Results in Social Psychology", "Building a cumulative psychological science", and "The replication crisis: Implications for linguistics" [1–5]. Researchers in the field of brain imaging, which often dovetails with psychology, have also published numerous works on the subject, with brain imaging organizations having become staunch supporters of efforts to address the problem, such organizations including the Stanford Center for Reproducible Neuroscience and the Organization for Human Brain Mapping (OHBM), the latter having created an annual award for the best replication study [6], regularly featuring informative events concerning the replication crisis and Open Science at its annual meetings [3,7]. The purpose of the *Brain Sciences* Special Issue "The Brain Imaging Replication Crisis" is to provide a forum for discussions concerning this replication crisis in light of the special challenges posed by brain imaging.

In John Ioannidis' widely cited article entitled "Why most published research findings are false", he convincingly argues that most published findings are indeed false, with relatively few exceptions [8–10]. He supports this claim using Bayes' theorem and some reasonable assumptions concerning published research findings. It follows from Bayes' theorem that when a hypothesis test is positive, the likelihood that this study finding is true (PPV, positive predictive value) depends on three variables: the α-level for statistical significance (where α is the probability of a positive test, given that the hypothesis is false), the power of the study ($1 - \beta$, where β is the probability of a negative test, given that the hypothesis is true), and the odds that the hypothesis is true (R, the ratio of the probability that the hypothesis is true to the probability that the hypothesis is false). This relationship is expressed with the equation $\text{PPV} = R(1 - \beta)/[\alpha + R(1 - \beta)]$. From this equation, it follows that any hypothesis will likely be false, even after a positive test, when $R < \alpha$. This situation applies to fields where tested hypotheses are seldom true, which could in part explain the low replication rates observed in cancer studies [11,12]. It also follows that when the study power is equal to α, the probability that the hypothesis is true remains the same as it was before the test. Thus, inadequately powered studies lack the capacity to advance our confidence in the tested hypotheses. The PPV can also be reduced by sources of bias that elevate the actual value of α above its nominal value, for example, when publication bias [13,14] causes only positive studies to be published for a given hypothesis. When published p-values are not corrected for multiple comparisons involving negative studies, actual p-values become much higher than the published ones.

Academic incentives regarding the publication of "interesting" findings in high-impact journals can further bias research towards the production of spurious, false–positive findings through multiple mechanisms [15]. Simmons et al. [16] demonstrated with computer simulations how four common variations in research methods and data analyses allowed the inflation of actual p-values via so-called p-hacking [14,17], from 0.05 to 0.61. Researchers incentivized to find their anticipated results might be biased towards choosing methods that yield those results [18]. In the same vein, methodological errors [19] might be found less frequently when they support the anticipated results. Additionally, after seeing the results of a study, researchers might be inclined to reconsider their original hypotheses to match the observed data, so-called HARKing (hypothesizing after the results are known) [20].

To counteract these deleterious academic incentives, Serra-Garcia and Gneezy [21] proposed disincentives for the publication of nonreplicable research findings. A problem with this approach is that it can take years and considerable research resources to identify such findings. Another problem is that the replicability of findings is not necessarily a good measure of study quality. High-quality studies have the capacity to sift out replicable from irreplicable hypotheses, for example, in confirmatory studies to provide a higher margin of certainty for hypotheses already considered likely to be true, and in exploratory studies to identify promising candidates for further research. Obviously, some such candidate hypotheses will not prove replicable. Conversely, a positive study of low quality, with no capacity to separate true from false hypotheses, could prove replicable if the tested hypothesis happened to be true.

Determining which hypotheses are replicable can be especially challenging in the field of brain imaging, with many experiments lacking the power to find the sought-after differences in neural activity due to limitations in the reliability of measures combined with cost considerations limiting sample sizes [22–30]. Nonetheless, the countless pipelines from available methods of analysis can provide the needed p-value to support practically any hypothesis [31,32]. HARKing also reliably yields positive findings, which can seem confirmatory. For example, if using functional connectivity (FC) to study brain differences between two groups that differ clinically in some way, one recipe for "success" is the following: (1) divide the brain into ~100 regions and find the FC between each pair of regions, yielding ~500 such pairs whose FC differs significantly between the two groups, with $\alpha = 0.05$; (2) select a pair of such brain regions that happens to correspond to existing findings in the literature related to the studied clinical group differences; (3) write the paper as if the selected pair had been the only pair of interest, based on the literature search, thereby giving the appearance that the study is a confirmation of an expected finding.

What can improve the replicability of research results? Theoretical considerations can help to sift out likely from unlikely hypotheses even before testing begins [33]. Judicious study design can improve power. Perhaps the most efficient means of improving replicability are those that address the inflation of p-values. The preregistration of study hypotheses and methods [3,7] can prevent p-hacking and HARKing, provided that methods are specified in enough detail to eliminate flexibility in the data collection and analysis. A detailed specification of methods in published articles allows other researchers to reproduce published studies and to double-check the authors' work if study data and software are also available. Many organizations now provide tools to facilitate such a preregistration of studies and storage of data and software. The Center for Open Science [34,35], for example, is a well-funded, nonprofit organization that provides these services at little to no cost to researchers.

We welcome the submission of papers contributing further ideas for how to address the replication crisis, including replication studies or papers describing refinements of brain imaging methods to improve study power. Additionally welcome are examples of excellent study quality involving (1) preregistration with detailed methods allowing an unambiguous study reproduction and (2) availability of data and software, if feasible. Please feel free to contact the guest editor (R.E.K.) to discuss a planned study, to learn if it would be considered suitable for publication, and if not, how to make it so.

References

1. Pashler, H.; Wagenmakers, E.J. Editors' Introduction to the Special Section on Replicability in Psychological Science: A Crisis of Confidence? *Perspect. Psychol. Sci.* **2012**, *7*, 528–530. [CrossRef] [PubMed]
2. Barch, D.M.; Yarkoni, T. Introduction to the special issue on reliability and replication in cognitive and affective neuroscience research. *Cogn. Affect. Behav. Neurosci.* **2013**, *13*, 687–689. [CrossRef] [PubMed]
3. Nosek, B.A.; Lakens, D. Registered reports: A method to increase the credibility of published results. *Soc. Psychol.* **2014**, *45*, 137–141. [CrossRef]
4. Sharpe, D.; Goghari, V.M. Building a cumulative psychological science. *Can. Psychol.* **2020**, *61*, 269–272. [CrossRef]
5. Sönning, L.; Werner, V. The replication crisis, scientific revolutions, and linguistics. *Linguistics* **2021**, *59*, 1179–1206. [CrossRef]
6. Gorgolewski, K.J.; Nichols, T.; Kennedy, D.N.; Poline, J.B.; Poldrack, R.A. Making replication prestigious. *Behav. Brain Sci.* **2018**, *41*, e131. [CrossRef] [PubMed]
7. Nosek, B.A.; Alter, G.; Banks, G.C.; Borsboom, D.; Bowman, S.D.; Breckler, S.J.; Buck, S.; Chambers, C.D.; Chin, G.; Christensen, G.; et al. Promoting an open research culture. *Science* **2015**, *348*, 1422–1425. [CrossRef] [PubMed]
8. Ioannidis, J.P.A. Why Most Published Research Findings Are False. *PLoS Med.* **2005**, *2*, e124. [CrossRef] [PubMed]
9. Ioannidis, J.P.A. Discussion: Why "An estimate of the science-wise false discovery rate and application to the top medical literature" is false. *Biostatistics* **2014**, *15*, 28–36. [CrossRef] [PubMed]
10. Jager, L.R.; Leek, J.T. An estimate of the science-wise false discovery rate and application to the top medical literature. *Biostatistics* **2014**, *15*, 1–12. [CrossRef] [PubMed]
11. Begley, C.; Ellis, L. Raise standards for preclinical cancer research. *Nature* **2012**, *483*, 531–533. [CrossRef] [PubMed]
12. Prinz, F.; Schlange, T.; Asadullah, K. Believe it or not: How much can we rely on published data on potential drug targets? *Nat. Rev. Drug Discov.* **2011**, *10*, 712–713. [CrossRef] [PubMed]
13. Young, N.S.; Ioannidis, J.P.A.; Al-Ubaydli, O. Why Current Publication Practices May Distort Science. *PLoS ONE* **2008**, *5*, 1418–1422. [CrossRef] [PubMed]
14. Brodeur, A.; Cook, N.; Heyes, A. Methods Matter: P-Hacking and Publication Bias in Causal Analysis in Economics. *Am. Econ. Rev.* **2020**, *110*, 3634–3660. [CrossRef]
15. Nosek, B.A.; Spies, J.R.; Motyl, M. Scientific Utopia: II. Restructuring Incentives and Practices to Promote Truth over Publishability. *Perspect. Psychol. Sci.* **2012**, *7*, 615–631. [CrossRef] [PubMed]
16. Simmons, J.P.; Nelson, L.D.; Simonsohn, U. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychol. Sci.* **2011**, *22*, 1359–1366. [CrossRef] [PubMed]
17. Head, M.L.; Holman, L.; Lanfear, R.; Kahn, A.T.; Jennions, M.D. The Extent and Consequences of P-Hacking in Science. *PLoS Biol.* **2015**, *13*, e1002106. [CrossRef] [PubMed]
18. Giner-Sorolla, R. Science or Art? How Aesthetic Standards Grease the Way Through the Publication Bottleneck but Undermine Science. *Perspect. Psychol. Sci.* **2012**, *7*, 562–571. [CrossRef] [PubMed]
19. Vul, E.; Harris, C.; Winkielman, P.; Pashler, H. Puzzlingly High Correlations in fMRI Studies of Emotion, Personality, and Social Cognition. *Perspect. Psychol. Sci.* **2009**, *4*, 274–290. [CrossRef] [PubMed]
20. Kerr, N.L. HARKing: Hypothesizing after the results are known. *Personal. Soc. Psychol. Rev.* **1998**, *2*, 196–217. [CrossRef] [PubMed]
21. Serra-Garcia, M.; Gneezy, U. Nonreplicable publications are cited more than replicable ones. *Sci. Adv.* **2021**, *7*, eabd1705. [CrossRef] [PubMed]
22. Mumford, J.A.; Nichols, T.E. Power calculation for group fMRI studies accounting for arbitrary design and temporal autocorrelation. *Neuroimage* **2008**, *39*, 261–268. [CrossRef] [PubMed]
23. Button, K.S.; Ioannidis, J.P.A.; Mokrysz, C.; Nosek, B.A.; Flint, J.; Robinson, E.S.J.; Munafò, M.R. Power failure: Why small sample size undermines the reliability of neuroscience. *Nat. Rev. Neurosci.* **2013**, *14*, 365–376. [CrossRef] [PubMed]
24. Carp, J. The secret lives of experiments: Methods reporting in the fMRI literature. *Neuroimage* **2012**, *63*, 289–300. [CrossRef] [PubMed]
25. Elliott, M.L.; Knodt, A.R.; Ireland, D.; Morris, M.L.; Poulton, R.; Ramrakha, S.; Sison, M.L.; Moffitt, T.E.; Caspi, A.; Hariri, A.R. What Is the Test-Retest Reliability of Common Task-Functional MRI Measures? New Empirical Evidence and a Meta-Analysis. *Psychol. Sci.* **2020**, *31*, 792–806. [CrossRef] [PubMed]

26. Geuter, S.; Qi, G.; Welsh, R.C.; Wager, T.D.; Lindquist, M.A. Effect Size and Power in fMRI Group Analysis. *bioRxiv* **2018**. [CrossRef]
27. Turner, B.O.; Paul, E.J.; Miller, M.B.; Barbey, A.K. Small sample sizes reduce the replicability of task-based fMRI studies. *Commun. Biol.* **2018**, *1*, 62. [CrossRef] [PubMed]
28. Masouleh, S.K.; Eickhoff, S.B.; Hoffstaedter, F.; Genon, S. Empirical examination of the replicability of associations between brain structure and psychological variables. *Elife* **2019**, *8*, e43464. [CrossRef] [PubMed]
29. Noble, S.; Scheinost, D.; Constable, R.T. A guide to the measurement and interpretation of fMRI test-retest reliability. *Curr. Opin. Behav. Sci.* **2021**, *40*, 27–32. [CrossRef]
30. Szucs, D.; Ioannidis, J.P. Sample size evolution in neuroimaging research: An evaluation of highly-cited studies (1990–2012) and of latest practices (2017–2018) in high-impact journals. *Neuroimage* **2020**, *221*, 2017–2018. [CrossRef]
31. Botvinik-Nezer, R.; Holzmeister, F.; Camerer, C.F.; Dreber, A.; Huber, J.; Johannesson, M.; Kirchler, M.; Iwanir, R.; Mumford, J.A.; Adcock, R.A.; et al. Variability in the analysis of a single neuroimaging dataset by many teams. *Nature* **2020**, *582*, 84–88. [CrossRef] [PubMed]
32. Bowring, A.; Maumet, C.; Nichols, T.E. Exploring the impact of analysis software on task fMRI results. *Hum. Brain Mapp.* **2019**, *40*, 3362–3384. [CrossRef] [PubMed]
33. Kelly, R.E., Jr.; Ahmed, A.O.; Hoptman, M.J.; Alix, A.F.; Alexopoulos, G.S. The Quest for Psychiatric Advancement through Theory, beyond Serendipity. *Brain Sci.* **2022**, *12*, 72. [CrossRef] [PubMed]
34. Nosek, B. *Center for Open Science: Strategic Plan*; Center for Open Science: Charlottesville, VA, USA, 2017. [CrossRef]
35. Foster, E.D.; Deardorff, A. Open Science Framework (OSF). *J. Med. Libr. Assoc.* **2017**, *105*, 203. [CrossRef]