Empowering peer reviewers with a checklist to improve transparency

Timothy H. Parker[1,2]
Simon C. Griffith[2]
Judith L. Bronstein[3]
Fiona Fidler[4,5]
Susan Foster[6]
Hannah Fraser[4]
Wolfgang Forstmeier[7]
Jessica Gurevitch[8]
Julia Koricheva[9]
Ralf Seppelt[10,11,12]
Morgan W. Tingley[13]
Shinichi Nakagawa[14]

[1] Department of Biology, Whitman College, Walla Walla, WA 99362, USA
[2] Department of Biological Sciences, Macquarie University, North Ryde, NSW 2109, Australia
[3] Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, Arizona 85721, USA
[4] School of BioSciences, University of Melbourne, Vic 3010, Australia
[5] History & Philosophy of Science, School of Historical & Philosophical Studies, University of Melbourne, Vic 3010, Australia
[6] Department of Biology, Clark University, Worcester, MA 01610-1477 USA
[7] Department of Behavioural Ecology and Evolutionary Genetics, Max Planck Institute for Ornithology, 82319 Seewiesen, Germany
[8] Department of Ecology and Evolution, Stony Brook University, Stony Brook, NY 11794-5245 USA
[9] School of Biological Sciences, Royal Holloway University of London, Egham, Surrey, TW20 0EX UK
[10] UFZ – Helmholtz Centre for Environmental Research, Department of Computational Landscape Ecology, 04318 Leipzig, Germany
[11] Martin-Luther-University Halle-Wittenberg, Institute of Geoscience and Geography, 06099 Halle (Saale), Germany
[12] iDiv – German Centre for Integrative Biodiversity Research Halle-Jena- Leipzig, Deutscher Platz 5e, 04103 Leipzig, Germany
[13] Department of Ecology and Evolutionary Biology, University of Connecticut, 75 N. Eagleville Rd U-3043, Storrs, CT 06269, USA
[14] Evolution & Ecology Research Centre, School of Biological, Earth and Environmental Sciences, University of New South Wales, Randwick, NSW 2052, Australia

Abstract

Peer review is widely considered fundamental to maintaining the rigour of science, but it often fails to ensure transparency and reduce bias in published papers, and this systematically weakens the quality of published inferences. In part, this is because many reviewers are unaware of important questions to ask with respect to the soundness of the design and analyses and the presentation of the methods and results; also some reviewers may expect others to be responsible for these tasks. We therefore present a reviewers' checklist of ten questions that address these critical components. Checklists are commonly used by practitioners of other complex tasks, and we see great potential for the wider adoption of checklists for peer review, especially to reduce bias and facilitate transparency in published papers. We expect that such checklists will be well-received by many reviewers.

Two important tasks facing peer reviewers are assessing the soundness of study design and evaluating the reporting of methods and results. Study soundness and reporting both bear directly on the reliability of the inferences that can be drawn from the papers that are ultimately published [1]. Other reviewing tasks include considering the placement of the study in a broader context, the writing, and the importance of the research, but these vary by journal and the expertise of the reviewer, and are often more subjective. We therefore focus on only the first two reviewing tasks. Our goal here is to explain particular components of this assessment process that we believe are too frequently ignored by peer reviewers, ultimately to the detriment of the scientific literature. We combine these components in a checklist that reviewers can use to improve transparency and reduce bias, and thus improve the reliability of scientific inferences.

We present this checklist as a series of ten questions (summarized in Table 1), each accompanied by suggestions for how the reviewer should proceed depending on the answer to that question. The checklist is not meant to be comprehensive. A longer checklist to help reviewers in ecology and evolutionary biology promote transparency was created as part of TTEE (Tools for Transparency in Ecology and Evolution; https://osf.io/y8aqx/) in an effort to help journals in ecology and evolutionary biology adopt TOP (Transparency and Openness Promotion) guidelines [2]. TTEE checklists, for both reviewers and authors, were designed to cover a broad swath of transparency issues. In contrast, the short checklist we present in this paper focusses on the subset of practices that we think are critically in need of improvement, and on which we think a concise checklist can achieve greatest impact. Our checklist provides reviewers with an efficient tool for promoting transparency in empirical research papers. Until now, no such tool has existed.

Why a checklist?

The use of checklists is well established among skilled practitioners working in complex systems. Checklists make flying complicated aircraft safer, they free architects to devote their mental energy to creativity, and they help surgeons focus on applying their skill without forgetting vital tasks [3,4]. Good checklists do not replace complex thought; they facilitate it. Of course, effective peer review requires expertise and critical thinking skills that no practical checklist can provide. However, this does not mean that checklists cannot be used to improve peer review, even dramatically, by calling attention to essential elements that are often overlooked.

Checklists can be of use to peer reviewers in two primary ways related to creating a more transparent and less biased literature: to help reviewers check (1) mundane but important details, and (2) both their own and the authors' potential biases. With regard to the first point, incomplete reporting of information hinders interpretation of studies and effective synthesis, and thus scientific progress [1,5]. We know from surveys of subsets of the ecology literature that approximately half of published papers omit important information such as sample size or variability associated with estimates [6-8]. Nearly all papers omitting this information were peer reviewed, suggesting that reviewers either overlooked these details, or felt that it was someone else's job to monitor them. Whether we notice omissions as reviewers depends on scrutiny that may vary unconsciously with factors such as whether we agree with the study's conclusions, our perception of the expertise of the authors, or whether we have used similar research designs. Regardless, the frequency of these omissions in the literature is evidence of a systematic problem, but one that could be resolved with the help of an appropriate checklist.

With regard to the second point, we need to explicitly address potential bias from authors and reviewers because all people, scientists included, are subject to biases that influence the information we

notice and how we interpret that information [9,10]. Such biases have been shown to have major impacts on the content of scientific papers [11-13], and so we expect that biases also influence the opinions we form when reviewing scientific papers. In fact, evidence suggests that peer review often suffers from a multitude of complex, systematic biases [14].

We hope that reviewers will find the questions in this checklist useful for most reviews. To facilitate the checklist's use, we provide some suggestions for reviewer responses to individual checklist questions, although we cannot provide a set of all possible answers to each checklist question. Occasionally reviewers will be uncertain about answers to one or more of these questions. Sometimes this uncertainty can be resolved by asking for additional information from the authors, and sometimes the reviewer should simply notify the editor so that she or he can seek additional reviewer expertise if needed. Of course some questions may not apply to some papers; it will be up to the reviewer to determine the relevance of each question. Determining its relevance may be aided by the explanation and justification that we provide following each particular checklist item.

The checklist

Questions to promote transparent reporting of methods and results

1. Were all sample sizes fully reported, including exact values for all subsets of data (e.g., each treatment group), and for all statistical analyses?
→If 'no', request that authors provide this information.

> Knowledge of sample size is essential for understanding the power of analyses (see below) and the reliability of estimates, and thus for interpreting results. It is also essential for later meta-analytic synthesis [5]. Yet, researchers fail to report sample sizes with troubling frequency [7,8]. Reporting a range (e.g. "9-12 replicates per treatment") is inadequate.

2. Are the methods for carrying out the study and analysing the results reported in sufficient detail to allow another researcher to gather the same data and run the identical analyses? When not in the paper itself, methodological details should be included in a supplement, or in many cases, archived in a publicly accessible and curated repository.
→If 'no', request that authors provide the relevant information.
→If you are uncertain about some aspect of the methods, state your uncertainty to the editor so that she or he can seek appropriate expertise as needed.

> By keeping replicability in mind while reading the methods, the reviewer can determine if methods have been reported in sufficient detail. Necessary details vary among studies with different methods. For instance, in the case of Bayesian analyses, authors should explicitly define their priors and report how their posterior distributions were derived, if applicable including Markov chain Monte Carlo specifications, and method of convergence (mixing) assessment. Archiving of details such as analysis code is essential if others are to understand how results were derived [15] (see also http://www.britishecologicalsociety.org/wp-content/uploads/2017/12/guide-to-reproducible-code.pdf) and, at least theoretically, be able to replicate the study, including the analyses. This information should be stored in curated archives. Temporary and uncurated repositories, including personal websites and the version-control site GitHub, are not viable for long-term storage. There will occasionally be valid justifications for not reporting certain information (e.g., population

locations for species threatened by illegal collection), but in most cases these exceptions should be explicitly addressed in the manuscript.

3. Are statistical results reported completely (considered in two parts below)?

3a. Are statistical results for each test reported in sufficient detail? What qualifies as 'sufficient detail' will differ among analyses. For most analyses, however, this will include (but not be limited to) basic parameter estimates of central tendency (e.g., means) or other basic estimates (e.g., regression or correlation coefficients) and variation (e.g., standard deviation) or associated estimates of uncertainty (e.g., confidence/credible intervals). For null hypothesis tests, reporting P-values and test statistics by themselves is almost always insufficient.
→If 'no', request that authors provide this information.
→If you are uncertain, state your uncertainty to the editor so that he or she can seek appropriate statistical expertise as needed. Remember that you may be the only reviewer looking carefully at this aspect of the manuscript.

3b. Are results from all variables and from all models reported? Complete reporting should include results related to all variables examined in preliminary models and all results from exploratory analyses. It will sometimes be appropriate to include these as supplementary materials. For analysis types that generate vast sets of results, it may be appropriate to place results in data archives.
→If 'no', request that authors provide this information.
→If you are uncertain, ask the authors to declare in the paper that all exploratory analyses are reported in full. We recommend using the 'Standard Reviewer Statement for Disclosure of Sample, Conditions, Measures, and Exclusions': "I request that the authors add a statement to the paper confirming whether, for all experiments, they have reported all measures, conditions, data exclusions, and how they determined their sample sizes. The authors should, of course, add any additional text to ensure the statement is accurate. This is the standard reviewer disclosure request endorsed by the Center for Open Science [see http://osf.io/hadz3]".

Insufficient reporting of results is one of the largest obstacles to an unbiased understanding of empirical progress [1,16]. Sometimes authors state that an analysis was conducted, but fail to provide all the relevant statistical outcomes such as slope estimates or estimates of variability [6-8,17]. At other times, authors conduct multiple analyses but do not explicitly acknowledge that they are reporting results from only a subset. Both practices may sometimes result from a direct request by the journal to shorten the text because of space limits or a desire for a concise story. Regardless, they weaken our ability to draw unbiased conclusions from the published literature. The failure to provide all relevant details from a reported analysis is often easily recognized by reviewers. In contrast, analyses that have been conducted but are completely unreported are more difficult, and sometimes even impossible, to recognize. However, there can be signs of unreported analyses: for instance different variables may be included in different models without obvious a priori justification, a subset of potential interactions may be provided without clear justification for the choice, or the authors may have failed to examine obvious predictions that are testable with available data. Each of these signs were found in a sample of literature in behavioural ecology, providing circumstantial evidence of unreported analyses [17]. Reviewers can prompt authors to include missing information in supplementary materials or in searchable, curated data archives. Asking authors to state whether all results from all analyses have been reported should lead authors to be more transparent about their exploratory work [18]. If necessary, authors should be directed to consult published recommendations regarding thorough reporting of results (and methodological

choices) from the type of analysis they have conducted [5]. Finally, it may help to remind authors that "not statistically significant" does not mean "not interesting or not important."

Questions to check biases of reviewers and authors

4. Were observers kept unaware of the experimental treatment imposed on the samples (e.g., organisms, plots) when recording observations or measurements so as to minimize unconscious bias?
→If not stated, then request clarification in the manuscript of whether methods were adopted that reduced the possibility of unconscious bias influencing observations.
→If no steps were taken to prevent observer bias, request that an explanation appear in the manuscript of how unconscious bias could have influenced observations.

It is now well demonstrated that researchers' observations are often influenced by what they expect to see [12,13]. For instance, when researchers were unaware of the colony of origin of the ants they were observing, they were > 3 times more likely to report aggression between colony mates than were researchers who knew the ants' colony of origin [12]. Keeping observers unaware of treatment categories or expected outcomes is not always possible or reasonable, but researchers should at least discuss the possibility of unconscious bias [19].

5. Did the authors explain how sample size was decided (e.g., based on a priori power analysis or logistical constraints), or when an experiment with pre-set sample sizes was terminated? If sample size or the end of the experiment was not decided prior to the initiation of the study, was there a decision rule for when to cease data collection?
→If not reported, request that authors provide this information.
→If the stopping rule included iterative statistical tests or examination of patterns as data accumulated, request that authors acknowledge the bias resulting from this process.

Cessation of data collection should never be made in response to reaching some threshold of statistical significance or effect. Such a practice leads to strong bias in favour of effects inflated by sampling error [20,21]. An explanation for the choice of stopping point should be provided (e.g., "we planned to harvest samples at the end of the second growing season").

6. Did the authors develop their analysis plan, including choices of variables, without looking at the data, for instance prior to gathering data or with a dummy data set? This is most easily determined by the existence of a pre-registered analysis plan. In the absence of pre-registration, a statement from the authors about the development of their analysis plan is still important.
→ If no, request that authors acknowledge the exploratory nature of their analyses and declare that they are reporting the complete set of results from all exploratory analyses.
→ If authors deviated from their analysis plan, request an explanation of why and how they deviated from the plan.

Choosing the analyses to present based on the strength of the effects derived from those analyses or models biases the distribution of presented results and can lead to presentation of entirely spurious relationships [20,22]. An ideal solution is to develop an analysis plan before examining the data and file it in a pre-registration archive such as offered by the Open Science Framework (https://cos.io/prereg/). One plausible alternative is an unusually detailed and publicly available grant proposal. Either way the pre-registration or proposal should be cited in the manuscript.

Researchers will sometimes need to deviate from pre-registered analysis plans. The pre-registration simply makes this transparent, and gives the reviewer, and later the reader, the opportunity to assess whether deviations were sufficiently justified. Regardless of the availability of an analysis plan, reporting all versions of all analyses is essential for avoiding bias.

7. How suitable do you find the research methods without considering the outcome? Evaluate the design and methods regardless of whether or not there was a finding of "statistical significance", or whether or not the results conform to a predicted pattern.
→If the methods appear to have been flawed, call attention to the problems and, if possible, recommend a better design. Deciding whether the problems with the methods are sufficient to justify a recommendation of rejection will require your expert judgement.
→ If uncertain about the suitability of some aspect of the methods, state your uncertainty to the editor so that she or he can seek appropriate methodological expertise as needed.

One driver of bias in the published literature is that we often evaluate the suitability of a study's methods based on the direction and strength of results [23]. This is especially true in cases of smaller samples or weaker study designs. In such cases, studies producing statistically significant or strong effects are sometimes incorrectly viewed as more plausible than those reporting weak or statistically non-significant results. There is a tendency among people we have talked with to assume that if a study found statistically significant results, sample sizes were sufficient or methodological weakness was not much of a problem. However, since 'strong' or 'significant' effects can often arise by chance [24] or be selected for reporting from among other unreported results [20,22], such results cannot be taken as proof that a study's methodological limitations were not a problem. Instead, the quality of the methods must be judged independent of the results. (Of course some studies include tests designed to assess a method's effectiveness rather than to assess the biological effect of primary interest, and those tests should be used to determine the quality of methods). Doubts about the reliability of the methods should be given equal strength regardless of the primary outcome.

8. Are the sample sizes large enough to justify the authors' conclusions? If presenting significance tests, how much power would this study have to detect statistically significant weak, moderate, and strong effects? (See Table 2 for examples of how sample size and effect size combine to determine power in two types of simple analyses.) Expectation of effect size can best be derived from average effect sizes presented in meta-analyses of similar topics. The effect size reported in the manuscript under review can be a poor estimate of the underlying effect size, especially if the sample size is small thus elevating sampling uncertainty. Statistical significance is a poor indicator of the reliability of an estimate across a wide range of sample sizes and common effect sizes (Table 2 provides insight into statistical power).
→If sample sizes are small in a system where effects are expected to be weak to moderate, request that authors avoid inferences based on threshold p-values, acknowledge uncertainty in effect size estimates, and acknowledge the need for further study.
→Do not use sample size as a criterion for recommending publication unless you do so regardless of study outcome (i.e., regardless of reported effect size and regardless of the outcomes of tests for significance).
→Do not use the failure to surpass a significance threshold as a reason to recommend rejection.

Presumably, nearly all ecologists and evolutionary biologists understand that there are problems with low power. However, it is clear that most of us would benefit from a reminder that type II error (false negatives) is only one of these problems. Because effect sizes are more variable with small

samples, inflated effect sizes are more likely, and thus large effects derived from small samples can be unreliable [25-27]. In fact, with low power caused by some combination of a small sample and relatively weak biological effect, studies are likely to reach statistical significance only if sampling error drives the observed effect size much higher than the true effect [25,27]. Unfortunately the weak to moderate-strength biological effects that contribute to low power are common in ecology and evolutionary biology, at least in some sub-disciplines [27,28]. However, biological effects can be larger in some types of studies and in some systems [27,29], and so what qualifies as a small sample in one study may be sufficiently large in another. Thus, evaluating sample size and power will benefit from knowledge of the typical effect sizes for the type of study in question, and this can most reliably be learned by consulting meta-analyses. If the study under review appears to have low power, we should not consider meeting a threshold p-value to be a reliable index of the validity of a pattern or a given effect size. In general, the reviewer should treat conclusions derived from low-powered studies as tentative, whether or not some significance threshold was met. However, studies with low power may often be worthy of publication, as some studies face major logistical obstacles regarding sample size, and it is only through publication and subsequent meta-analysis of a series of studies with small samples that we build a robust understanding of the true effect size [27].

9. What does the size of the estimated effect (e.g., slope, correlation coefficient, difference in means) suggest about its biological or practical importance, and what does uncertainty around that effect estimate suggest about the estimate's precision? Depending on the biological question, weak effects may be either biologically important or of limited interest; authors should justify their interpretation accordingly. Uncertainty around effects can be estimated with standard error (SE), 95% confidence intervals (approximately 2 x SE), or with other statistics. As sample size increases (see checklist question 8 above) and variance decreases, SE decreases and we gain confidence in the precision of the effect estimate.

→If the authors do not interpret their results in terms of the biological relevance of the effect and the uncertainty surrounding their effect estimate, request that they consider doing so.

Evaluating results based on the size of the effect estimate and the associated uncertainty rather than based on a p-value provides more direct insight into the biological phenomenon of interest [30]. Too often, interpretation of results focusses on statistical significance rather than on biological significance, and thus we can be led astray regarding our understanding of their relevance.

10. How unexpected would you judge these results to be in light of prior empirically derived understanding? Effects that are more surprising in light of robust prior information are those that had a lower prior probability of being correct. When testing unlikely hypotheses, the chance that a statistically significant result is a false positive rises dramatically (Table 3, Fig. 1). $P < 0.05$ is a poor threshold for evaluating the significance of an unexpected discovery and should be presented as no more than suggestive evidence for such discoveries.

→If a result is unexpected in light of prior evidence and is not supported by very strong new evidence (e.g., multiple lines of convincing evidence), do not recommend against publication on these grounds, but request that the authors acknowledge the tentative nature of their results.

Findings should be interpreted in light of previously published information, and the more robust the body of pre-existing information, the more caution authors should exercise when interpreting the implications of their contradictory results. To quote Carl Sagan, "Extraordinary claims require extraordinary evidence". For instance, many researchers in biology are unaware that the strength of evidence presented by a p-value depends on the prior probability of the outcome. When testing

moderately unlikely hypotheses (those with a 10% chance of being true) in a test with high statistical power, more than 1/3 of statistically 'significant' effects below the $p < 0.05$ threshold will be false positives (Table 3, Fig. 1)[21]. Thus, if robust pre-existing information makes a result unlikely, that result should be held to a higher standard of evidence than would be appropriate for a hypothesis that has already been empirically supported and thus has a higher prior probability [31]. For instance, a finding that parental diet influenced offspring phenotype is consistent with previously published findings and theory, but a finding that parental diet influenced grand-offspring phenotype more strongly than it influenced offspring phenotype would be extraordinary. Extraordinary results may be correct, but relative to results with a high prior probability, the extraordinary results are more likely to be false positives. We are not suggesting that reviewers estimate prior probabilities. However, a qualitative consideration of this issue is important for thoroughly evaluating the link between evidence and inference presented in a manuscript.

Conclusions

We have designed this checklist for the use of reviewers, but we also hope that editors and authors will find it useful. Of course, reviewers are also authors, and many editors are also researchers, and understanding of the issues raised here can contribute to excellence in scientific publication in ecology and evolution in many ways. Currently, a small number of journals where ecologists and evolutionary biologists publish have adopted checklists for authors that rigorously address some of the issues we raise here (e.g., Nature journals [https://www.nature.com/authors/policies/ReportingSummary.pdf]), Conservation Biology [https://mc.manuscriptcentral.com/societyimages/conbio/checklist_26.08.2016.docx]). These are important steps forward. As such checklists become more widespread, they should reduce the need for separate reviewer checklists. However, until rigorous author checklists designed to promote transparency and reduce bias are standard across journals, checklists such as this one will continue to play an important role. And even when author checklists become widespread, reviewers will still have an important function since, in their role as reviewers, they are not subject to the incentives that might lead authors or editors to publish biased subsets of results or be insufficiently transparent.

How will peer review checklists be received by peer reviewers? Journal editors often struggle to recruit the necessary two or three reviewers per manuscript, and so editors are legitimately reluctant to do anything that makes reviewing seem more burdensome. However, even if journal editors decide not to make review checklists mandatory, they can still make them readily available to reviewers. Our discussions with new reviewers (e.g. senior PhD students and post-docs) suggest that there is strong demand for this sort of guidance in peer reviewing papers.

Our checklist questions are practical tools. We hope that these questions, along with peer review checklists that address a broader set of topics (e.g., TTEE; https://osf.io/y8aqx/), will improve transparency in the published literature and thus reduce bias therein. More transparency and less bias should mean more reliable inferences in published papers and later in the meta-analyses based on those published papers[1]. As we improve peer review, we improve the quality of science.

References

1 Parker, T. H. *et al.* Transparency in ecology and evolution: real problems, real solutions. *Trends in Ecology & Evolution* **31**, 711-719, doi:10.1016/j.tree.2016.07.002 (2016).

2 TTEE_Working_Group. *Tools for Transparency in Ecology and Evolution (TTEE)*, <https://osf.io/g65cb/> (2016).
3 Arriaga, A. F. *et al.* Simulation-based trial of surgical-crisis checklists. *New England Journal of Medicine* **368**, 246-253, doi:10.1056/NEJMsa1204720 (2013).
4 Gawande, A. A. *The Checklist Manifesto: How to Get Things Right*. (Metropolitan Books, New York, 2009).
5 Gerstner, K. *et al.* Will your paper be used in a meta-analysis? Make the reach of your research broader and longer lasting. *Methods in Ecology and Evolution* **8**, 777-784, doi:10.1111/2041-210X.12758 (2017).
6 Ferreira, V. *et al.* A meta-analysis of the effects of nutrient enrichment on litter decomposition in streams. *Biological Reviews* **90**, 669-688, doi:10.1111/brv.12125 (2015).
7 Fidler, F., Burgman, M. A., Cumming, G., Buttrose, R. & Thomason, N. Impact of criticism of null-hypothesis significance testing on statistical reporting practices in conservation biology. *Conservation Biology* **20**, 1539-1544, doi:10.1111/j.1523-1739.2006.00525.x (2006).
8 Zhang, Y., Chen, H. Y. H. & Reich, P. B. Forest productivity increases with evenness, species richness and trait variation: a global meta-analysis. *Journal of Ecology* **100**, 742-749, doi:10.1111/j.1365-2745.2011.01944.x (2012).
9 Nickerson, R. S. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology* **2**, 175-220, doi:10.1037/1089-2680.2.2.175 (1998).
10 Fischhoff, B. Hindsight not equal to foresight – effect of outcome knowledge on judgment under uncertainty. *Journal of Experimental Psychology–Human Perception and Performance* **1**, 288–299 (1975).
11 Kozlov, M. V., Zverev, V. & Zvereva, E. L. Confirmation bias leads to overestimation of losses of woody plant foliage to insect herbivores in tropical regions. *PeerJ* **2**, e709, doi:10.7717/peerj.709 (2014).
12 van Wilgenburg, E. & Elgar, M. A. Confirmation bias in studies of nestmate recognition: a cautionary note for research into the behaviour of animals. *PLoS ONE* **8**, e53548, doi:10.1371/journal.pone.0053548 (2013).
13 Holman, L., Head, M. L., Lanfear, R. & Jennions, M. D. Evidence of experimental bias in the life sciences: why we need blind data recording. *PLoS Biol* **13**, e1002190, doi:10.1371/journal.pbio.1002190 (2015).
14 Lee, C. J., Sugimoto, C. R., Zhang, G. & Cronin, B. Bias in peer review. *Advances in Information Science* **64**, 2-17, doi:10.1002/asi.22784 (2013).
15 Mislan, K. A. S., Heer, J. M. & White, E. P. Elevating the status of code in ecology. *Trends in Ecology & Evolution* **31**, 4-7, doi:10.1016/j.tree.2015.11.006 (2016).
16 Fidler, F. *et al.* Metaresearch for evaluating reproducibility in ecology and evolution. *BioScience* **67**, 282-289, doi:10.1093/biosci/biw159 (2017).
17 Parker, T. H. 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. *Biological Reviews* **88**, 511-536 (2013).
18 Simmons, J. P., Nelson, L. D. & Simonsohn, U. A 21 word solution. *Dialogue: Newsletter of the SPSP* **26**, 4-7 (2012).
19 Kardish, M. R. *et al.* Blind trust in unblinded observation in ecology, evolution and behavior. *Frontiers in Ecology and Evolution* **3**, 51, doi:10.3389/fevo.2015.00051 (2015).
20 Simmons, J. P., Nelson, L. D. & Simonsohn, U. False positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science* **22**, 1359-1366 (2011).

21 Forstmeier, W., Wagenmakers, E.-J. & Parker, T. H. Detecting and avoiding likely false-positive findings – a practical guide. *Biological Reviews* **92**, 1941-1968, doi:10.1111/brv.12315 (2017).
22 Forstmeier, W. & Schielzeth, H. Cryptic multiple hypotheses testing in linear models: overestimated effect sizes and the winner's curse. *Behavioral Ecology and Sociobiology* **65**, 47-55, doi:10.1007/s00265-010-1038-5 (2011).
23 Palmer, A. R. Quasireplication and the contract of error: lessons from sex ratios, heritabilities and fluctuating asymmetry. *Annual Review of Ecology and Systematics* **31**, 441-480 (2000).
24 Halsey, L. G., Curran-Everett, D., Vowler, S. L. & Drummond, G. B. The fickle *P* value generates irreproducible results. *Nature Methods* **12**, 179, doi:10.1038/nmeth.3288 (2015).
25 Gelman, A. & Weakliem, D. Of beauty, sex, and power. *American Scientist* **97**, 310-316 (2009).
26 Barto, E. K. & Rillig, M. C. Dissemination biases in ecology: effect sizes matter more than quality. *Oikos* **121**, 228-235, doi:10.1111/j.1600-0706.2011.19401.x (2012).
27 Lemoine, N. P. *et al.* Underappreciated problems of low replication in ecological field studies. *Ecology* **97**, 2554-2561, doi:10.1002/ecy.1506 (2016).
28 Møller, A. P. & Jennions, M. D. How much variance can be explained by ecologists and evolutionary biologists? *Oecologia* **132**, 492-500 (2002).
29 Duffy, J. E., Godwin, C. M. & Cardinale, B. J. Biodiversity effects in the wild are common and as strong as key drivers of productivity. *Nature* **549**, 261-264, doi:10.1038/nature23886 (2017).
30 Nakagawa, S. & Cuthill, I. C. Effect size, confidence interval and statistical significance: a practical guide for biologists. *Biological Reviews* **82**, 591-605, doi:10.1111/j.1469-185X.2007.00027.x (2007).
31 Benjamin, D. J. *et al.* Redefine statistical significance *Nature Human Behaviour* **2**, 6-10, doi:10.1038/s41562-017-0189-z (2018).

Highlighted references:

Barto, E. K. & Rillig, M. C. Dissemination biases in ecology: effect sizes matter more than quality. Oikos 121, 228-235, doi:10.1111/j.1600-0706.2011.19401.x (2012).

Barto and Rillig provide evidence that various forms of bias, rather than concerns about data quality, have often influenced publication patterns in ecology.

Forstmeier, W., Wagenmakers, E.-J. & Parker, T. H. Detecting and avoiding likely false-positive findings – a practical guide. Biological Reviews 92, 1941-1968, doi:10.1111/brv.12315 (2017).

Forstmeier et al. present insights that can help reviewers recognize and guide authors away from potentially biased and unreliable reporting.

Lemoine, N. P. et al. Underappreciated problems of low replication in ecological field studies. Ecology 97, 2554-2561, doi:10.1002/ecy.1506 (2016).

Lemoine et al. discuss how bias can emerge from low powered studies, and also how bias can be avoided, even in systems where low power is inevitable due to logistical constraints.

475 Table 1. Concise version of ten questions reviewers can use to improve transparency and reduce bias in
476 the empirical literature. See the text for details.
477

Questions to promote transparent reporting of methods and results	
1	Were all sample sizes fully reported, including exact values for all subsets of data (e.g., each treatment group), and for all statistical analyses?
2	Are the methods reported in sufficient detail to allow another researcher to gather the same data and run the identical analyses?
3	Are statistical results reported completely (considered in two parts below)?
3a	Are statistical results for each test reported in sufficient detail? What qualifies as 'sufficient detail' will differ among analyses.
3b	Are results from all variables and from all models reported? Complete reporting should include results related to all variables examined in preliminary models and all results from exploratory analyses.
Questions to check biases of reviewers and authors	
4	Were observers kept unaware of the experimental treatment imposed on the samples (e.g., organisms, plots) when recording observations or measurements so as to minimize unconscious bias?
5	Did the authors explain how sample size was decided (e.g., based on a priori power analysis or logistical constraints), or when an experiment with pre-set sample sizes was terminated? If sample size or the end of the experiment was not decided prior to the initiation of the study, was there a decision rule for when to cease data collection?
6	Did the authors develop their analysis plan, including choices of variables, without looking at the data, for instance prior to gathering data or with a dummy data set? This is most easily determined by the existence of a pre-registered analysis plan. In the absence of pre-registration, a statement from the authors about the development of their analysis plan is still important.
7	How suitable do you find the research methods without considering the outcome? Evaluate the design and methods regardless of whether or not there was a finding of "statistical significance", or whether or not the results conform to a predicted pattern.
8	Are the sample sizes large enough to justify the authors' conclusions? If presenting significance tests, how much power would this study have to detect statistically significant weak, moderate, and strong effects? Expectation of effect size can best be derived from average effect sizes presented in meta-analyses of similar topics. The effect size reported in the manuscript under review can be a poor estimate of the underlying effect size, especially if the sample size is small, which elevates sampling uncertainty. Statistical significance is a poor indicator of the reliability of an estimate across a wide range of sample sizes and common effect sizes.
9	What does the size of the estimated effect (e.g., slope, correlation coefficient, difference in means) suggest about its biological or practical importance, and what does uncertainty around that effect estimate suggest about the estimate's precision?
10	How unexpected would you judge these results to be in light of prior empirically derived understanding? Effects that are more surprising in light of robust prior information are those that had a lower prior probability of being correct.

478
479

Table 2. Power to detect a true biological effect as statistically significant ($p < 0.05$) as a function of sample size and actual effect size for two types of simple analysis. High power is typically considered 0.8, or an 80% chance of detecting an effect (designated with * below) if the effect exists. Note that obtaining high power to detect small to medium effects (those most common in ecology and evolution) requires sample sizes much larger than are typical.

	effect size		sample size					
			10	20	50	100	200	500
correlation	r		power (to detect a true effect)					
	0.1	small	0.06	0.07	0.11	0.17	0.29	0.61
	0.3	medium	0.14	0.26	0.57	0.86*	>0.99*	>0.99*
	0.5	large	0.33	0.64	0.97*	>0.99*	>0.99*	>0.99*

			sample size (summed across both treatments in balanced design)					
			10	20	50	100	200	500
comparison of means (e.g., t-test)	Hedge's d		power (to detect a true effect)					
	0.2	small	0.06	0.07	0.11	0.17	0.29	0.61
	0.5	medium	0.11	0.18	0.41	0.7	0.94*	>0.99*
	0.8	large	0.2	0.4	0.79*	0.98*	>0.99*	>0.99*

Table 3. False positive report probability (the probability that a statistically significant result is a false positive – in other words, the probability that, in the case of a statistically significant rejection of the null, the null hypothesis is actually true) as a function of prior probability and statistical power. Note that for unlikely hypotheses, larger portions of statistically significant findings will be false positives. This table assumes a significance threshold of $p < 0.05$.

	power			
	0.1	0.2	0.5	0.8
prior	false positive report probability			
0.01	0.98	0.96	0.91	0.86
0.1	0.82	0.69	0.47	0.36
0.25	0.60	0.43	0.23	0.16
0.5	0.33	0.20	0.09	0.06
0.75	0.14	0.08	0.03	0.02

Figure 1. The relationship between prior probability, statistical power, and the false positive report probability. The false positive report probability is the probability of a statistically significant result being a false positive (in other words, the probability that, in the case of a statistically significant rejection of the null, the null hypothesis is actually true). Note that for unlikely hypotheses, large portions of statistically significant findings will be false positives even with high power. This figure is based on a significance threshold of $p < 0.05$.

Acknowledgements

We thank two anonymous reviewers as well as A. Moore and P. Goymer for suggestions that improved the manuscript.

Contributions

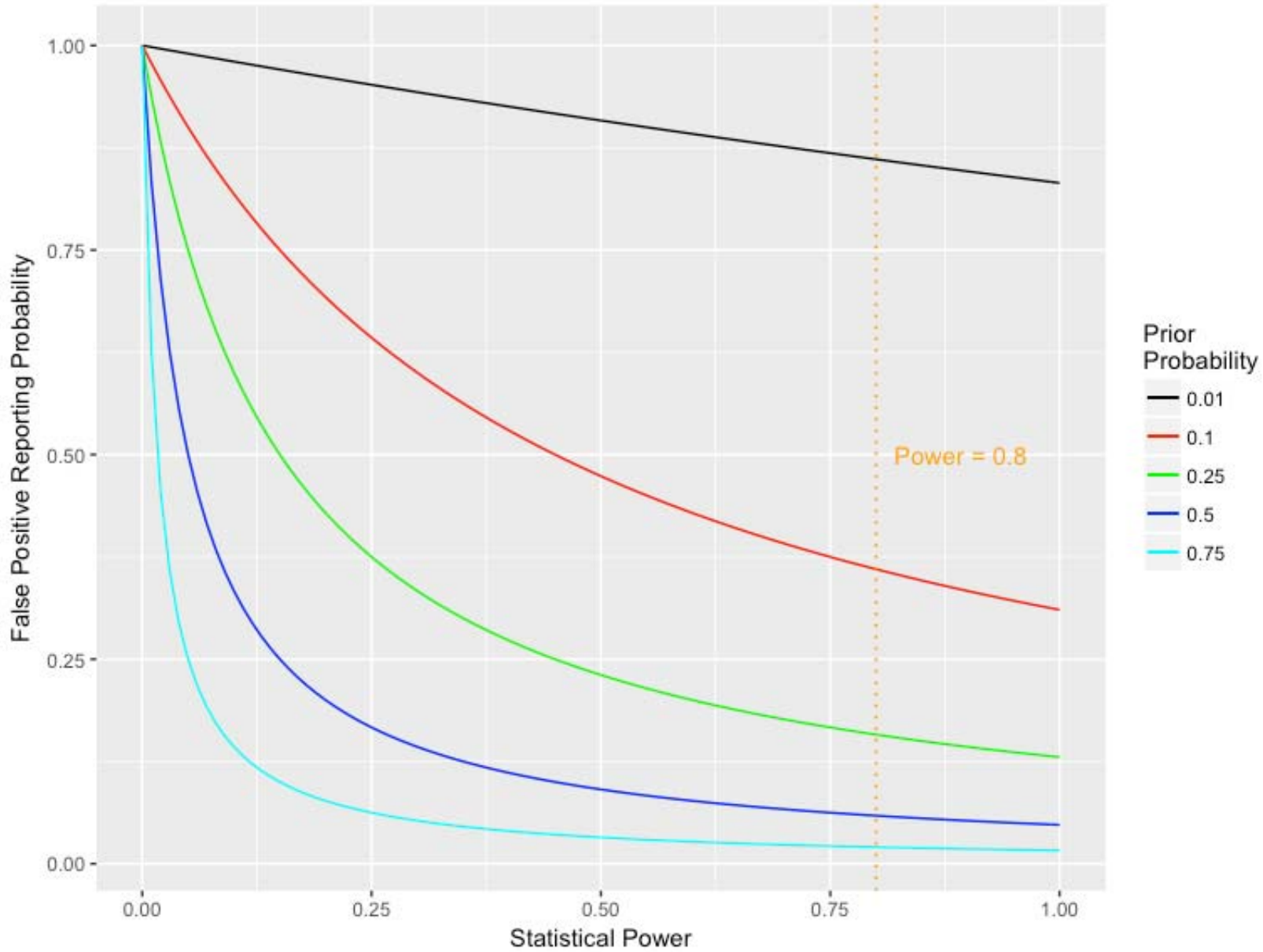
T.H.P. composed the original draft of this manuscript in consultation with S.C.G. and S.N. S.N. made the figure. The manuscript was edited substantially over multiple rounds with input from all co-authors.

Competing interests

The authors declare no competing financial interests.

Corresponding author

Correspondence to Timothy Parker.


1.00
0.75
0.50
0.25
0.00
False Positive Reporting Probability
0.00
0.25
0.50
0.75
1.00
Statistical Power
Power = 0.8
Prior Probability
0.01
0.1
0.25
0.5
0.75