Metaresearch for Evaluating Reproducibility in Ecology and Evolution

FIONA FIDLER, YUNG EN CHEE, BONNIE C. WINTLE, MARK A. BURGMAN, MICHAEL A. MCCARTHY AND ASCELIN GORDON

Recent replication projects in other disciplines have uncovered disturbingly low levels of reproducibility, suggesting that those research literatures may contain unverifiable claims. The conditions contributing to irreproducibility in other disciplines are also present in ecology. These include a large discrepancy between the proportion of "positive" or "significant" results and the average statistical power of empirical research, incomplete reporting of sampling stopping rules and results, journal policies that discourage replication studies, and a prevailing publish-or-perish research culture that encourages questionable research practices. We argue that these conditions constitute sufficient reason to systematically evaluate the reproducibility of the evidence base in ecology and evolution. In some cases, the direct replication of ecological research is difficult because of strong temporal and spatial dependencies, so here, we propose metaresearch projects that will provide proxy measures of reproducibility.

Keywords: metaresearch, reproducibility, publication bias, transparency, open science

Evaluating the reproducibility of scientific results

Reproducibility is a broad term used to describe the ability to replicate an experiment or study and/or its outcomes (see box 1). Most philosophies of science agree that it is a core component of scientific theory testing (Schmidt 2009). Although concern over the reproducibility of scientific results is not itself new, large-scale metaresearch projects aimed at directly evaluating the reproducibility of entire fields of research are a relatively new and growing phenomenon. So far, the results of such projects in other disciplines have amplified anxiety over the state of the scientific evidence base.

For example, the Open Science Collaboration (2015) recently conducted a large metaresearch project in psychology that directly replicated published studies. Only 39% of replications reproduced the results of the original studies, with replication effect sizes averaging only half those of the originals. Similar metaresearch evaluations of biomedical research have produced a range of equally discouraging reproducibility estimates, from approximately 11% (Begley and Ellis 2012) to 49% (Freedman et al. 2015).

To date, there have been no equivalent metaresearch projects in ecology and evolution. However, as ecological analyses are increasingly complex in their statistical approaches, there have been several calls for greater methodological transparency over at least a decade (e.g., Ellison 2006, 2010, Parker and Nakagawa 2014). A strong case has been made for the existence of related problems in the discipline (Parker et al. 2016b), and in 2016, a disciplinary specific set of transparency and openness promotion (TOP) guidelines, known as tools for transparency in ecology and evolution (TTEE; *https://osf.io/g65cb*), were compiled. Editorials promoting these guidelines have now appeared in seven journals in the discipline, including *Ecology Letters* (Parker et al. 2016a) and *Conservation Biology* (Parker et al. 2016c). This growing interest and awareness suggests that the discipline is now ready to meet metaresearch challenges.

In some areas of ecology, the feasibility of direct replication projects that have characterized metaresearch in other disciplines is severely limited (Schnitzer and Carson 2016). Ecological processes often operate and vary over large spatial scales and long time horizons, and temporal and spatial dependencies can make re-collecting appropriate data difficult—and in some cases impossible. However, there are compelling arguments that in some subfields, such as behavioral ecology, direct or at least close partial replications are feasible (Nakagawa and Parker 2015), and their absence in the published literature is problematic (Kelly 2006). We agree and suggest that it is time for the discipline to assess

BioScience 67: 282–289.

doi:10.1093/biosci/biw159

Advance Access publication 13 January 2017

Box 1. Defining replication and reproducibility.

It is by replicating a study that we determine whether or not its results are reproducible. A range of concepts and definitions relating to replication and reproducibility already exist (e.g., Cassey and Blackburn 2006), as do more finely grained typologies (e.g., Nagakawa and Parker 2015). Here, we focus on two broad categories, which include direct and conceptual replication, in line with Schmidt (2009).

Direct replication adheres as closely as possible to original study. The Reproducibility Project Psychology is an example; the Open Science Collaboartion (2015) repeated the full experimental procedure of 100 published studies, including data collection and analysis, using the same (or similar) protocols as the original study. Direct replication projects pose the greatest challenge for ecology, especially in subfields in which temporal and spatial dependencies are strong. *Direct re-analysis* projects involve identical (or very close) repetition of the analytic procedure, starting from the same raw data as the original. Related to this, *direct computational reproducibility* refers to the ability to reproduce particular analysis outcomes from the same data set using the same code and software.

Conceptual replication repeats a test of theory or hypothesis made in past research but does so using different methods. Conceptual replications aim to test the underlying concepts or hypotheses as the original study but may operationalize concepts differently and use different measurements, statistical techniques, interventions, and/or instruments to see whether they lead to the same conclusion. *Conceptual re-analysis* involves analysis of the same raw data set but permits the use of justified alternative approaches, methods, and models (see, e.g., Silberzahn and Uhlmann 2015).

Both direct and conceptual replications help establish the generalizability of facts, but they fulfill different scientific functions. Direct replications control for sampling error, artifacts, and fraud, providing crucial information about the reliability and validity of prior empirical work. Conceptual replications help corroborate the underlying theory or substantive (as opposed to statistical) hypothesis in question and contribute to our understanding of concepts and mechanisms.

the reproducibility of its published literature wherever possible. The use of “direct replication” (box 1) is not the only means of evaluation, and in the remainder of this article, we describe other ways to take stock of the problem.

Why would ecology and evolution have a reproducibility problem?

Outright fraud and fake data obviously result in reproducibility problems, and there is some evidence that the frequency of such cases is increasing in other disciplines, such as biomedicine (Fang et al. 2013). Fang and colleagues (2013) estimated that that 43.4% of retracted articles are withdrawn because of fraud. However, roughly half a million biomedical articles are published annually, and only about 400 per year are retracted (Oranksy 2015), so this amounts to a very small proportion of (approximately 0.1%) of the literature. In short, fraud is not the main source of irreproducibility in those disciplines, nor is it likely to be in ecology. So what else contributes to irreproducibility, and why suspect those factors exist in ecology and evolution?

Freedman and colleagues’ (2015) analysis of biomedical research estimated that around half (51%) of the irreproducible findings in the literature are the result of poor study design and inadequate data analysis and reporting. Other commentators have suggested that the contribution of inadequate data reporting to irreproducible results may be even higher than this (Ioannidis interviewed in Baker 2015). Reproducibility problems are most likely to emerge where institutionalized publication bias toward “significant” results is combined with a publish-or-perish research culture (Ioannidis 2005, Fanelli 2010a, Necker 2014). These conditions characterize ecology as much as they do biomedical and psychological research. Along with these other sciences, ecology also suffers from incomplete reporting of methods and results, and insufficient incentives to share materials, code, and data. Although this alone is not evidence of low reproducibility of ecological research (or a “reproducibility crisis” as the problem has been labeled in other disciplines), we believe it does constitute evidence that the discipline is *at risk* and that a systematic evaluation of the evidence base is worthwhile. In the following sections, we discuss the existing evidence that conditions of (a) publication bias, (b) questionable research practices in a publish-or-perish research culture, (c) incomplete reporting of methods and results, and (d) insufficient incentives for sharing materials, code, and data are all present in ecology, and we examine how they contribute to irreproducibility.

Publication bias. Over a decade ago, Jennions and Møller (2002) warned of widespread publication bias in ecology. Applying trim and fill assessments on 40 meta-analyses, they found that 38% of data sets (15 of 40) showed evidence of “missing” nonsignificant studies. Although 95% of meta-analyses showed statistically significant outcomes (38 of 40), after correcting for publication bias 15%–21% of those meta-analyses that originally showed statistically significant outcomes were no longer significant. Publications bias has been discussed by ecologists since then (e.g., Lortie et al. 2007), but more comprehensive and recent measures of the extent of the problem are needed.

In an unbiased literature, the proportion of significant studies should roughly match the average statistical power of the published research. When the proportion of significant studies in the literature exceeds the average power, bias is probably in play. Publication bias can result in a

Table 1. Existing estimates of the statistical power of ecology research.

		Power estimate for effect sizes (ES)		
Source	Research field	Small ES	Medium ES	Large ES
Parris and McCarthy (2001)	Effects of toe-clipping frogs (<10 studies)	6%–10%	8%–21%	15%–60%
Jennions and Møller (2003)	Behavioural Ecology (1362 tests from 697 articles in 10 journals)	13%–16%	40%–47%	65%–72%
Smith et al. (2011)	Animal Behaviour (278 tests in *Animal Behaviour*)	7%–8%	23%–26%	–

false positive error rate for the literature well beyond what is expected from the disclosed, accepted false positive rate (typically 5% in standard statistical tests), and it can result in the overestimation of effect sizes (Ioannidis 2005).

Fanelli (2010b, 2012) estimated that the proportion of "positive" results in the published environment or ecology literature was 74%. In the related field of plant and animal sciences, the estimated proportion was similar (78%). Both are well above the expected average statistical power of these fields, which the available evidence suggests is at best 40%–47% for medium effects (see table 1). This suggests an excess of statistical significance and therefore a higher-than-expected false-positive rate in the literature.

"Registered reports" offer an alternative to the traditional peer-review process, in which journals commit to a policy of undertaking peer review and making manuscript publication decisions on the basis of the introduction, method, and planned analysis sections alone, with actual results submitted later. Under this policy, reviewers and editors cannot be swayed by the significance or otherwise of results and must make their decisions on the basis of the study's rationale (i.e., how important is it to know the answer to this question?) and methods (i.e., is the proposed research design and analysis capable of answering the question?). Over 30 journals in different disciplines have now implemented registered reports in some form (*https://osf.io/8mpji/wiki/home*).

Questionable research practices in a "publish or perish" research culture. Statistically nonsignificant or "failed" studies used to be merely relegated to the file drawer (Rosenthal 1979). But in a publish-or-perish culture, these same studies are often resuscitated back to statistical significance through exercising "researcher degrees of freedom" (Simmons et al. 2011), also known as "Questionable Research Practices" (QRPs, see table 2; John et al. 2012). QRPs refer to activities such as *p*-hacking, cherry-picking, and hypothesizing after results are known (HARKing), which are well documented in fields such as psychology and medicine. Using computer simulations of experimental psychology data, Simmons and colleagues (2011) demonstrated how four common forms of undisclosed flexibility in choosing among dependent or response variables, sampling stopping rules, and reporting subsets of experimental conditions, can systematically inflate the false positive rate.

In a survey of over 2000 psychologists, 63% of the control group admitted that they have failed to report statistically nonsignificant dependent or response variables in their manuscripts, and 56% said they had checked the statistical significance of their results before deciding whether to collect more data (John et al. 2012). These percentages rose to 66% and 58%, respectively, in a group who were given incentives for honest reporting (John et al. 2012). In a survey of 426 economists, 38% admitted stopping a statistical analysis when they obtained a desired result, 36% admitted searching for control variables until they got the desired results, and 32% admitted presenting empirical findings selectively to confirm an argument (Necker 2014). Much smaller percentages of the respondents admitted to excluding data (e.g., outliers) without reporting it and using other tricks to increase t-value, R^2, or other statistics (3% and 7%, respectively). Evidence of widespread *p*-hacking in the medical literature has also been recently reported (Head et al. 2015). Fanelli (2012) reported a 22 percentage-point increase in the proportion of significant results between 1990 and 2007, which arguably corresponds to an increase in external pressure to publish (for grants and other forms of remuneration) over that time, and increased QRPs.

As the same hurdles and incentives based on publication and grant-funding track records exist in ecology and evolution as in other disciplines, there is every reason to expect that QRPs are widespread in this field too. Such biases may enter a research program insidiously, without any overt intent by the researcher to bias the outcomes. Some have even attained a general level of social acceptability among scientists, with, for example, many psychological researchers in John and colleagues' (2012) survey openly endorsing them. Some QRPs pertain only to research based on Frequentist statistical significance testing, but counterparts in other paradigms may also exist (see box 2).

Preregistration databases, which can perhaps be considered a precursor or alternative to registered reports, offer a repository in which researchers publicly commit to research questions, hypotheses or expectations, methods, and planned analysis *prior* to data collection and date-stamp this commitment (*https://cos.io/prereg*). Preregistration protocols can be applied broadly to all kinds of studies, not just those reliant on hypothesis testing, and have been strongly advocated in other disciplines as a strategy for immediately curbing QRPs, especially HARKing.

Incomplete reporting of methods and analysis. In addition to joint conditions of publication bias and publish or

Table 2. Questionable Research Practices (QRPs) that can inflate the false positive rate in the literature and result in less reproducible research (adapted from John et al. 2012).

p-hacking	• Checking the statistical significance of results before deciding whether to collect more data • Stopping data collection early because results reached statistical significance • Deciding whether to exclude data points (e.g., outliers) only after foreshadowing the impact on statistical significance and not reporting the impact of the data exclusion Rounding off a *p* value to meet a statistical significance threshold (e.g., presenting 0.053 as $p < .05$)
Cherry-picking	• Failing to report dependent or response variables or relationships that did not reach statistical significance or other threshold • Failing to report conditions or treatments that did not reach statistical significance or other threshold
HARKing (hypothesizing after the results are known)	Presenting a *post hoc* finding as though it had been hypothesized all along

perish, ecological research often lacks sufficient transparency around methods. Incomplete reporting manifests in many forms, and some of the more important omissions include failure to disclose the sample sizes for all treatment conditions and how those sample sizes were determined, the methods used to choose the subjects or allocate the treatments, whether blinding was used, the methods for handling missing data or censoring data, the full details of effect sizes or parameters and their corresponding measures of variation, and whether hypotheses were *post hoc* or *a priori*. Incomplete reporting poses a serious obstacle to reproducibility, limits the usefulness of data for meta-analysis, and can bias or weaken meta-analysis and systematic review processes (Koricheva et al. 2013).

Unfortunately, existing evidence suggests incomplete reporting is very common.

For example, less than 10% of articles in four leading ecology and conservation biology journals (see box 2 for more detail) report the statistical power or other sampling stopping rule of their research, despite the fact that close to 90% rely primarily on some form of statistical significance testing (Fidler et al. 2006, survey updated in 2010; see also Low-Décarie et al. 2014). Many articles also fail to report effect sizes or appropriate error bars (Anderson et al. 2001, Fidler et al. 2006). Data provided in supplemental materials or online repositories are often incorrectly or insufficiently described by the authors, with Gilbert and colleagues (2012) finding this to be the case in 35% of the 60 molecular ecology data sets they examined, making re-analysis difficult. Together, this evidence suggests that current journal guidelines, standards of enforcement, and incentives are insufficient for complete and transparent reporting of methods and analysis. The Tools for Transparency in Ecology and Evolution (TTEE) mentioned earlier, aim to support good practice by providing a checklist journals can provide to authors, reviewers and editors to facilitate compliance with transparent reporting. Simmons and colleagues (2012) suggested that a simple solution would be to require all methods sections to include and satisfy the following declaration: “We report how we determined our sample size, all data exclusions (if any), all manipulations, and all measures in the study.” However, we also acknowledge that some surveys of published research reported here are now several years old, and there may have been subsequent changes in practice. More recent surveys are therefore needed.

Insufficient incentives to share materials, data and, code. It is now widely acknowledged that lack of access to data, research materials (e.g., surveys containing the complete wording of questions and response scales), source code, or software is a fundamental obstacle to reproducing research and also to building on research in the future (Ince et al. 2012, Costello et al. 2013). Funding agencies such as the National Science Foundation (NSF) now require all submitted research proposals to include data-management plans that describe how research results and data will be disseminated and shared (*www.nsf.gov/bfa/dias/policy/dmp.jsp*). New data journals are emerging, such as *Nature’s Scientific Data* (*http://nature.com/sdata*), and services such as DataCite (*http://datacite.org*), figshare (*http://figshare.com*), the Dataverse Project (*http://dataverse.org*), and Dryad (*http://datadryad.org*) are making it easier to for researchers to archive, share and cite data sets. The growing popularity of code repositories such as GitHub (*https://github.com*) provides a powerful platform for researchers to efficiently collaborate, version and provide open access to their source code. When combined with digital repositories such as figshare and Zenodo (*https://zenodo.org*), data and code can be archived and assigned a license and a persistent digital object identifier (DOI), making it citeable, discoverable, and reuseable indefinitely (Mislan et al. 2016).

Many scientists who use computational methods are self-taught and often unaware of tools and best programming practices for writing versioned, reliable, efficient and maintainable code (Wilson et al. 2014) that aids reproducibility. This gap is being addressed by guidance on best practices in scientific computing, metadata are becoming more commonplace (e.g., Michener 2006, 2015, Sandve et al. 2013, Osborne et al. 2014, Wilson et al. 2014), and initiatives such as Software Carpentry (*http://software-carpentry.org*) teach skills in scientific computing via online resources and in-person workshops.

One indication of the future of this rapidly evolving area is the Open Science Framework (OSF; *http://osf.io*), maintained by the Center for Open Science (COS), which provides a platform to archive, share, preregister, and collaboratively

Box 2. Reproducibility beyond statistical significance testing.

Some of the problems we have discussed here are specific to null hypothesis significance testing (NHST)–based research, but we also stress that the reproducibility challenge applies more broadly. For example, some questionable research practices are specific to NHST research (e.g., *p*-hacking), but others are not (e.g., cherry picking and HARKing). Even in the former case, there may be parallel offences in other frameworks. For example, some have argued that Bayesian methods are also sensitive to undisclosed stopping rules and show error-rate inflation as a result of checking the data for some particular outcome and stopping once it has been found (Sanborn et al 2014, Yu 2014). Others have contested these findings and argue that optional stopping poses no risk within a Bayesian framework (Rouder 2014). The matter is far from resolved, and we urge users of Bayesian methods and other alternative modeling methods to consider and document reproducibility issues relevant to them.

Outside the domain of hypothesis testing (in either its Bayesian or Frequentist form), there are other types of reproducibility issues to consider. Conservation science, for example, can involve elements of decision theory, cost-effectiveness analysis, optimization, and scientific computing methods. Computational reproducibility (see box 1; Stodden 2015) of such research is equally crucial for detecting errors, testing software reliability, and verifying its fitness for reuse (Ince et al. 2012).

We have focused on highlighting reproducibility issues related to NHST in this article given its continued widespread use in ecology. In an unpublished update on our earlier survey (Fidler et al. 2006) of NHST use in the journals *Ecology*, *Journal of Ecology*, *Biological Conservation*, and *Conservation Biology*, we found little evidence of use waning. In 2005, 84% (n = 167 out of 200 articles) reported *p* values; in 2010, the corresponding figure was 90% (n = 153 out of 170 articles).

undertake research projects, as well as integrating with many existing services such as figshare and GitHub. The COS also provides a free online consulting service to support scientists in the use of tools, workflows and statistical methods to increase the reproducibility of their work (see *https://cos.io/stats_consulting*).

A number of journals now recognize the importance of preserving data and making them available for future use and promote public data archiving (PDA) with explicit policies such as the Joint Data Archiving Policy (JDAP; *http://datadryad.org/pages/jdap*). Journals that adopt this policy require as a condition for publication that the data, code, and other material used in a study be archived in an appropriate public repository, such as Dryad, figshare, GitHub, TreeBASE, GenBank, or the Open Science Framework (OSF).

However, although most journals now offer the option to upload supplemental material (which may include raw data, details of measurement materials and instruments, and source code or software), the uptake by authors is uneven. A survey of environmental biology publications produced from NSF-funded projects in the United States found that public data sharing was highest for genetic data (43% of publications) but very low (only 8%) for nongenetic ecological data (Hampton et al. 2013). This was attributed to different norms around PDA in fields that produce genetic data compared with fields that don't (Hampton et al. 2013).

In ecology and evolution, many data sets are collected at great effort over multiple locations and over a long period of time. Such data sets may have a "long shelf life" and may be used to test multiple hypotheses (Roche et al. 2014, p. 1). Data sharing and PDA provide many substantial benefits to the research and broader community (e.g., enabling data reuse, which improves the return per research dollar, and enabling errors to be detected and corrected) for just the modest cost of maintaining public repositories (Roche et al. 2014).

Although the benefits accrue to the community, the costs are seen to be borne by individual researchers: first in the loss of exclusive, priority access to data sets (which may be perceived as a loss of competitive advantage) and second in the significant additional effort required to archive data in a way that makes them properly reusable—that is to say complete, accompanied by adequate metadata, and preferably in both human- and machine-readable file formats (see, e.g., Michener 2006, 2015, Gilbert et al. 2012, Roche et al. 2015, Stodden 2015). This asymmetry in real or perceived costs and benefits to the community versus individual researchers creates understandable tensions regarding data sharing and PDA.

However, there is good evidence that mandating data archiving, such as requiring an explicit data-availability statement, vastly improves data availability almost 1000-fold compared with having no journal policy (Vines et al. 2013). Even voluntary "opt-in" incentives such as the Centre for Open Science's Open Data badge have resulted in a large increase of data availability in journals that have adopted this scheme (Kidwell et al. 2016).

Metaresearch projects for ecology and evolution

The primary goal of our article is to promote metaresearch in ecology and evolution, to systematically evaluate the evidence base of our discipline. Here, we propose four categories of metaresearch projects designed to take indicator measures of the likely reproducibility of published ecology and evolution research. These will be especially applicable in areas where direct or close replication is not feasible and/or which do not even necessarily rely on experimental data.

Re-analysis projects. When full direct replication is not possible (box 1), re-analysis often is. The simplest type of re-analysis is *computational reproducibility*, in which author-supplied data, analysis code and full details of platforms, required software versions, and auxiliary files are used to verify that results can be reproduced (Peng 2009, 2011). Given that scientific practices are increasingly reliant on computational software, tools (e.g., simulations and visualization techniques), and code, this basic form of re-analysis is worthy of attention. As the following examples from two different disciplines show, the expected outcomes of this seemingly straightforward form of re-analysis cannot be taken for granted.

In economics, Chang and Li (2015) attempted a computational re-analysis of 67 papers from 13 well-regarded journals, using author-supplied data and code replication files. Without author assistance, they were able to computationally reproduce key results of only 33% of the papers (22 of 67), with the success rate rising to 43% (29 of 67) when author assistance was sought. Reasons for replication failure included missing or incorrect data and code, missing software, and proprietary data.

Gilbert and colleagues (2012) re-analyzed data from 19 molecular ecology papers (containing 30 analyses) using the same freely available and widely used software program (STRUCTURE) that had been used in the original studies to infer genetic clustering and found that 30% of results could not be reproduced. They attribute this to a combination of inadequate analysis and reporting *and* inherently stochastic statistical methods. In practice, those two factors are very difficult to disentangle.

Conceptual re-analysis takes the same raw data as the starting point, but re-analysis may then employ a different statistical framework or different assumptions, methods, and models. An example of a conceptual re-analysis project comes once again from psychology. Silberzahn and Uhlmann (2015) recruited 29 teams of skilled researchers to address the same simple research question, “Are soccer referees more likely to give red cards to dark-skinned players?” using the exact same large data set compiled by a sports statistics firm across four major soccer leagues. Teams made their own decision about how best to analyze the data and reported what variables and models they used and why. The result was independent and diverse choices about what each team considered to be appropriate analysis, such as Bayesian cluster analysis, logistic regression, and linear modeling. Twenty of the 29 teams found a statistically significant correlation between skin color and red cards, but estimates of the effect ranged from a *slight negative* relationship to a *very large positive* relationship between the tendency for referees to give more red cards to dark-skinned players. When researchers were invited to critique the full results, some approaches were considered less defensible than others, but there was no consensus on what might constitute a single, best or correct approach. This study underlines the sensitivity of conclusions to subjective, but nevertheless justifiable, choices at the analysis stage of a study.

We suggest that many more re-analysis projects are needed in ecology, both *direct re-analysis* projects, which help highlight the impact of disclosed and undisclosed statistical assumptions and versions or editions of software among other things (see Peng 2011, White 2015), and *conceptual re-analysis* such as Silberzahn and Ulhmann’s (2015) crowd-sourced approach, which provided the opportunity to assess the extent of variability among analytic approaches and choices.

Computational reproducibility is a fixable problem, with a clear solution, and some journals have adopted policies to address it. In the journal *Biostatistics*, verification of the computational reproducibility of results by the associate editor for reproducibility (AER) is offered as an option for authors, and an article is kite-marked with an *R* if the AER is able to execute the code and data provided to produce the stated results that are claimed to be reproducible. The *American Journal of Political Science* applies a more stringent policy, requiring the submission of replication materials, with final acceptance of a manuscript contingent on successful replication of the results.

Conceptual re-analysis projects could be instigated and supported by journals, in which the highest-impact studies from the journal are selected for crowd-sourced re-analysis challenges. Re-analysis work could be done by interested volunteer researchers in exchange for publication of their results or as part of student group projects in methodology and statistics courses (Grahe et al 2014). This motivates researchers and students to conduct re-analysis studies and provides a published record of the results.

Quantifying publication bias. In table 1, we quoted estimates of statistical power (e.g., Jennions and Møller 2003, Smith 2011) and compared them to general counts of positive results from Fanelli (2010b, 2012). A more direct and compelling measure of publication bias would show the average statistical power of a sample of published research (e.g., all articles published in specified journals over a 12-month period) and the proportion of statistically significant studies *in that same sample*. However, as we also mentioned above, less than 10% published articles sampled in conservation biology and ecology journals report the statistical power of their own research (Fidler et al. 2006, updated survey in 2010). This means that power calculations will need to be done “from scratch” (based on expected, not obtained, effect sizes and relevant details reported in the study, such as degrees of freedom). This is a tractable but labor-intensive statistical exercise and will probably have more impact with endorsement from the editors of the journals being evaluated.

Measuring questionable research practices in ecology and evolution. As we discussed earlier, psychologists have developed methods to obtain honest survey responses to questions about undesirable activities, and these have successfully been deployed to measure levels of Questionable Research practices in Psychology itself (John et al. 2012). Similar surveys could be easily done in ecology and evolution. These measures would serve as indicators, because the reproducibility of a discipline's results will have an inverse relationship to the extent of QRPs among its researchers.

Assessing the completeness and transparency of methodological and statistical reporting in journals. We propose extensive journal surveys: systematically recording statistical practices and methodology descriptions in published journal articles (substantially extending and updating Fidler et al. 2006) and also documenting the sharing and reusability of materials, codes, and data. Incomplete reporting is a barrier not only to direct replication and meta-analysis but also to direct re-analysis projects (in which no new data are collected but a published study's data are subjected to independent statistical analysis following original protocols). Some aspects of statistical reporting accuracy can now be checked using automated procedures, such as statcheck (Nuijten et al. 2015). Such projects would help highlight the areas of journal's statistical reporting policies that are most in need of attention.

Conclusions

Research expenditure on irreproducible studies in preclinical biomedicine was recently estimated to be $28 billion per year in the United States alone (Freedman et al. 2015). Although the reproducibility rates in ecology are not currently known, if they were to approximate the rates in biomedicine or psychology, then we might expect that up to half of the current research expenditure in our own field has funded irreproducible research. The financial cost of these avoidable errors may be staggering, let alone the environmental costs. For these reasons, we believe the scientific community who undertake ecological research should urgently begin engaging in projects to evaluate the reproducibility of its evidence base.

In summary, we have argued that replication projects, such as the Reproducibility Project in Psychology, suit only some specific research areas of ecology. However, there are other tractable means of critically evaluating the remaining scientific base of the field. We have outlined four indicator measures of the likely reproducibility of results, and argued that such metaresearch projects will help us better understand the quality of current scientific evidence base.

Acknowledgments

Fiona Fidler is supported by an Australian Research Council Future Fellowship (FT150100297). Ascelin Gordon is supported an ARC Discovery Project (DP150103122). Yung En Chee is supported by the Melbourne Waterway Research–Practice Partnership. Bonnie Wintle is supported by the Templeton World Charity Foundation.

References cited

Anderson DR, Link WA, Johnson DH, Burnham KP. 2001. Suggestions for presenting the results of data analysis. Journal of Wildlife Management 65: 373–378.

Barker M. 2015. Irreproducible research costs put at $28 billion per year. Nature. 9 June. doi:10.1038/nature.2015.17711

Begley CG, Ellis LM. 2012. Raise standards for preclinical cancer research. Nature 483: 531–533.

Chang AC, Li P. 2015. Is Economics Research Replicable? Sixty Published Papers from Thirteen Journals Say "Usually Not." Finance and Economics Discussion Series. Board of Governors of the Federal Reserve System.

Costello MJ, Michener WK, Gahegan M, Zhang Z, Bourne PE. 2013. Biodiversity data should be published, cited, and peer reviewed. Trends in Ecology and Evolution 28: 454–461.

Ellison AM. 2010. Repeatability and transparency in ecological research. Ecology 91: 2536–2539.

Ellison AM, et al. 2006. Analytic web supports the synthesis of ecological data sets. Ecology 87: 1345–1358.

Fanelli D. 2010a. Do pressures to publish increase scientists' bias? An empirical support from US states data. PLOS ONE 5 (art. e10271).

———. 2010b. "Positive" results increase down the hierarchy of the sciences. PLOS ONE 5: e10068.

———. 2012. Negative results are disappearing from most disciplines and countries. Scientometrics 90: 891–904.

Fang FC, Steen RG, Casadevall A. 2013. Misconduct accounts for the majority of retracted scientific publications. Proceedings of the National Academy of Sciences 109: 17028–17033.

Fidler F, Burgman M, Cumming G, Buttrose R, Thomason N. 2006. Impact of criticism of null-hypothesis significance testing on statistical reporting practices in conservation biology. Conservation Biology 20: 1539–1544.

Freedman LP, Cockburn IM, Simcoe TE. 2015. The economics of reproducibility in preclinical research. PLOS Biology 13 (art. e1002165).

Gilbert KJ, et al. 2012. Recommendations for utilizing and reporting population genetic analyses: The reproducibility of genetic clustering using the program STRUCTURE. Molecular Ecology 21: 4925–4930.

Grahe J, Brandt M, IJerzman H, Cohoon J. 2014. Replication education. Association for Psychological Science Observer. (11 November 2016; *www.psychologicalscience.org/observer/replication-education*)

Hampton SE, et al. 2013. Big data and the future of ecology. Frontiers in Ecology and the Environment 11: 156–162.

Head ML, Holman L, Lanfear R, Kahn AT, Jennons MD. 2015. The extent and consequences of p-hacking in science. PLOS Biology 13: 1–15.

Ioannidis JP. 2005. Why most published research findings are false. PLOS Medicine 2: 696–701.

Ince DC, Hatton L, Graham-Cumming J. 2012. The case for open computer programs. Nature 482: 485–488.

Jennions MD, Møller AP. 2002. Publication bias in ecology and evolution: An empirical assessment using the "trim and fill" method. Biological Reviews of the Cambridge Philosophical Society 77: 211–22.

———. 2003. A survey of the statistical power of research in behavioral ecology and animal behavior. Behavioral Ecology 14: 438–445.

John LK, Loewenstein G, Prelec D. 2012. Measuring the prevalence of questionable research practices with incentives for truth telling. Psychological Science 23: 524–532.

Kidwell M, et al. 2016. Badges to acknowledge open practices: A simple, low cost, effective method for increasing transparency. PLOS Biology 14 (e1002456). (11 November 2016; *http://dx.doi.org/10.1371/journal.pbio.1002456*)

Koricheva J, Gurevitch J, Mengersen K. 2013. Handbook of Meta-Analysis in Ecology and Evolution. Princeton University Press.

Lortie CJ, Aarssen LW, Budden AE, Koricheva JR, Leimu R, Tregenza T. 2007. Publication bias and merit in ecology. Oikos 116: 1247–1253.

Low-Décarie E, Chivers C, Granados M. 2014. Rising complexity and falling explanatory power in ecology. Frontiers in Ecology and the Environment 12: 412–418.

McNutt M. 2014. Journals unite for reproducibility. Nature 515: 679.

Michener WK. 2006. Meta-information concepts for ecological data management. Ecological Informatics 1: 3–7.

———. 2015. Ecological data sharing. Ecological Informatics 29: 33–44.

Mislan KAS, Heer JM, White EP. 2016. Elevating the status of code in ecology. Trends in Ecology and Evolution 31: 4–7.

Necker S. 2014. Scientific misbehavior in economics. Research Policy 43: 1747–1759.

Nuijten MB, Hartgerink CHJ, van Assen MALM, Epskamp S, Wicherts JM. 2015. The prevalence of statistical reporting errors in psychology (1985–2013). Behavior Research Methods 48: 1–22. doi:10.3758/s13428-015-0664-2

Open Science Collaboration. 2015. Estimating the reproducibility of psychological science. Science 349 (art. aac4716). (11 November 2016; *http://dx.doi.org/10.1126/science*)

Oransky I. 2015. Half of biomedical studies don't stand up to scrutiny—and what we need to do about that. Conversation. (11 November 2016; *http://theconversation.com/half-of-biomedical-research-studies-dont-stand-up-to-scrutiny-and-what-we-need-to-do-about-that-45149*)

Osborne JM, et al. 2014. Ten simple rules for effective computational research. PLOS Computational Biology 10 (art. e1003506). doi:10.1371/journal.pcbi.1003506

Parker TH, Nakagawa S. 2014. Mitigating the epidemic of type I error: Ecology and evolution can learn from other disciplines. Frontiers in Ecology and Evolution 2: 1–3.

Parker TH, Nakagawa S, Gurevitch J, IIEE workshop participants. 2016a. Promoting transparency in evolutionary biology and ecology. Ecology Letters 19: 726–728. doi:10.1111/ele.12610

Parker TH, Forstmeier W, Koricheva J, Fidler F, Hadfield JD, Chee YE, Kelly CD, Gurevitch J, Nakagawa S. 2016b. Transparency in ecology and evolution: Real problems, real solutions. Trends in Ecology and Evolution 31: 711–719. doi:10.1016/j.tree.2016.07.002

Parker TH, Main E, Nakagawa S, Gurevitch J, Jarrad F, Burgman M. 2016c. Promoting transparency in conservation science. Conservation Biology 30: 1149–1150. doi:10.1111/cobi.12760

Parris KM, McCarthy MA. 2001. Identifying effects of toe clipping on anuran return rates: The importance of statistical power. Amphibia–Reptilia 22: 275–289.

Peng RD. 2011. Reproducible research in computation science. Science 334: 1226–1227.

Roche DG, Lanfear R, Binning SA, Haff TM, Schwanz LE, Cain KE, Kokko H, Jennions MD, Kruuk LEB. 2014. Troubleshooting public data archiving: Suggestions to increase participation. PLOS Biology 12 (art. e1001779).

Roche DG, Kruuk LEB, Lanfear R, Binning SA. 2015. Public data archiving in ecology and evolution: How well are we doing? PLOS Biology 13 (art. e1002295).

Rosenthal R. 1979. The file drawer problem and tolerance for null results. Psychological Bulletin 86: 638–641.

Rouder JN. 2014. Optional stopping: No problem for Bayesians. Psychonomic Bulletin and Review 21: 301–308.

Sanborn AN, Hills TT. 2014. The frequentist implications of optional stopping on Bayesian hypothesis tests. Psychonomic Bulletin and Review 21: 283–300.

Sandve GK, Nekrutenko A, Taylor J, Hovig E. 2013. Ten simple rules for reproducible computational research. PLOS Computational Biology 9: e1003285–e1003285.

Schmidt S. 2009. Shall we really do it again? The powerful concept of replication is neglected in the social sciences. Review of General Psychology 13: 90–100.

Schnitzer SA, Carson WP. 2016. Would ecology fail the repeatability test. BioScience 66: 98–99. doi:10.1093/biosci/biv176

Schooler JW. 2014. Metascience could rescue the "replication crisis." Nature 515: 9.

Silberzahn R, Uhlmann EL. 2015. Crowdsourced research: Many hands make tight work. Nature 526: 189–191. doi:10.1038/526189a

Simmons JP, Nelson LD, Simonsohn U. 2011. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. Psychological Science 22: 1359–1366.

Simmons JP, Nelson LD, Simonsohn U. 2012. A 21 Word Solution. Social Science Research Network. (11 November 2016; *http://dx.doi.org/10.2139/ssrn.2160588*)

Smith DR, Hardy ICW, Gammell MP. 2011. Power rangers: No improvement in the statistical power of analyses published in *Animal Behaviour*. Animal Behaviour 81: 347–352.

Stodden V. 2015. Reproducing statistical results. Annual Review of Statistics and Its Application 2: 1–19.

Vines TH, et al. 2013. Mandated data archiving greatly improves access to research data. FASEB Journal 27: 1304–1308.

White EP. 2015. Some thoughts on best publishing practices for scientific software. Ideas in Ecology and Evolution 8: 50–54.

Wilson G, et al. 2014. Best practices for scientific computing. PLOS Biology 12: 1–7.

Yu EC, Sprenger AM, Thomas RP, Dougherty MR. 2014. When decision heuristics and science collide. Psychonomic Bulletin and Review 21: 268–282.

Associate Professor Fiona Fidler (fidlerfm@unimelb.edu.au) holds a joint appointment in the School of BioSciences and the School of Historical and Philosophical Studies (History and Philosophy of Science Discipline) at the University of Melbourne, Australia; Fiona is interested in how scientists and experts make decisions. Bonnie C. Wintle is a postdoctoral fellow and Mark Burgman and Michael McCarthy are professors in the School of BioSciences at the University of Melbourne, Australia; they are interested in a broad range of topics related to environmental decisionmaking. Bonnie Wintle is now a research fellow at the Centre for Research in the Arts, Social Sciences and Humanities, University of Cambridge. Yung En Chee is a senior research fellow in the School of Ecosystem and Forest Sciences at the University of Melbourne, Australia; Yung applies ecological and decision-analytic theory and models to conservation problems. Ascelin Gordon is a senior research fellow in the Interdisciplinary Conservation Science Research Group in the School of Global, Urban, and Social Studies at RMIT University, in Melbourne, Australia; Ascelin is broadly interested in modeling approaches for understanding the impacts of environmental policies. FF, YC, BW, MB and MM were involved in discussion group about reproducibility and type 1 errors in ecology in 2014, which helped develop the outline for this article. AG and FF independently discussed the application of open science initiatives in ecology. FF wrote the first draft; YC wrote sections on data and code sharing with substantial input from AG. BW, MB, and MM made edits throughout.