



Mitigating the epidemic of type I error: ecology and evolution can learn from other disciplines

Timothy H. Parker^{1*} and Shinichi Nakagawa²

¹ Biology Department, Whitman College, Walla Walla, WA, USA

² National Centre for Growth and Development, Department of Zoology, University of Otago, Dunedin, New Zealand

*Correspondence: parkerth@whitman.edu

Edited by:

François Criscuolo, Centre National de la Recherche Scientifique, France

Reviewed by:

Cristian Pasquaretta, Centre National de la Recherche Scientifique - Institute Pluridisciplinaire Hubert Curien, France

Keywords: bias, p-hacking, registered reports, replication, reproducibility, type I error

SECTION

In probabilistic disciplines from psychology to cancer biology and behavioral ecology, a disturbing quantity of empirically derived understanding has been challenged and found wanting (Begley and Ellis, 2012; Carpenter, 2012; Parker, 2013). Recently, it was reported that 47 of 53 “landmark” cancer studies from the past decade could not be reproduced (Begley and Ellis, 2012). Ongoing attempts to replicate results in psychology (Carpenter, 2012) have found that substantial portions do not stand subsequent tests (Reproducibility Project: <https://osf.io/ezcuj/>). Although some well-publicized cases of data fabrication have plagued that field recently (Vogel, 2011), much of the lack of repeatability is expected to result from less nefarious forms of bias (Ioannidis, 2005). Closer to home, a recent meta-analysis of studies of plumage color in a European songbird has substantially clouded what had been hailed as a model for the understanding of plumage color and sexual selection (Parker, 2013). The crux of the problem is that the published literature, especially in highly probabilistic systems, suffers from inflated type I error (false positive) rates, and careful replication is too rare to reliably separate the robust results from those resulting from error (Ioannidis, 2005; Parker, 2013). Thus, many published results are incorrect, and these results are too rarely discredited. Concerns about problems of empirical error are receiving attention from prestigious journals (e.g., Nature; Nuzzo, 2014), and in the popular press (e.g., Lehrer, 2010; Anonymous, 2013)

they have stimulated a discourse that may be eroding public confidence in science.

Strategies to reduce the problems of inflated error and infrequent replication are emerging in psychology, neuroscience, and medicine (Baker, 2012; Carpenter, 2012). High rates of type I error and low rates of replication may appear to result primarily from the decisions of individual researchers. These researchers are, however, responding to institutional incentive structures. For instance, funding bodies support novel projects to the exclusion of replications, and high impact journals also place a premium on novelty (Palmer, 2000; Kelly, 2006). As another example, most journals select articles based on study outcome rather than just soundness of hypothesis, predictions, and methods (Chambers, 2013). Thus, researchers often choose to report the most interesting subsets of results or pursue other forms of biased reporting rather than reporting the entire set of outcomes (John et al., 2012). Institutions also promote bias, and possibly even academic dishonesty, by basing professional evaluation and remuneration on number of publications and the stature of the journals in which they are published (Qiu, 2010; John et al., 2012). Thus, effective strategies will come from changes in the institutions that influence our research practices, such as professional societies (including journals) and funding agencies (Parker, 2013). It is precisely at this institutional level that psychology and medicine are tackling the challenge of reducing bias and increasing replication. Initiatives in these other disciplines are not necessarily

templates directly transferable to ecology and evolution. Yet, such examples should serve to stimulate discussion and they clearly demonstrate that redesigning incentive structures is possible.

Reducing incomplete and biased reporting of results may be accomplished by encouraging or requiring registration of studies at their initiation (Schooler, 2011). Since 2000, the US government has provided a registry for clinical trials of medical interventions (ClinicalTrials.gov). Registration prior to initiation is a requirement of many funding agencies and medical journals, and thus has become “standard practice” (Huser and Cimino, 2013). Although results from approximately half of registered trials end up unpublished, about a third of the unpublished studies post some results in the registry (Ross et al., 2009). Further, the registry facilitates a more precise estimate of reporting bias, and provides contact information for researchers with unpublished work. Thus, the bias in available results has dropped along with our ignorance of this bias. These are highly desirable outcomes.

A conceptually similar idea is the “registered report” initiated by the neuroscience journal *Cortex* in 2013 (Chambers, 2013). To publish in the registered report section of the journal, researchers submit a study plan for peer review and conditional acceptance prior to gathering data (<http://www.elsevier.com/journals/cortex/0010-9452/guide-for-authors>). This counteracts several forms of publication bias, including editors’ preferences for statistically significant or novel outcomes, and

the tendency of researchers to selectively report the more interesting facets of their results (Chambers, 2013). This option for publication remains rare, but if widely adopted it could serve as an important tool for reducing bias.

A prestigious journal in psychology has taken an alternate approach to reducing incomplete and biased reporting. Following the suggestions of Simmons et al. (2011) in their prominent paper on inflated false positive rates, *Psychological Science*, as of 2014, requires authors to confirm that they are reporting on their full data set, including “all independent variables or manipulations” and “all dependent variables or measures,” and “how sample size was determined.” This requirement rests on the assumption that many researchers who might otherwise be willing to report a biased subset of their results would not willingly make false statements, either because of the clear moral implication or because of the risk to one’s career (Simmons et al., 2011). Although it is too early to determine success, employing such statements is a promising strategy for reducing reporting bias.

Bias in reporting is clearly problematic, but equally problematic is the pervasive lack of sufficient replication to identify robust patterns (Palmer, 2000; Kelly, 2006). The Reproducibility Initiative is a private organization that facilitates and incentivizes replication (Baker, 2012). Researchers can submit an experiment and the Reproducibility Initiative locates an appropriate lab, anonymous to the original researchers, to conduct the replication. The researchers pay for this service, but if the original results are reproduced, their work can carry an “independently validated” badge (<http://reproducibilityinitiative.org>). The open access journal PLOS ONE has joined the initiative with a pledge to publish replications (Baker, 2012). Further, at least some replication will be funded independently. In 2013, the Reproducibility Initiative received a 1.3 million dollar grant from the Center for Open Science to replicate a series of high profile cancer studies (<http://centerforopenscience.org/pr/2013-10-16/>). It is not yet clear whether a certificate of independent validation will serve as a sufficient incentive

to promote widespread replication, but at least for researchers with substantial financial stakes in getting their research right, the appeal of independent validation is strong (Phillips, 2012).

Some of the most extensive replication efforts are currently underway in psychology, also partly funded by the Center for Open Science. The Reproducibility Project: Psychology (distinct from the Reproducibility Initiative described above) currently involves over 150 researchers volunteering to replicate studies published in 2008 in three well-respected psychology journals (<https://osf.io/ezcuj/wiki/home/>). The failure to replicate a number of published results justifies this ongoing effort to increase study replication, but more important, the open and collaborative model pursued in these replications serve as a potential model for pursuing replication more widely. In a related project from the Center for Open Science, an entire issue of *Social Psychology* earlier this year was devoted to reporting replications of important studies (Nosek and Lakens, 2014) dating back as far as the 1930’s (Klein et al., 2014). Some of the original studies were supported and some were not, and others appeared more complex than previously realized (Nosek and Lakens, 2014). As proposed more than a decade ago (Palmer, 2000), funding replications and allocating journal space to publishing them appears to increase their frequency. In the case of psychology, the Center for Open Science’s strong and multi-faceted institutional support for replication has clearly also been important.

Other proposals that may reduce biased reporting and increase replication abound. For instance, major funding agencies could devote a portion of their budgets to support worthy replications (Palmer, 2000) or could preferentially fund proposals that rest on better-replicated foundations (Parker, 2013). Simply ensuring that authors report sufficient methodological and statistical details (Nakagawa and Cuthill, 2007) is a useful step. To this end, standard guidelines are gaining support and endorsements in (bio-)medical sciences (e.g., ARRIVE—Animal Research: Reporting of *In Vivo* Experiments; Kilkenny et al., 2010) and even ecology (Hillebrand and Gurevitch, 2013). Such guidelines are inspired by various

motives, but their common thread is that they should reduce selective reporting and facilitate replication. Providing publishing outlets that evaluate research based on the quality of the methods and inferences rather than on the appeal of the outcome should also help, but such journals may be of most use when complemented by incentives to publish negative results (http://www.scilogs.com/communication_breakdown/negative-results-plos-one/).

Unfortunately, we lack model strategies for reducing the effects of some important negative institutional incentives. For instance, we know of no movements to counteract the growing trend for universities, research institutes, and funding agencies to evaluate researchers based on number of publications or impact factors of the journals in which they publish (Qiu, 2010). Given that this trend tends not to originate in or to be controlled by decisions at the level of the discipline, it may be more difficult to counteract. Widespread grassroots opposition to these evaluation methods could lead to advocacy by influential people and institutions, and thus ultimately to a reduction in the practice of evaluating researchers in this simplistic manner. Certainly without a public discussion of the perverse incentives imposed by these evaluation methods, they seem unlikely to change.

Where do the fields of evolution and ecology stand? Although the published discussion of the problem of biased reporting and poor replication (Palmer, 2000; Kelly, 2006; Nakagawa and Cuthill, 2007; Forstmeier and Schielzeth, 2011; Parker, 2013) is not new, and our conversations with colleagues suggest aspects of these problems are relatively widely recognized, little has yet been done. One exception is that in 2011, four prominent journals in our fields began requiring authors to deposit their raw data in publicly accessible databases (e.g., Whitlock et al., 2010), and more journals are joining this movement. Unfortunately, data archiving is expected to go only a small way toward addressing biased reporting, not only because thorough re-analyses of such data sets will be time consuming and thus probably rare, but also because authors can still readily publish (and post raw data from) a biased subset of their work (Simmons et al., 2011). Further,

data archiving does not create incentives to replicate important findings. Although data archiving itself has a number of issues and has been controversial (Roche et al., 2014), its adoption demonstrates that evolutionary biologists and ecologists and the institutions they constitute can accept and promote substantial changes in the way research is conducted and published. This is a hopeful sign.

Which of the strategies discussed above, if any, are right for evolution and ecology? Unlike psychology and medicine, evolution and ecology consider the entire spectrum of organisms and living systems. Clearly, such breadth of study subject raises distinct challenges. For instance, a field biologist cannot simply arrange for a laboratory-for-hire to replicate her/his experiments. Yet, this difficulty does not mean that we should give up on replicating important studies (Kelly, 2006). Instead, we need to gather our collective experiences and insights and develop plans suitable for our own disciplines and sub-disciplines. We may find that some proposals, such as the development of voluntary hypothesis testing registries (Schooler, 2011), guidelines for improved statistical reporting (Hillebrand and Gurevitch, 2013), or devoting sections of journals to replication (Palmer, 2000) would face relatively few practical obstacles to implementation in evolution and ecology. Indeed, we expect that more ideas well-tailored to our disciplines will emerge from an open and active discussion. If we ignore these issues of biased reporting and a lack of replication and continue as we have, we do so at our peril. Other disciplines have responded to the crisis with bold steps. Let's figure out the ways forward for evolution and ecology.

AUTHOR CONTRIBUTIONS

Timothy H. Parker conceived of this paper and developed it in consultation with Shinichi Nakagawa. Timothy H. Parker drafted and revised the original manuscript and Shinichi Nakagawa contributed important additional content and provided editorial suggestions.

ACKNOWLEDGMENTS

We are grateful for the thoughtful comments of anonymous reviewers and to our colleagues for stimulating conversations on this topic.

REFERENCES

- Anonymous. (2013, October 19). Trouble at the lab. *The Economist*.
- Baker, A. (2012, August 14). Independent labs to verify high-profile papers. *Nature News*.
- Begley, C. G., and Ellis, L. M. (2012). Raise standards for preclinical cancer research. *Nature* 483, 531–533. doi: 10.1038/483531a
- Carpenter, S. (2012). Psychology's bold initiative. *Science* 335, 1558–1561. doi: 10.1126/science.335.6076.1558
- Chambers, C. D. (2013). Registered reports: a new publishing initiative at cortex. *Cortex* 49, 609–610. doi: 10.1016/j.cortex.2012.12.016
- Forstmeier, W., and Schielzeth, H. (2011). Cryptic multiple hypotheses testing in linear models: over-estimated effect sizes and the winner's curse. *Behav. Ecol. Sociobiol.* 65, 47–55. doi: 10.1007/s00265-010-1038-5
- Hillebrand, H., and Gurevitch, J. (2013). Reporting standards in experimental studies. *Ecol. Lett.* 16, 1419–1420. doi: 10.1111/ele.12190
- Huser, V., and Cimino, J. J. (2013). Linking ClinicalTrials.gov and PubMed to track results of interventional human clinical trials. *PLoS ONE* 8:e68409. doi: 10.1371/journal.pone.0068409
- Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Med.* 2:124. doi: 10.1371/journal.pmed.0020124
- John, L. K., Loewenstein, G., and Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychol. Sci.* 23, 524–532. doi: 10.1177/0956797611430953
- Kelly, C. D. (2006). Replicating empirical research in behavioral ecology: how and why it should be done but rarely ever is. *Q. Rev. Biol.* 81, 221–236. doi: 10.1086/506236
- Kilkenny, C., Browne, W. J., Cuthill, I. C., Emerson, M., and Altman, D. G. (2010). Improving Bioscience Research Reporting: The ARRIVE Guidelines for Reporting Animal Research. *PLoS Biol* 8:e1000412. doi: 10.1371/journal.pbio.1000412
- Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B. Jr., Bahník, Š., Bernstein, M. J., et al. (2014). Investigating variation in replicability. *Soc. Psychol.* 45, 142–152. doi: 10.1027/1864-9335/a000178
- Lehrer, J. (2010, December 13). The truth wears off. *The New Yorker*, pp. 52–57.
- Nakagawa, S., and Cuthill, I. C. (2007). Effect size, confidence interval and statistical significance: a practical guide for biologists. *Biol. Rev.* 82, 591–605. doi: 10.1111/j.1469-185X.2007.00027.x
- Nosek, B. A., and Lakens, D. (2014). Registered reports. *Soc. Psychol.* 45, 137–141. doi: 10.1027/1864-9335/a000192

- Nuzzo, R. (2014). Scientific method: statistical errors. *Nature* 506, 150–152. doi: 10.1038/506150a
- Palmer, A. R. (2000). Quasireplication and the contract of error: lessons from sex ratios, heritabilities and fluctuating asymmetry. *Annu. Rev. Ecol. Syst.* 31, 441–480. doi: 10.1146/annurev.ecolsys.31.1.441
- Parker, T. H. (2013). What do we really know about the signalling role of plumage colour in blue tits? A case study of impediments to progress in evolutionary biology. *Biol. Rev.* 88, 511–536. doi: 10.1111/brv.12013
- Phillips, M. L. (2012, August 21). Initiative tackles scientific study validation. *BioTechniques: News*.
- Qiu, J. (2010). Publish or perish in China. *Nature* 463, 142–143. doi: 10.1038/463142a
- Roche, D. G., Lanfear, R., Binning, S. A., Haff, T. M., Schwanz, L. E., Cain, K. E., et al. (2014). Troubleshooting public data archiving: suggestions to increase participation. *PLoS Biol* 12:e1001779. doi: 10.1371/journal.pbio.1001779
- Ross, J. S., Mulvey, G. K., Hines, E. M., Nissen, S. E., and Krumholz, H. M. (2009). Trial publication after registration in ClinicalTrials.gov: a cross-sectional analysis. *PLoS Med.* 6:e1000144. doi: 10.1371/journal.pmed.1000144
- Schooler, J. (2011). Unpublished results hide the decline effect. *Nature* 470, 437–437. doi: 10.1038/470437a
- Simmons, J. P., Nelson, L. D., and Simonsohn, U. (2011). False positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychol. Sci.* 22, 1359–1366. doi: 10.1177/0956797611417632
- Vogel, G. (2011). Psychologist accused of fraud on “astounding scale”. *Science* 334, 579. doi: 10.1126/science.334.6056.579
- Whitlock, M. C., McPeck, M. A., Rausher, M. D., Rieseberg, L., and Moore, A. J. (2010). Data archiving. *Am. Nat.* 175, E145–146. doi: 10.1086/650340

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 01 July 2014; accepted: 07 November 2014; published online: 25 November 2014.

Citation: Parker TH and Nakagawa S (2014) Mitigating the epidemic of type I error: ecology and evolution can learn from other disciplines. *Front. Ecol. Evol.* 2:76. doi: 10.3389/fevo.2014.00076

This article was submitted to *Behavioral and Evolutionary Ecology*, a section of the journal *Frontiers in Ecology and Evolution*.

Copyright © 2014 Parker and Nakagawa. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.