Editorial: Special Issue on Replicability in Cognitive Science

Brent Strickland (CNRS; Institut Jean Nicod; Ecole Normale Superieure)
Helen De Cruz (Saint Louis University)

1. Introduction

This special issue on what some regard as a crisis of replicability in cognitive science (i.e. the observation that a worryingly large proportion of experimental results across a number of areas cannot be reliably replicated) is informed by three recent developments. First, philosophers of mind and cognitive science rely increasingly on empirical research, mainly in the psychological sciences, to back up their claims. This trend has been noticeable since the 1960s (see Knobe, 2015). This development has allowed philosophers to draw on a wider range of relevant resources, but it also makes them vulnerable to relying on claims that may not survive further scrutiny. If we have reasons to believe that a large proportion of findings in the psychological sciences cannot be reliably replicated, this would be a problem for philosophers who use such findings in their work.

Second, philosophers are increasingly designing and carrying out their own experiments to back up claims, or to test claims earlier made from the armchair, for example, on the perceived permissibility of diverting trolleys or on the nature of free will. This growing field of experimental philosophy has diversified the intellectual field in philosophy, but may also be vulnerable to issues of replicability that philosophers did not face before.

Third, the recent evidence of apparently widespread non-replicability in the social sciences (and other fields) has forced philosophers of science to grapple with long standing questions from their field from a new perspective. To what extent does replicability matter for theory construction? How do the notions of replicability and scientific progress interact? How can normative insights from philosophy of science be used in order to improve scientific practice?

It is with these three developments in mind—the increasing importance of empirically-informed philosophy, of experimental philosophy, and of philosophy of science around replicability—that this special issue has been conceived.

2. Background

The epidemiologist and meta-scientist John Ioannidis' groundbreaking work in the medical and biological sciences shone light on the fact that a worryingly large percentage of scientific findings that inform everyday medical practice may not be reproducible (2005a; 2005b). This work was the catalyst for large-scale meta-analyses of prior studies and a raft of new studies seeking to quantify and understand the root

causes of non-replicability in other fields like economics (Camerer et al., 2016; Camerer et al., 2018; Ioannidis et al., 2017), marketing (Hunter, 2001; Aichner et al., 2016), sports science (Halperin et al, 2018), water resource management (Stagge et al., 2019), computer science (Ferrari Dacrema, Cremonesi, & Jannach, D., 2019; Ekstrand et al., 2011), and psychological science (Open Science Collaboration, 2015; Klein, 2018; Stanley, Carter, & Doucouliagos, H., 2018; Sprouse, 2011).

Specifically in the psychological sciences (the scientific field to which the guest editors and *ROPP* are most closely connected), recent work documenting unsuccessful attempts at replication has been fascinating from a sociological, methodological, and philosophical point of view. Despite some high-profile instances of fraud, such as the so-called Stapel affair (Carpenter, 2012), much of the focus in this literature has been on the large-scale production of non-reliable findings that may be a product of structurally systematic causes that go beyond individual fraudulent acts. The Open Science Collaboration project (OSC, 2015) estimated the replicability of 100 experiments from three of experimental psychology's most prominent journals through collectively organized direct replication attempts. Their overall finding was that only 36% of the replication attempts produced statistically significant findings using conventional methods of statistical analysis (i.e., Null Hypothesis Significance Testing). This is compared to 97% of studies with reported statistically significant findings, using the same method of analysis, in the original published set. Camerer et al. (2018) carried out a similar large-scale study examining the replicability of 21 papers from the social and behavioral sciences appearing in *Nature* and *Science*, and found that only 13 replicated according to similar criteria (61%). A final project from the Center for Open Science (Klein et al, 2018) put together a team of 186 researchers from 60 different laboratories to conduct direct replications of 28 high profile classic and modern findings from psychology. The authors found that only half of the studies replicated, but that amongst those that did replicate, they tended to do so in most samples.

Together this massive body of work compellingly documents some of the challenges of modern experimental psychology with regard to the interpretability and reliability of its findings. Complementary work has attempted to understand the root causes of wide-scale non-replicability, and proposes practical solutions to help rectify the problem. Root causes that have been discussed can be divided into "ultimate causes" which relate to underlying motivations and rewards, and "proximate causes" which relate to specific sub-optimal practices that are present as experiments are being designed, conducted, and analyzed. One useful rule of thumb in thinking about this distinction is that ultimate causes, unlike proximate causes, typically predict not only that there will be error but also the direction error, as we will see below.

Prominent ultimate causes that have received attention in recent years include publication bias and the related "file drawer problem" (Rosenthal, 1979; Ioannidis, 2005a), whereby null results are less likely to be published than positive results; financial and career incentives for positive and surprising findings (Heesen, 2018); high

costs associated with direct replication attempts (Everett & Earp, 2015); confirmation and experimenter bias (Strickland & Suben, 2012; Rosenthal & Fode, 1963); and even a desire on the part of the experimenter to see the truth propagated (Bright, 2017). All of these types of causes predict a specific direction of expected error or bias. For example, as Bright (2017) argued, if scientists have an earnest desire to sway the opinions of their peers toward what they think is true, they may be incentivized to illegimately produce results in line with their perceived truth. Thus the direction of expected error here would be towards conforming with experimenters' prior beliefs. Proximate causes, in contrast to ulimate causes, do not offer insight into the direction of expected error. They instead explain why results may end up being imperfect or noisy, regardless of directionality. Prominent proximate causes which have been discussed include substandard analytic and statistical practices (such as allowing too many degrees of freedom in statistical analyses; Simmons, Nelson, & Simonsohn, 2011); low statistical power (Ioannidis, 2005a); and suboptimal measurement practice (e.g. Doyen et al., 2012).

During the beginning of the replication crisis, Ioannidis (2012) advanced the idea that widespread non-replicability was evidence that science is not necessarily self-correcting, or at least not as self-correcting as one would hope. However the wide scale response in the years that have followed directly brings this idea into question, and in our view actually shows clear evidence of self-correction, at least in terms of the norms and practices that respond to systemic issues around reliability (even if not every non-reliable study ever published will undergo explicit correction, nor would this be necessarily desirable given the huge amount of resources it would take to replicate every study).

Along these lines, a number of solutions and practical responses to the problem of non-replicability have been proposed and, in some cases, adopted at scale. Standards are now generally higher for required statistical power in order to publish (Cumming, 2012), and are also higher for norms of reporting and public availability of data (Shrout & Rodgers, 2018). Pre-registration, meant to curb problematic flexibility in statistical analysis as well as data cherry picking, is more and more widely practiced in the psychological sciences (Simmons, Nelson, & Simonsohn, 2011; Kupferschmidt, 2018). Journals also often encourage direct replications prior to publication and are more tolerant of publishing non-replications (Nelson, Simmons, & Simonsohn, 2018), though this latter trend needs to be balanced against the general informativity and interest of published findings. Finally, many journals have moved towards accepting registered reports prior to the results of a study being known, which additionally helps address the file drawer problem. It is highly likely that these practices will serve to drop the overall non-replicability rate of published findings in the field (Stromland, 2019), though it remains to be seen what costs, particularly in terms of innovation, these new approaches may incur (Goldin-Meadow, 2016).

Building on this family of responses to the replicability crisis, a next generation of new practices and tools is being created that pushes the boundaries (in terms of scale and breadth) in how the social and behavioral sciences are conducted. These include for example prediction markets that identify candidate findings that merit direct replication attempts (Dreber et al., 2015), meta-analytic methods that can help identify "p-hacking" (van Aert, Wicherts, & van Assen, 2019), and large scale "conceptual" replication efforts that, unlike the direct replication studies described above, help give a sense of how theoretically robust a finding is across a number of experimental parameters where one would expect a given type of finding to appear (Landy et al., 2020).

3. Content of this special issue

Our special issue appears in this broader context. Experimental philosophers, philosophers of science, and philosophically-minded psychologists have offered unique insights into the problem of replicability, what its causes might be, and what responses might be promoted in order to improve replicability rates.

The special issue includes one large-scale, collectively organized replicability study which concentrates specifically on experiments coming from the field of experimental philosophy (x-phi) (Knobe, 2015). This study, by Cova et al., was proposed to the editorial board at *Review of Philosophy and Psychology* as part of the proposal for the current special issue. It involved coordinating between 20 research teams across 8 countries in order to directly replicate 40 individual experiments from the field. The study found a successful replication rate of about 70% in the field according to a range of criteria. Quantifying replicability in x-phi has proven useful as it strongly suggests that it is not inevitable that replicability rates hover at or below the 50% mark that has so far been observed in the social and behavioral sciences. While 70% may sound very good, it is hard to provide a benchmark for what percentage of studies should be replicable in a field, since a lot depends on tradeoffs regarding experimental design, how easy it is to recruit participants, and other factors. Furthermore, the study provides some potential insight into why replicability may vary from field to field and thus helps us improve overall practice. In particular, x-phi studies are generally run online using tools and methods where replication is cheap and easy, for example, using Amazon's Mechanical Turk (MTurk) platform for crowd-sourcing tasks. This dynamic may mean that many groups replicate internally before publishing, and also may change the cost benefit analysis of publishing "risky" results. Given the low cost of replication, researchers may rationally think there is a good chance that others will attempt to reproduce their work, making them more cautious.

The study provides some additional understanding of more particular proximate factors that may explain high vs. low replicability rates. In particular, when the main finding of a study was an observed difference (for example, in the mean value of a distribution of scores) between experimental conditions within a given sample, such a

finding tended to replicate at a high rate. However when the main finding of a study highlighted a difference between samples drawn from different populations (e.g. looking at cultural variability), this tended to have lower replicability rates.

While this study provides reasons to be optimistic about the field of experimental philosophy, Andrea Polonioli, Mariana Vega-Mendoza, Brittany Blankinship and David Carmel caution in their paper "Reporting in Experimental Philosophy: Current Standards and Recommendations for Future Practice" that the field relies extensively on null hypothesis statistical significance testing, but has only partially adopted additional measures that help to bolster the results (especially in the light of proposed shortcomings of significance testing against the null; see Trafimow & Earp, 2017; Cumming, 2012). In their review of 134 recent experimental philosophy papers, they find that only 53% of the papers report an effect size, 28% confidence intervals, 1% examined prospective statistical power and 5% report observed statistical power. Intriguingly, the extent to which these additional measures are adopted does not impact how often a paper is cited.

Other articles in the special issue examined a range of ultimate and proximate causes for non-replicability within the social and behavioral sciences at large, and proposed some novel solutions that should be considered as candidate scalable solutions. In “The Alpha War” Edouard Machery argues in favor of decreasing the significance level threshold for publishability by an order of magnitude, a measure that would be practical in many cases, likely effective, and could thus be broadly implemented.

Deborah Mayo's "Significance tests: Vitiated or Vindicated by the Replication Crisis in Psychology?" pushes back against the idea that statistical significance testing would be to blame for unreplicable results, because statistical significance testing makes it too easy to find effects. However, such claims frequently miss the mark of what statistical significance testing is supposed to do. As Mayo argues, even Ronald Fisher, a statistician who contributed to the theory behind null hypothesis testing in the 1930s, already cautioned that one cannot demonstrate a genuine experimental phenomenon from just a single small p-value. Yet this is what critics of statistical significance testing say is happening collectively. Moreover, alternatives to significance testing, such as likelihood ratios, Bayes Factors, or Bayesian updating, do not fare better than statistical significance based on alpha thresholds and p-values, and might give bias a free pass, because such methods, unlike statistical significance tests, cannot pick up on how data-dredging alters the capabilities of tests to distinguish genuine effects from noise.

Relatedly, Lincoln John Colling and Dénes Szűcs compare the frequentist and Bayesian approach as two very different perspectives on evidence and inference in their paper "Statistical Reform and the Replication Crisis". They argue that the frequentist approach prioritizes error control, and the Bayesian approach offers a formal method for quantifying the relative strength of evidence for hypotheses.

Finally Mary Amon and John Holden advocate a general systems framework that can serve as a complement to standard inferential statistics, and better accommodate intrinsic fluctuations and contextual adaptations.

In our view, this collection of articles provides a unique perspective on the problem of replicability from philosophers of science and experimental philosophers. We hope that this special issue might spur further dialogue within these communities around this important topic.

References:

Aichner, T, Coletti, P., Forza, C., Perkmann, U., Trentin, A. (2016). Effects of subcultural differences on country and product evaluations: A replication study. *Journal of Global Marketing*. 29 (3): 115–127. doi:10.1080/08911762.2015.1138012.

Bohannon, J. (2016). About 40% of Economics Experiments Fail Replication Survey. *Science*. doi:10.1126/science.aaf4141

Bright, L. K. (2017). On fraud. *Philosophical Studies*, 174(2), 291-310.

Camerer, C., Dreber, A., Forsell, E., Ho, T., Huber, J., Johannesson, M., Kirchler, M., Almenberg, J., Altmejd, A. (2016). Evaluating replicability of laboratory experiments in economics. *Science*, 351 (6280), 1433–1436.

Camerer, C., Dreber, A., et al. (2018). Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour*. 2 (9), 637–644. doi:10.1038/s41562-018-0399-z.

Carpenter, S. (2012). Harvard psychology researcher committed fraud, U.S. investigation concludes', *Science*, 6 September. Available at https://www.sciencemag.org/news/2012/09/harvard-psychology-researcher-committed-fraud-us-investigation-concludes (accessed 8 Dec. 2020).

Cumming, G. (2012). Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis. New York, NY US: Routledge/Taylor & Francis Group.

Doyen, S., Klein, O., Pichon, C.-L., & Cleeremans, A. (2012). Behavioral priming: It's all in the mind, but whose mind? *PLoS ONE*, 7(1), e29081.

Dreber A., et al. (2015). Using prediction markets to estimate the reproducibility of scientific research. *Proc. Natl. Acad. Sci. U.S.A.*, 112, 15343-15347.

Ekstrand, M., Ludwig, M., Konstan, J., & Riedl, J. (2011). *Rethinking the Recommender Research Ecosystem: Reproducibility, Openness, and LensKit*. *Proceedings of the Fifth ACM Conference on Recommender Systems*. RecSys '11. New York, NY, USA: ACM. pp. 133–140. doi:10.1145/2043932.2043958.

Everett, J.A.C., & Earp, B.D. (2015). A tragedy of the (academic) commons: Interpreting the replication crisis in psychology as a social dilemma for early-career researchers. *Frontiers in Psychology, 6*(1152), 1-4.

Ferrari Dacrema, M., Cremonesi, P., Jannach, D. (2019). Are we really making much progress? A worrying analysis of recent neural recommendation approaches. *Proceedings of the 13th ACM Conference on Recommender Systems*. ACM: 101–109. arXiv:1907.06902. doi:10.1145/3298689.3347058.

Goldin-Meadow, S. (2016). Why pre-registration makes me nervous. *Association for Psychological Science Observer*. https://www.psychologicalscience.org/observer/why-preregistration-makes-me-nervous

Halperin, I., Vigotsky, A., Foster, C., & Pyne, D. (2018). Strengthening the practice of exercise and sport-science research. *International Journal of Sports Physiology and Performance*. 13 (2): 127–134. doi:10.1123/ijspp.2017-0322.

Heesen, R. (2018). Why the reward structure of science makes reproducibility problems inevitable. *The Journal of Philosophy*. 115(12):661-674. DOI: 10.5840/jphil20181151239

Hunter, J.. (2001). The desperate need for replications. *Journal of Consumer Research*. 28 (1): 149–158. doi:10.1086/321953.

Ioannidis, J. (2005a). Why most published research findings are false. *PLOS Medicine*. 2 (8): e124.

Ioannidis J. (2005b). Contradicted and initially stronger effects in highly cited clinical research. *JAMA*. 294 (2): 218–228.

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

Ioannidis, J., Stanley, T., Doucouliagos, H. (2017). The power of bias in economics research. *The Economic Journal*. 127 (605): F236–F265. doi:10.1111/ecoj.12461.

Ioannidis, J. (2012). Why science is not necessarily self-correcting. *Perspectives on Psychological Science*. 7 (6): 645–654.

Klein, R. (2018). Many Labs 2: Investigating variation in replicability across samples and settings. *Advances in Methods and Practices in Psychological Science*. 1 (4): 443–490. doi:10.1177/2515245918810225.

Knobe, J. (2015). Philosophers are doing something different now: Quantitative data. *Cognition*, 135, 36-38.

Kupferschmidt, K. (2018). More and more scientists are pre-registering their studies. Should you? *Science*. doi:10.1126/science.aav4786 Accessed Dec. 10, 2020.

Landy, J.F., et al., (2020). Crowdsourcing hypothesis tests: Making transparent how design choices shape research results. *Psychological Bulletin*. *146*(5), 451–479. https://doi.org/10.1037/bul0000220.

Nelson, L. D., Simmons, J. & Simonsohn, U. (2018). Psychology's renaissance. *Annual Review of Psychology*, 69, 511-534, doi:10.1146/annurev-psych-122216-011836.

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

Rosenthal, R. (1979), 'The file drawer problem and tolerance for null results', Psychological Bulletin, 86, 638-641

Rosenthal, R., & K. Fode. (1963). The effect of experimenter bias on performance of the albino rat. *Behavioral Science*, 8: 183–189.

Shrout, P. E., Rodgers, J. L. (2018). Psychology, science, and knowledge construction: Broadening perspectives from the replication crisis. *Annual Review of Psychology, 69*, 487–510. doi:10.1146/annurev-psych-122216-011845

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*, 1359–1366.

Sprouse, J. (2011). A validation of Amazon Mechanical Turk for the collection of acceptability judgments in linguistic theory. *Behavior research methods*, *43*(1), 155-167.

Stagge, J., Rosenberg, D., Abdallah, A., Akbar, H., Attallah, N., & James, R. (2019). Assessing data availability and research reproducibility in hydrology and water resources. *Scientific Data*. 6: doi:10.1038/sdata.2019.30.

Stanley, T., Carter, E., & Doucouliagos, H. (2018). What meta-analyses reveal about the replicability of psychological research. *Psychological Bulletin*. 144(12): 1325–1346. doi:10.1037/bul0000169.

Strickland, B., & Suben, A. (2012). Experimenter philosophy: The problem of experimenter bias in experimental philosophy. *Review of Philosophy and Psychology*, 3(3), 457-467.

Stromland, E. (2019). Pre-registration and reproducibility. *Journal of Economic Psychology*, 75(a), doi.org/10.1016/j.joep.2019.01.006.

Trafimow, D., & Earp, B. D. (2017). Null hypothesis significance testing and Type I error: The domain problem. *New Ideas in Psychology*, 45, 19-27.

van Aert, R., Wicherts, J., & van Assen, M. (2019) Publication bias examined in meta- analyses from psychology and medicine: A meta-meta-analysis. *PLoS ONE, 14*, e0215052, doi:10.1371/journal.pone.0215052.