Empowering peer reviewers to improve transparency

Timothy H. Parker[1,2]
Simon C. Griffith[2]
Judith 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 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] Ecology & 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

Peer review is widely considered fundamental to maintaining the rigour of science, but it is an imperfect process (Henderson, 2010; Jefferson et al., 2002). In that context, it is noteworthy that formal standards or guidelines for peer reviews themselves are rarely discussed in many disciplines, including ecology and evolutionary biology. Some may argue that a dearth of explicit guidelines is not a problem. After all, a tremendous amount of effective peer reviewing happens every day. However, there are reasons to expect that well-constructed guidelines in the form of checklists could be useful for improving certain aspects of peer review (Cobo et al., 2011; Cobo et al., 2007), such as promoting transparency of reviewed manuscripts, and that such checklists might be widely and enthusiastically adopted by many reviewers. Although some journals already provide checklists to reviewers, most of these checklists are quite limited in scope and do not substantially improve the rigor of the review process. There are also

guidelines that seek to explain the general process of peer review (Baier and Baker, 2013). Instead we propose a short list of important questions that reviewers can use to help authors produce more transparent and reliable manuscripts. We want to empower excellent peer review because it helps promote the production of high quality scientific publications.

Peer reviewers are typically expected to assess the soundness of study design and analyses and the presentation of methods and results, as well as the placement of the study in a broader context, the appropriateness of the writing, and the novelty and importance of the research. We focus on the first two issues because the others vary with journal and are often more subjective. To assess soundness, a useful review will evaluate the study itself on topics such as the experimental or observational methods, sample size, evidence of measurement reliability, choice of statistical approach, and plausibility of the assumptions that link the results to the conclusions. To complete this assessment, and to facilitate later interpretation of the study, a review will also insist that the manuscript contain sufficient details of methods and results. Our goal here is to identify particular components of this complex assessment that we believe are too frequently ignored by peer reviewers to the detriment of the published literature.

We root our argument in the concept of peer review as a complex undertaking. 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, precisely because peer review is complex. The use of checklists is well established among highly skilled practitioners working in complex systems. Checklists make flying complicated aircraft safe, they free architects to devote their mental energy to creativity, and they help surgeons focus on applying their skill without forgetting vital tasks (Arriaga et al., 2013; Gawande, 2009). Good checklists do not replace complex thought, they facilitate it.

Checklists would be of use to peer reviewers in two primary ways related to creating a more transparent and less biased literature: (1) to help reviewers check for mundane but important details, and (2) to help reviewers check both their own and the author's potential biases. Reviewers need to check for details because, mundane though they may be, such details are not trivial, and they can have major impacts on the interpretation of the study or its inclusion in later meta-analyses (Gerstner et al., 2017). A substantial obstacle to effective synthesis, and thus scientific progress, is incomplete and biased reporting of information (Parker et al., 2016). 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 (Ferreira et al., 2015; Fidler et al., 2006; Zhang et al., 2012). Most papers omitting this information were peer reviewed, and so reviewers must either not know any better or fail to notice. Whether we notice omissions as reviewers depends on scrutiny that may vary unconsciously with factors such as whether we agree with the study's conclusion 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.

We need to explicitly address potential bias from authors and reviewers because all people, we scientists included, are subject to biases that influence the information we notice and how we interpret that information (Fischhoff, 1975; Nickerson, 1998). Such biases have been shown to have major impacts on the information presented in scientific papers (Holman et al., 2015; Kozlov et al., 2014; van Wilgenburg and Elgar, 2013), and there is no reason to expect that biases do not 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 (Lee et al., 2013).

How would 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. In this scenario, those reviewers who wish to use checklists will do so, and those who find checklists burdensome will ignore them. Our discussions with PhD students and post-docs suggest that there is strong demand for this sort of guidance in peer reviewing papers, and the guidance of a checklist could make reviewing feel less burdensome and thus actually encourage more early-career researchers to engage in peer review. Most scientists never receive formal training in peer review and checklists can provide guidance as to the range of important issues to consider when evaluating a manuscript. Even if young scientists were the only ones to adopt these checklists, checklists could still have a major impact on the quality of reviews now, and an even larger impact in the future if the use of checklists became part of the culture of peer review.

Checklists for peer review already exist. For instance, TTEE (Tools for Transparency in Ecology and Evolution; https://osf.io/y8aqx/) was recently created to help journals in ecology and evolutionary biology adopt TOP (Transparency and Openness Promotion) guidelines and includes checklists for authors and for reviewers to help facilitate transparency in published work (TTEE_Working_Group, 2016) ; see also the Equator Network [http://www.equator-network.org/] which serves as a central location for reporting checklists in the health sciences and see the excellent checklist for authors submitting to Nature Ecology and Evolution [https://www.nature.com/authors/policies/ReportingSummary.pdf]). We encourage the use of the TTEE checklist in its entirety, and we include several TTEE questions here, but also novel items designed to help check cognitive bias and to call attention to important issues that are often overlooked by reviewers in ecology and evolution. Obviously, neither the TTEE checklist nor the checklist we present here is meant to address every consideration that a reviewer may have. It would be reasonable to include questions about, for instance, the validity of inferences, and some journals already ask reviewers such questions. We focus here on questions specifically related to *transparency* and *reducing bias* because we believe that these are issues that can be improved dramatically by the use of a relatively concise checklist.

We present below a short checklist of questions especially designed to promote transparency and reduce bias. Each question is accompanied by instructions for how to proceed depending on how the question is answered, and this is then followed by a more detailed explanation of and justification for that particular checklist item. Note that we exclude citations from the checklist questions themselves, but cite relevant sources in the explanations that follow. We encourage scientists to use any or all of these as well as other checklist questions (e.g., TTEE) as guides when conducting peer review, and we also encourage journal editors to bring these checklist questions to the attention of reviewers as tools to empower more effective reviewing.

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 (Gerstner et al., 2017), but researchers fail to report sample sizes with troubling frequency (Fidler et al., 2006; Zhang et al., 2012).

2. Are the methods reported in sufficient detail that would allow another researcher to gather the same data and run the identical analyses? Some methodological details, such as analysis code, should be archived in a publicly accessible and curated repository. 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.
→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 this simple question in mind while reading the methods, the reviewer can determine if methods have been reported in sufficient detail. Archiving of details such as analysis code is essential if others are to understand how results were derived (Mislan et al., 2016) and, at least theoretically, be able to replicate the study. 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 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, but for most analyses this includes (but is not limited to) basic parameter estimates of central tendency (e.g., means) or other basic estimates (e.g., regression or correlation coefficients) and variability (e.g., standard deviation) or associated estimates of uncertainty (e.g., confidence/credible intervals). For null hypothesis tests, P-values and test statistics by themselves are insufficient in most cases.
→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 some analysis types which generate vast sets of results, it may be appropriate to present quantitative summaries or 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 (Fidler et al., 2017; Parker et al., 2016). 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 (Ferreira et al., 2015; Fidler et al., 2006; Parker, 2013; Zhang et al., 2012). 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 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 first scenario is easy to recognize as a reviewer. The second scenario is more difficult, and sometimes even impossible, to recognize. However, there can be signs of unreported analyses, for instance different variables included in different models without obvious a priori justification, presentation of a subset of potential interactions without clear justification for the choice, or failure 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 (Parker, 2013). Authors can be prompted 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 (see wording for authors suggested in Simmons et al., 2012). It may help to be reminded that "not statistically significant" does not mean "not interesting or important."

Questions to check biases of reviewers and authors

4. Were observers blind to the experimental treatment imposed on the samples (e.g., organisms, plots) when recording observations or measurements?
→If not stated, then request clarification in the manuscript of the study's blinding practices.
→If no, request that an explanation of blinding practices appear in the manuscript.

It is now well demonstrated that the observations of researchers are often influenced by what they expect to see (Holman et al., 2015; van Wilgenburg and Elgar, 2013). For instance, when researchers were blinded to the colony of origin of the ants they were observing, they were > 3 times more likely to report aggression between colony mates than researchers who knew the ants' colony of origin (van Wilgenburg and Elgar, 2013). Thus, blinding observers to the expected observation reduces bias in that observation. Blinding is not always possible or reasonable, but researchers should at least address their blinding practices (Kardish et al., 2015).

5. Did the authors explain how sample size was decided (e.g., based on a priori power analysis or logistical constraints). If sample size was not decided prior to the initiation of the study, was there a decision rule for ceasing data collection?
→If not reported, request that authors provide this information.
→If 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 (Forstmeier et al., 2016; Simmons et al., 2011).

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, but 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 even generate entirely spurious relationships (Forstmeier and Schielzeth, 2011; Simmons et al., 2011). Thus either developing an analysis plan before examining the data (and ideally filing it in a pre-registration archive such as the Open Science Framework: https://cos.io/prereg/) or reporting all versions of exploratory analyses are essential for avoiding bias. Researchers will sometimes have to deviate from pre-registered analysis plans for various reasons, and the pre-registration simply makes this transparent, and gives the reviewer, and later the reader, the opportunity to assess whether deviation was sufficiently well justified to consider the analyses 'pre-registered'.

7. How suitable do you find the research methods and sample size without considering the results? Try this exercise: If the results are statistically significant, imagine how you would view the validity of the study if the results were not statistically significant. Alternatively, if the results were not statistically significant, imagine how you would view the validity of the study if the results were statistically significant.
→If the methods appear to have been unsuitable, call attention to the problems and make recommendations for an improved design. Deciding whether the problems with the methods are sufficient to justify a recommendation of rejection will require your expert judgement.

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 (Palmer, 2000). This is especially true in cases of smaller samples or weaker study designs. In such cases, studies producing statistically significant or strong effects may be 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, the quality of the methods must be judged independent of the results (though 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 (Table 1)? Unless the authors present evidence of strong effects in this or a similar system from robust prior studies, we should expect that most effects are weak to moderate. In the absence of such

evidence, if the study under review reports a strong effect based on a small sample, this effect is likely to be inflated due to sampling error. If the study under review reports a non-significant effect, keep in mind that studies across a wide range of sample sizes lack power to consistently detect statistical significance for effects of the typical size (Table 1 provides insight into what to consider a 'small' sample).

→If sample sizes are small, request that authors treat all results as preliminary and avoid inferences based on threshold p-values.

Table 1. Power to detect a true biological effect as statistically significant ($p < 0.05$) as a function of sample size and actual effect size. 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 effects of sizes typical (small to medium) 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*

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 are often unreliable (Barto and Rillig, 2012; Gelman and Weakliem, 2009; Lemoine et al., 2016). In fact, with low power studies are likely to reach statistical significance only if sampling error drives the observed effect size much higher than the true effect (Gelman and Weakliem, 2009; Lemoine et al., 2016). Further, we need to remember that power is also a function of the strength of the underlying biological effect, and many effects we study in ecology and evolutionary biology are weak to moderate (Lemoine et al., 2016; Møller and Jennions, 2002), though they can be larger in some types of studies (Duffy et al., 2017; Lemoine et al., 2016). Without good evidence to the contrary (such as effect sizes based on large samples derived from exploratory work in this system or average effect sizes from multiple well-designed experiments in similar systems) we should assume that studies are looking for effects that fall in this weak to moderate range. Thus unless we have good a priori evidence for a strong effect, we should typically not consider meeting a threshold p-value to be a reliable index of the validity of a pattern or a given effect size unless the sample size is sufficient to provide relatively high power to detect a relatively weak effect. In general the reviewer should be sceptical of studies with small samples, but scepticism should not translate to intolerance, 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 (Lemoine et al., 2016).

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 its precision? Depending on the biological question, weak effects may be biological important, or weak effects may be of limited interest, and authors should justify their interpretation accordingly. Effects should be considered unreliable if they are associated with substantial uncertainty. Uncertainty around effects is most commonly estimated as standard error (SE) or 95% confidence intervals (approximately 2 x SE). As sample size increases (see checklist question 8 above) and variance decreases, SE decreases and we gain confidence in the mean 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, request that they do so.

Evaluating results based on the size of the effect and the associated uncertainty rather than based on a p-value provides more direct insight in the biological phenomenon of interest (Nakagawa and Cuthill, 2007). 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.

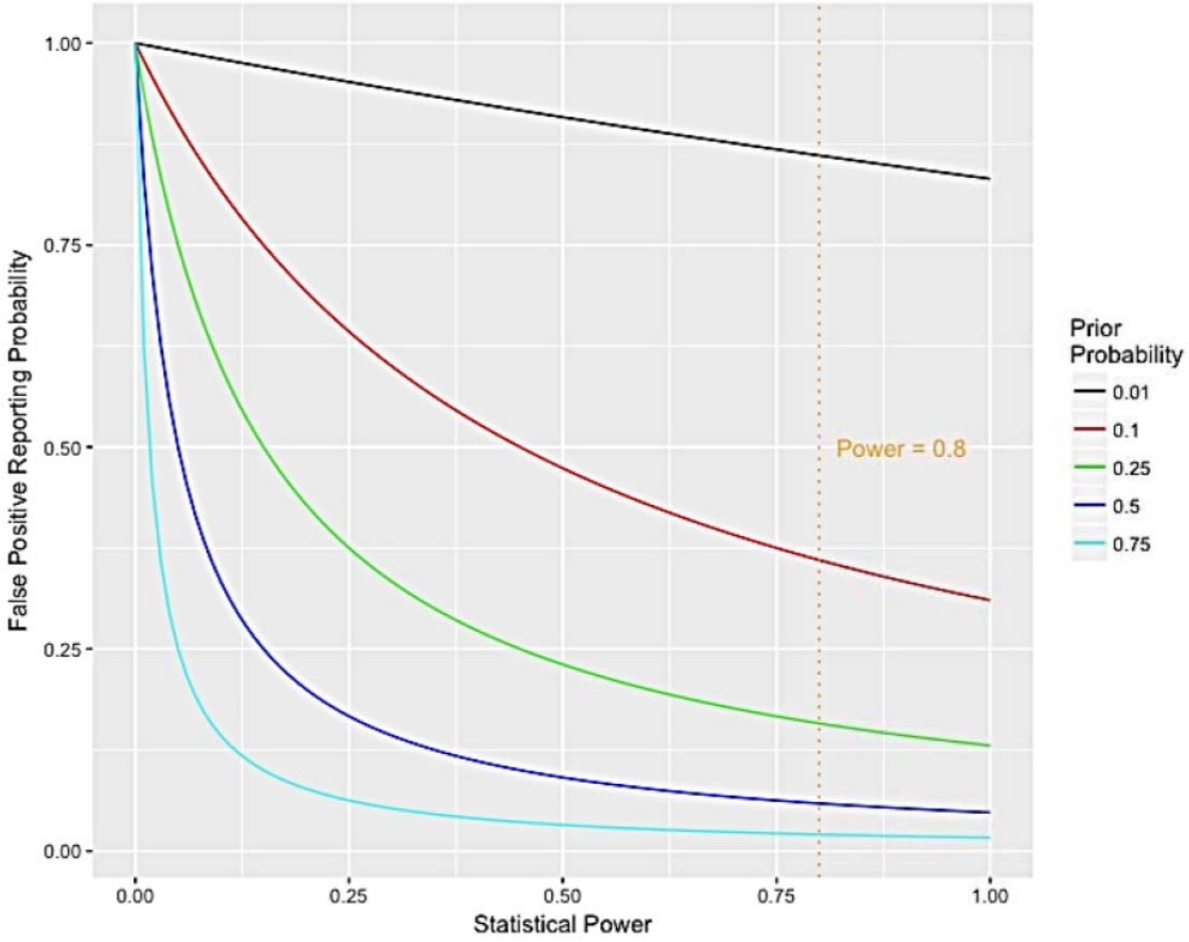
10. How unexpected would you judge these results to be in light of prior empirically derived understanding? Effects that are more surprising in light of robust prior information are those that had a lower prior probability of being correct. When testing unlikely hypotheses, the chance that a statistically significant result is a false positive rises dramatically (Table 2, Fig. 1). $P < 0.05$ is a poor threshold for evaluating the significance of an unexpected discovery and should not be presented as anything but suggestive evidence for such discoveries. To quote Carl Sagan, "Extraordinary claims require extraordinary evidence".
→If a result is unexpected in light of prior evidence and is not supported by very strong evidence (e.g., $0.05 > p > 0.005$), request that the authors acknowledge the tentative nature of their evidence.

Table 2. False positive report probability (the probability of a statistically significant result being a false positive) 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. 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 $p < 0.05$.



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 2, Fig. 1) (Forstmeier et al., 2016). 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. For instance, if using a null hypothesis test, a better significance threshold for results without prior support may be $p < 0.005$ (Benjamin et al., in press). Determining prior probability is an imperfect undertaking, and even experts can be deceived, for instance by bias in the existing literature (Palmer, 2000). However, considering this issue is important for thoroughly evaluating the evidence presented in a manuscript.

Conclusion

These checklist questions are meant to be practical tools for improving transparency and reducing bias, but this list is not comprehensive. Other peer review checklists include questions that address a broader sets of topics (e.g., TTEE; https://osf.io/y8aqx/), and we encourage reviewers to consult those lists as well. However, even if biologists (and researchers from any number of other disciplines) consult only a subset of the checklist questions embedded in this paper while reviewing manuscripts, we expect this will improve transparency in the published literature and thus reduce bias therein. Ideally, consulting these checklist questions will not only improve the individual manuscripts under review, but in the process will also help spread awareness of the issues addressed by these questions. Further, we hope that young biologists who use this checklist or other similar checklists will then adopt checklist questions as a useful tool, thus facilitating their integration into the culture of peer review. As we improve the culture of peer review, we improve the quality of science.

Literature Cited

Arriaga AF, Bader AM, Wong JM, Lipsitz SR, Berry WR, Ziewacz JE, Hepner DL, Boorman DJ, Pozner CN, Smink DS, Gawande AA, 2013. Simulation-based trial of surgical-crisis checklists. New England Journal of Medicine 368:246-253. doi: 10.1056/NEJMsa1204720.

Baier A, Baker L, 2013. A Guide to Peer Review in Ecology and Evolution. British Ecological Society.

Barto EK, Rillig MC, 2012. Dissemination biases in ecology: effect sizes matter more than quality. Oikos 121:228-235. doi: 10.1111/j.1600-0706.2011.19401.x.

Benjamin DJ, Berger JO, Johannesson M, Nosek BA, Wagenmakers E-J, Berk R, Bollen KA, Brembs B, Brown L, Camerer C, Cesarini D, Chambers CD, Clyde M, Cook TD, De Boeck P, Dienes Z, Dreber A, Easwaran K, Efferson C, Fehr E, F. F, Field AP, Forster M, George EI, Gonzalez R, Goodman S, Green E, Green DP, Greenwald A, Hadfield JD, Hedges LV, Held L, Ho T-H, Hoijtink H, Jones JH, Hruschka DJ, Imai K, Imbens G, Ioannidis JPA, Jeon M, Kirchler M, Laibson D, List J, Little R, Lupia A, Machery E, Maxwell SE, McCarthy M, Moore D, Morgan SL, Munafó M, Nakagawa S, Nyhan B, Parker TH, Pericchi L, Perugini M, Rouder J, Rousseau J, Savalei V, Schönbrodt FD, Sellke T, Sinclair B, Tingley D, Van Zandt T, Vazire S, Watts DJ, Winship C, Wolpert RL, Xie Y, Young C, Zinman J, Johnson VE, in press. Redefine statistical significance Nature Human Behaviour.

Cobo E, Cortés J, Ribera JM, Cardellach F, Selva-O'Callaghan A, Kostov B, García L, Cirugeda L, Altman DG, González JA, Sànchez JA, Miras F, Urrutia A, Fonollosa V, Rey-Joly C, Vilardell M, 2011. Effect of using reporting guidelines during peer review on quality of final manuscripts submitted to a biomedical journal: masked randomised trial. BMJ 343. doi: 10.1136/bmj.d6783.

Cobo E, Selva-O'Callagham A, Ribera J-M, Cardellach F, Dominguez R, Vilardell M, 2007. Statistical reviewers improve reporting in biomedical articles: a randomized trial. PLOS ONE 2:e332. doi: 10.1371/journal.pone.0000332.

Duffy JE, Godwin CM, Cardinale BJ, 2017. Biodiversity effects in the wild are common and as strong as key drivers of productivity. Nature 549:261-264. doi: 10.1038/nature23886.

Ferreira V, Castagneyrol B, Koricheva J, Gulis V, Chauvet E, Graça MAS, 2015. A meta-analysis of the effects of nutrient enrichment on litter decomposition in streams. Biological Reviews 90:669-688. doi: 10.1111/brv.12125.

Fidler F, Burgman MA, 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. doi: 10.1111/j.1523-1739.2006.00525.x.

Fidler F, Chee YE, Wintle BC, Burgman MA, McCarthy MA, Gordon A, 2017. Metaresearch for evaluating reproducibility in ecology and evolution. BioScience 67:282-289. doi: 10.1093/biosci/biw159.

Fischhoff B, 1975. Hindsight not equal to foresight – effect of outcome knowledge on judgment under uncertainty. Journal of Experimental Psychology–Human Perception and Performance 1:288–299.

Forstmeier W, Schielzeth H, 2011. 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.

Forstmeier W, Wagenmakers E-J, Parker TH, 2016. Detecting and avoiding likely false-positive findings – a practical guide. Biological Reviews. doi: 10.1111/brv.12315.

Gawande AA, 2009. The Checklist Manifesto: How to Get Things Right. New York: Metropolitan Books.

Gelman A, Weakliem D, 2009. Of beauty, sex, and power. American Scientist 97:310-316.

Gerstner K, Moreno-Mateos D, Gurevitch J, Beckmann M, Kambach S, Jones HP, Seppelt R, 2017. 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.

Henderson M, 2010. End of the peer review show? BMJ: British Medical Journal 340:738-740.

Holman L, Head ML, Lanfear R, Jennions MD, 2015. Evidence of experimental bias in the life sciences: why we need blind data recording. PLoS Biol 13:e1002190. doi: 10.1371/journal.pbio.1002190.

Jefferson T, Wager E, Davidoff F, 2002. Measuring the quality of editorial peer review. JAMA 287:2786-2790. doi: 10.1001/jama.287.21.2786.

Kardish MR, Mueller UG, Amador-Vargas S, Dietrich EI, Ma R, Barrett B, Fang C-C, 2015. Blind trust in unblinded observation in ecology, evolution and behavior. Frontiers in Ecology and Evolution 3:51. doi: 10.3389/fevo.2015.00051.

Kozlov MV, Zverev V, Zvereva EL, 2014. 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.

Lee CJ, Sugimoto CR, Zhang G, Cronin B, 2013. Bias in peer review. Advances in Information Science 64:2-17. doi: 10.1002/asi.22784.

Lemoine NP, Hoffman A, Felton AJ, Baur L, Chaves F, Gray J, Yu Q, Smith MD, 2016. Underappreciated problems of low replication in ecological field studies. Ecology 97:2554-2561. doi: 10.1002/ecy.1506.

Mislan KAS, Heer JM, White EP, 2016. Elevating the status of code in ecology. Trends in Ecology & Evolution 31:4-7. doi: 10.1016/j.tree.2015.11.006.

Møller AP, Jennions MD, 2002. How much variance can be explained by ecologists and evolutionary biologists? Oecologia 132:492-500.

Nakagawa S, Cuthill IC, 2007. 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.

Nickerson RS, 1998. Confirmation bias: A ubiquitous phenomenon in many guises. Review of General Psychology 2:175-220. doi: 10.1037/1089-2680.2.2.175.

Palmer AR, 2000. Quasireplication and the contract of error: lessons from sex ratios, heritabilities and fluctuating asymmetry. Annual Review of Ecology and Systematics 31:441-480.

Parker TH, 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. Biological Reviews 88:511-536.

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

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. Dialogue: Newsletter of the SPSP 26:4-7.

TTEE_Working_Group, 2016. Tools for Transparency in Ecology and Evolution (TTEE). Open Science Framework. doi: 10.17605/OSF.IO/G65CB.

van Wilgenburg E, Elgar MA, 2013. 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.

Zhang Y, Chen HYH, Reich PB, 2012. 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.