

1 Empowering peer reviewers to improve transparency

2

3 Timothy H. Parker^{1,2}

4 Simon C. Griffith²

5 Judith Bronstein³

6 Fiona Fidler^{4,5}

7 Susan Foster⁶

8 Hannah Fraser⁴

9 Wolfgang Forstmeier⁷

10 Jessica Gurevitch⁸

11 Julia Koricheva⁹

12 Ralf Seppelt^{10,11,12}

13 Morgan Tingley¹³

14 Shinichi Nakagawa¹⁴

15

16 ¹ Department of Biology, Whitman College, Walla Walla, WA 99362, USA

17 ² Department of Biological Sciences, Macquarie University, North Ryde, NSW 2109, Australia

18 ³ Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, Arizona 85721, USA

19 ⁴ School of BioSciences, University of Melbourne, Vic 3010, Australia

20 ⁵ History & Philosophy of Science, School of Historical & Philosophical Studies, University of Melbourne,
21 Vic 3010, Australia

22 ⁶ Department of Biology, Clark University, Worcester, MA 01610-1477 USA

23 ⁷ Department of Behavioural Ecology and Evolutionary Genetics, Max Planck Institute for Ornithology,
24 82319 Seewiesen, Germany

25 ⁸ Department of Ecology and Evolution, Stony Brook University, Stony Brook, NY 11794-5245 USA

26 ⁹ School of Biological Sciences, Royal Holloway University of London, Egham, Surrey, TW20 0EX UK

27 ¹⁰ UFZ – Helmholtz Centre for Environmental Research, Department of Computational Landscape
28 Ecology, 04318 Leipzig, Germany

29 ¹¹ Martin-Luther-University Halle-Wittenberg, Institute of Geoscience and Geography, 06099 Halle
30 (Saale), Germany

31 ¹² iDiv – German Centre for Integrative Biodiversity Research Halle-Jena- Leipzig, Deutscher Platz 5e,
32 04103 Leipzig, Germany

33 ¹³ Ecology & Evolutionary Biology, University of Connecticut, 75 N. Eagleville Rd U-3043, Storrs, CT
34 06269, USA

35 ¹⁴ Evolution & Ecology Research Centre, School of Biological, Earth and Environmental Sciences,
36 University of New South Wales, Randwick, NSW 2052, Australia

37

38

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

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

53
54 Peer reviewers are typically expected to assess the soundness of study design and analyses and the
55 presentation of methods and results, as well as the placement of the study in a broader context, the
56 appropriateness of the writing, and the novelty and importance of the research. We focus on the first
57 two issues because the others vary with journal and are often more subjective. To assess soundness, a
58 useful review will evaluate the study itself on topics such as the experimental or observational methods,
59 sample size, evidence of measurement reliability, choice of statistical approach, and plausibility of the
60 assumptions that link the results to the conclusions. To complete this assessment, and to facilitate later
61 interpretation of the study, a review will also insist that the manuscript contain sufficient details of
62 methods and results. Our goal here is to identify particular components of this complex assessment that
63 we believe are too frequently ignored by peer reviewers to the detriment of the published literature.

64
65 We root our argument in the concept of peer review as a complex undertaking. Effective peer review
66 requires expertise and critical thinking skills that no practical checklist can provide. However, this does
67 not mean that checklists cannot be used to improve peer review, even dramatically, precisely because
68 peer review is complex. The use of checklists is well established among highly skilled practitioners
69 working in complex systems. Checklists make flying complicated aircraft safe, they free architects to
70 devote their mental energy to creativity, and they help surgeons focus on applying their skill without
71 forgetting vital tasks (Arriaga et al., 2013; Gawande, 2009). Good checklists do not replace complex
72 thought, they facilitate it.

73
74 Checklists would be of use to peer reviewers in two primary ways related to creating a more transparent
75 and less biased literature: (1) to help reviewers check for mundane but important details, and (2) to help
76 reviewers check both their own and the author's potential biases. Reviewers need to check for details
77 because, mundane though they may be, such details are not trivial, and they can have major impacts on
78 the interpretation of the study or its inclusion in later meta-analyses (Gerstner et al., 2017). A
79 substantial obstacle to effective synthesis, and thus scientific progress, is incomplete and biased
80 reporting of information (Parker et al., 2016). We know from surveys of subsets of the ecology literature
81 that approximately half of published papers omit important information such as sample size or
82 variability associated with estimates (Ferreira et al., 2015; Fidler et al., 2006; Zhang et al., 2012). Most
83 papers omitting this information were peer reviewed, and so reviewers must either not know any better
84 or fail to notice. Whether we notice omissions as reviewers depends on scrutiny that may vary
85 unconsciously with factors such as whether we agree with the study's conclusion or whether we have
86 used similar research designs. Regardless, the frequency of these omissions in the literature is evidence
87 of a systematic problem but one that could be resolved with the help of an appropriate checklist.

88
89 We need to explicitly address potential bias from authors and reviewers because all people, we
90 scientists included, are subject to biases that influence the information we notice and how we interpret
91 that information (Fischhoff, 1975; Nickerson, 1998). Such biases have been shown to have major
92 impacts on the information presented in scientific papers (Holman et al., 2015; Kozlov et al., 2014; van
93 Wilgenburg and Elgar, 2013), and there is no reason to expect that biases do not also influence the
94 opinions we form when reviewing scientific papers. In fact, evidence suggests that peer review often
95 suffers from a multitude of complex, systematic biases (Lee et al., 2013).

96

97 How would peer review checklists be received by peer reviewers? Journal editors often struggle to
98 recruit the necessary two or three reviewers per manuscript, and so editors are legitimately reluctant to
99 do anything that makes reviewing seem more burdensome. However, even if journal editors decide not
100 to make review checklists mandatory, they can still make them readily available to reviewers. In this
101 scenario, those reviewers who wish to use checklists will do so, and those who find checklists
102 burdensome will ignore them. Our discussions with PhD students and post-docs suggest that there is
103 strong demand for this sort of guidance in peer reviewing papers, and the guidance of a checklist could
104 make reviewing feel less burdensome and thus actually encourage more early-career researchers to
105 engage in peer review. Most scientists never receive formal training in peer review and checklists can
106 provide guidance as to the range of important issues to consider when evaluating a manuscript. Even if
107 young scientists were the only ones to adopt these checklists, checklists could still have a major impact
108 on the quality of reviews now, and an even larger impact in the future if the use of checklists became
109 part of the culture of peer review.

110
111 Checklists for peer review already exist. For instance, TTEE (Tools for Transparency in Ecology and
112 Evolution; <https://osf.io/y8aqx/>) was recently created to help journals in ecology and evolutionary
113 biology adopt TOP (Transparency and Openness Promotion) guidelines and includes checklists for
114 authors and for reviewers to help facilitate transparency in published work (TTEE_Working_Group,
115 2016) ; see also the Equator Network [<http://www.equator-network.org/>] which serves as a central
116 location for reporting checklists in the health sciences and see the excellent checklist for authors
117 submitting to Nature Ecology and Evolution
118 [<https://www.nature.com/authors/policies/ReportingSummary.pdf>]). We encourage the use of the TTEE
119 checklist in its entirety, and we include several TTEE questions here, but also novel items designed to
120 help check cognitive bias and to call attention to important issues that are often overlooked by
121 reviewers in ecology and evolution. Obviously, neither the TTEE checklist nor the checklist we present
122 here is meant to address every consideration that a reviewer may have. It would be reasonable to
123 include questions about, for instance, the validity of inferences, and some journals already ask reviewers
124 such questions. We focus here on questions specifically related to *transparency* and *reducing bias*
125 because we believe that these are issues that can be improved dramatically by the use of a relatively
126 concise checklist.

127
128 We present below a short checklist of questions especially designed to promote transparency and
129 reduce bias. Each question is accompanied by instructions for how to proceed depending on how the
130 question is answered, and this is then followed by a more detailed explanation of and justification for
131 that particular checklist item. Note that we exclude citations from the checklist questions themselves,
132 but cite relevant sources in the explanations that follow. We encourage scientists to use any or all of
133 these as well as other checklist questions (e.g., TTEE) as guides when conducting peer review, and we
134 also encourage journal editors to bring these checklist questions to the attention of reviewers as tools to
135 empower more effective reviewing.

136 137 **Checklist**

138 139 **Questions to promote transparent reporting of methods and results**

140
141 1. Were all sample sizes fully reported including exact values for all subsets of data (e.g., each treatment
142 group), and for all statistical analyses?

143 →If 'no', request that authors provide this information.

144

145 Knowledge of sample size is essential for understanding the power of analyses (see below) and the
146 reliability of estimates, and thus for interpreting results. It is also essential for later meta-analytic
147 synthesis (Gerstner et al., 2017), but researchers fail to report sample sizes with troubling frequency
148 (Fidler et al., 2006; Zhang et al., 2012).

149
150 2. Are the methods reported in sufficient detail that would allow another researcher to gather the same
151 data and run the identical analyses? Some methodological details, such as analysis code, should be
152 archived in a publicly accessible and curated repository. Necessary details vary among studies with
153 different methods. For instance, in the case of Bayesian analyses, authors should explicitly define their
154 priors and report how their posterior distributions were derived, if applicable including Markov chain
155 Monte Carlo specifications, and method of convergence (mixing) assessment.

156 →If 'no', request that authors provide the relevant information.

157 →If you are uncertain about some aspect of the methods, state your uncertainty to the editor so that
158 she or he can seek appropriate expertise as needed.

159
160 By keeping this simple question in mind while reading the methods, the reviewer can determine if
161 methods have been reported in sufficient detail. Archiving of details such as analysis code is
162 essential if others are to understand how results were derived (Mislán et al., 2016) and, at least
163 theoretically, be able to replicate the study. This information should be stored in curated archives.
164 Temporary and uncurated repositories, including personal websites and the version-control site
165 GitHub, are not viable for long-term storage. There will occasionally be valid justifications for not
166 reporting certain information (e.g., population locations for species threatened by illegal collection),
167 but these exceptions should be explicitly addressed in the manuscript.

168
169 3. Are statistical results reported completely (considered in two parts below)?

170
171 3a. Are statistical results for each test reported in sufficient detail? What qualifies as 'sufficient detail'
172 will differ among analyses, but for most analyses this includes (but is not limited to) basic parameter
173 estimates of central tendency (e.g., means) or other basic estimates (e.g., regression or correlation
174 coefficients) and variability (e.g., standard deviation) or associated estimates of uncertainty (e.g.,
175 confidence/credible intervals). For null hypothesis tests, P-values and test statistics by themselves are
176 insufficient in most cases.

177 →If 'no', request that authors provide this information.

178 →If you are uncertain, state your uncertainty to the editor so that he or she can seek appropriate
179 statistical expertise as needed. Remember that you may be the only reviewer looking carefully at this
180 aspect of the manuscript.

181
182 3b. Are results from all variables and from all models reported? Complete reporting should include
183 results related to all variables examined in preliminary models and all results from exploratory analyses.
184 It will sometimes be appropriate to include these as supplementary materials. For some analysis types
185 which generate vast sets of results, it may be appropriate to present quantitative summaries or to place
186 results in data archives.

187 →If 'no', request that authors provide this information.

188 →If you are uncertain, ask the authors to declare in the paper that all exploratory analyses are reported
189 in full. We recommend using the 'Standard Reviewer Statement for Disclosure of Sample, Conditions,
190 Measures, and Exclusions': "I request that the authors add a statement to the paper confirming
191 whether, for all experiments, they have reported all measures, conditions, data exclusions, and how
192 they determined their sample sizes. The authors should, of course, add any additional text to ensure the

193 statement is accurate. This is the standard reviewer disclosure request endorsed by the Center for Open
194 Science [see <http://osf.io/hadz3>].

195

196 Insufficient reporting of results is one of the largest obstacles to an unbiased understanding of
197 empirical progress (Fidler et al., 2017; Parker et al., 2016). Sometimes authors state that an analysis
198 was conducted, but fail to provide all the relevant statistical outcomes such as slope estimates or
199 estimates of variability (Ferreira et al., 2015; Fidler et al., 2006; Parker, 2013; Zhang et al., 2012). At
200 other times, authors conduct multiple analyses but do not explicitly acknowledge that they are
201 reporting results from only a subset. Both practices may sometimes result from a direct request by
202 the journal because of space limits or a desire for a concise story. Regardless, they weaken our
203 ability to draw unbiased conclusions from the published literature. The first scenario is easy to
204 recognize as a reviewer. The second scenario is more difficult, and sometimes even impossible, to
205 recognize. However, there can be signs of unreported analyses, for instance different variables
206 included in different models without obvious a priori justification, presentation of a subset of
207 potential interactions without clear justification for the choice, or failure to examine obvious
208 predictions that are testable with available data. Each of these signs were found in a sample of
209 literature in behavioural ecology, providing circumstantial evidence of unreported analyses (Parker,
210 2013). Authors can be prompted to include missing information in supplementary materials or in
211 searchable curated data archives. Asking authors to state whether all results from all analyses have
212 been reported should lead authors to be more transparent about their exploratory work (see
213 wording for authors suggested in Simmons et al., 2012). It may help to be reminded that “not
214 statistically significant” does not mean “not interesting or important.”

215

216

217 **Questions to check biases of reviewers and authors**

218

219 4. Were observers blind to the experimental treatment imposed on the samples (e.g., organisms, plots)
220 when recording observations or measurements?

221 →If not stated, then request clarification in the manuscript of the study’s blinding practices.

222 →If no, request that an explanation of blinding practices appear in the manuscript.

223

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

231

232 5. Did the authors explain how sample size was decided (e.g., based on a priori power analysis or
233 logistical constraints). If sample size was not decided prior to the initiation of the study, was there a
234 decision rule for ceasing data collection?

235 →If not reported, request that authors provide this information.

236 →If stopping rule included iterative statistical tests or examination of patterns as data accumulated,
237 request that authors acknowledge the bias resulting from this process.

238

239 Cessation of data collection should never be made in response to reaching some threshold of
240 statistical significance or effect. Such a practice leads to strong bias in favour of effects inflated by
241 sampling error (Forstmeier et al., 2016; Simmons et al., 2011).
242

243 6. Did the authors develop their analysis plan, including choices of variables, without looking at the data,
244 for instance prior to gathering data or with a dummy data set? This is most easily determined by the
245 existence of a pre-registered analysis plan, but in the absence of pre-registration, a statement from the
246 authors about the development of their analysis plan is still important.

247 → If no, request that authors acknowledge the exploratory nature of their analyses and declare that
248 they are reporting the complete set of results from all exploratory analyses

249 → If authors deviated from their analysis plan, request an explanation of why and how they deviated
250 from the plan

251
252 Choosing the analyses to present based on the strength of the effects derived from those analyses
253 or models biases the distribution of presented results and can even generate entirely spurious
254 relationships (Forstmeier and Schielzeth, 2011; Simmons et al., 2011). Thus either developing an
255 analysis plan before examining the data (and ideally filing it in a pre-registration archive such as the
256 Open Science Framework: <https://cos.io/prereg/>) or reporting all versions of exploratory analyses
257 are essential for avoiding bias. Researchers will sometimes have to deviate from pre-registered
258 analysis plans for various reasons, and the pre-registration simply makes this transparent, and gives
259 the reviewer, and later the reader, the opportunity to assess whether deviation was sufficiently well
260 justified to consider the analyses 'pre-registered'.
261

262 7. How suitable do you find the research methods and sample size without considering the results?
263 Try this exercise: If the results are statistically significant, imagine how you would view the validity of the
264 study if the results were not statistically significant. Alternatively, if the results were not statistically
265 significant, imagine how you would view the validity of the study if the results were statistically
266 significant.

267 → If the methods appear to have been unsuitable, call attention to the problems and make
268 recommendations for an improved design. Deciding whether the problems with the methods are
269 sufficient to justify a recommendation of rejection will require your expert judgement.
270

271 One driver of bias in the published literature is that we often evaluate the suitability of a study's
272 methods based on the direction and strength of results (Palmer, 2000). This is especially true in
273 cases of smaller samples or weaker study designs. In such cases, studies producing statistically
274 significant or strong effects may be viewed as more plausible than those reporting weak or
275 statistically non-significant results. There is a tendency among people we have talked with to
276 assume that if a study found statistically significant results, sample sizes were sufficient or
277 methodological weakness was not much of a problem. However, the quality of the methods must be
278 judged independent of the results (though of course some studies include tests designed to assess a
279 method's effectiveness rather than to assess the biological effect of primary interest, and those
280 tests should be used to determine the quality of methods). Doubts about the reliability of the
281 methods should be given equal strength regardless of the primary outcome.
282

283 8. Are the sample sizes large enough to justify the authors' conclusions? If presenting significance tests,
284 how much power would this study have to detect statistically significant weak, moderate, and strong
285 effects (Table 1)? Unless the authors present evidence of strong effects in this or a similar system from
286 robust prior studies, we should expect that most effects are weak to moderate. In the absence of such

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

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

294
 295 Table 1. Power to detect a true biological effect as statistically significant ($p < 0.05$) as a function of
 296 sample size and actual effect size. High power is typically considered 0.8, or an 80% chance of detecting
 297 an effect (designated with * below) if the effect exists. Note that obtaining high power to detect effects
 298 of sizes typical (small to medium) in ecology and evolution requires sample sizes much larger than are
 299 typical.

300

		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	Hedge's d	power (to detect a true effect)							
of means	0.2	small	0.06	0.07	0.11	0.17	0.29	0.61	
(e.g., t-test)	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*	

301 Presumably nearly all ecologists and evolutionary biologists understand that there are problems
 302 with low power. However, it is clear that most of us would benefit from a reminder that type II error
 303 (false negatives) is only one of these problems. Because effect sizes are more variable with small
 304 samples, inflated effect sizes are more likely, and thus large effects derived from small samples are
 305 often unreliable (Barto and Rillig, 2012; Gelman and Weakliem, 2009; Lemoine et al., 2016). In fact,
 306 with low power studies are likely to reach statistical significance only if sampling error drives the
 307 observed effect size much higher than the true effect (Gelman and Weakliem, 2009; Lemoine et al.,
 308 2016). Further, we need to remember that power is also a function of the strength of the underlying
 309 biological effect, and many effects we study in ecology and evolutionary biology are weak to
 310 moderate (Lemoine et al., 2016; Møller and Jennions, 2002), though they can be larger in some
 311 types of studies (Duffy et al., 2017; Lemoine et al., 2016). Without good evidence to the contrary
 312 (such as effect sizes based on large samples derived from exploratory work in this system or average
 313 effect sizes from multiple well-designed experiments in similar systems) we should assume that
 314 studies are looking for effects that fall in this weak to moderate range. Thus unless we have good a
 315 priori evidence for a strong effect, we should typically not consider meeting a threshold p-value to
 316 be a reliable index of the validity of a pattern or a given effect size unless the sample size is sufficient
 317 to provide relatively high power to detect a relatively weak effect. In general the reviewer should be
 318 sceptical of studies with small samples, but scepticism should not translate to intolerance, as some
 319

320 studies face major logistical obstacles regarding sample size, and it is only through publication and
321 subsequent meta-analysis of a series of studies with small samples that we build a robust
322 understanding of the true effect size (Lemoine et al., 2016).

323
324 9. What does the size of the estimated effect (e.g., slope, correlation coefficient, difference in means)
325 suggest about its biological or practical importance and what does uncertainty around that effect
326 estimate suggest about its precision? Depending on the biological question, weak effects may be
327 biological important, or weak effects may be of limited interest, and authors should justify their
328 interpretation accordingly. Effects should be considered unreliable if they are associated with
329 substantial uncertainty. Uncertainty around effects is most commonly estimated as standard error (SE)
330 or 95% confidence intervals (approximately 2 x SE). As sample size increases (see checklist question 8
331 above) and variance decreases, SE decreases and we gain confidence in the mean effect estimate.
332 →If the authors do not interpret their results in terms of the biological relevance of the effect and the
333 uncertainty surrounding their effect, request that they do so.

334
335 Evaluating results based on the size of the effect and the associated uncertainty rather than based
336 on a p-value provides more direct insight in the biological phenomenon of interest (Nakagawa and
337 Cuthill, 2007). Too often interpretation of results focusses on statistical significance rather than on
338 biological significance, and thus we can be led astray regarding our understanding of their relevance.

339
340 10. How unexpected would you judge these results to be in light of prior empirically derived
341 understanding? Effects that are more surprising in light of robust prior information are those that had a
342 lower prior probability of being correct. When testing unlikely hypotheses, the chance that a statistically
343 significant result is a false positive rises dramatically (Table 2, Fig. 1). $P < 0.05$ is a poor threshold for
344 evaluating the significance of an unexpected discovery and should not be presented as anything but
345 suggestive evidence for such discoveries. To quote Carl Sagan, “Extraordinary claims require
346 extraordinary evidence”.

347 →If a result is unexpected in light of prior evidence and is not supported by very strong evidence (e.g.,
348 $0.05 > p > 0.005$), request that the authors acknowledge the tentative nature of their evidence.

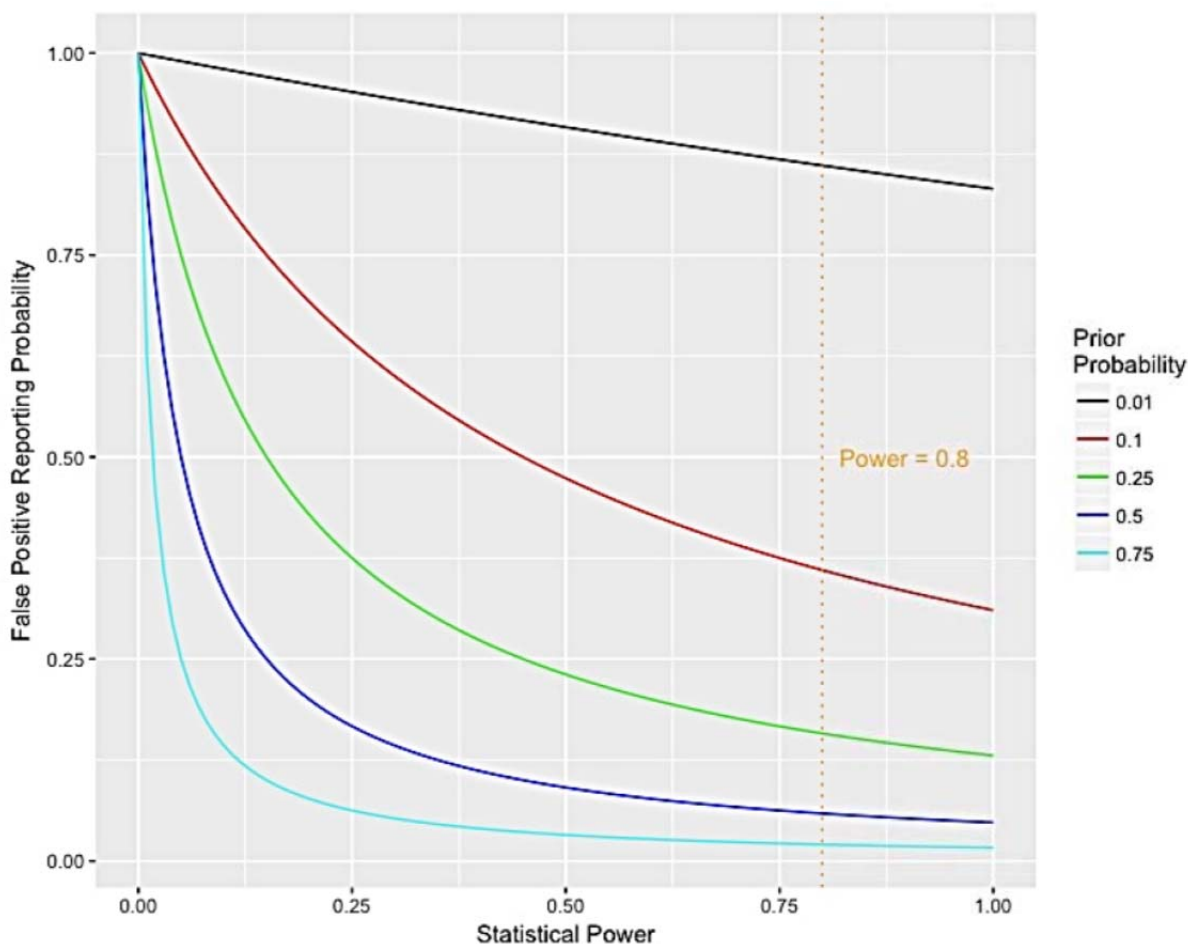
349
350 Table 2. False positive report probability (the probability of a statistically significant result being a false
351 positive) as a function of prior probability and statistical power. Note that for unlikely hypotheses, larger
352 portions of statistically significant findings will be false positives. This table assumes a significance
353 threshold of $p < 0.05$.

354

	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

355
356
357 Figure 1. The relationship between prior probability, statistical power, and the false positive report
358 probability. The false positive report probability is the probability of a statistically significant result being

359 a false positive. Note that for unlikely hypotheses, large portions of statistically significant findings will
360 be false positives even with high power. This figure is based on a significance threshold of $p < 0.05$.
361
362



363
364
365
366
367 Many researchers in biology are unaware that the strength of evidence presented by a p-value depends
368 on the prior probability of the outcome. When testing moderately unlikely hypotheses (those with a
369 10% chance of being true) in a test with high statistical power, more than 1/3 of statistically 'significant'
370 effects below the $p < 0.05$ threshold will be false positives (Table 2, Fig. 1) (Forstmeier et al., 2016).
371 Thus, if robust pre-existing information makes a result unlikely, that result should be held to a higher
372 standard of evidence than would be appropriate for a hypothesis that has already been empirically
373 supported and thus has a higher prior probability. For instance, if using a null hypothesis test, a better
374 significance threshold for results without prior support may be $p < 0.005$ (Benjamin et al., in press).
375 Determining prior probability is an imperfect undertaking, and even experts can be deceived, for
376 instance by bias in the existing literature (Palmer, 2000). However, considering this issue is important for
377 thoroughly evaluating the evidence presented in a manuscript.

378
379 **Conclusion**
380

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

392
393

394 **Literature Cited**

395

- 396 Arriaga AF, Bader AM, Wong JM, Lipsitz SR, Berry WR, Ziewacz JE, Hepner DL, Boorman DJ, Pozner CN,
397 Smink DS, Gawande AA, 2013. Simulation-based trial of surgical-crisis checklists. *New England*
398 *Journal of Medicine* 368:246-253. doi: 10.1056/NEJMsa1204720.
- 399 Baier A, Baker L, 2013. *A Guide to Peer Review in Ecology and Evolution*. British Ecological Society.
- 400 Barto EK, Rillig MC, 2012. Dissemination biases in ecology: effect sizes matter more than quality. *Oikos*
401 121:228-235. doi: 10.1111/j.1600-0706.2011.19401.x.
- 402 Benjamin DJ, Berger JO, Johannesson M, Nosek BA, Wagenmakers E-J, Berk R, Bollen KA, Brembs B,
403 Brown L, Camerer C, Cesarini D, Chambers CD, Clyde M, Cook TD, De Boeck P, Dienes Z, Dreber
404 A, Easwaran K, Efferson C, Fehr E, F. F, Field AP, Forster M, George EI, Gonzalez R, Goodman S,
405 Green E, Green DP, Greenwald A, Hadfield JD, Hedges LV, Held L, Ho T-H, Hoijtink H, Jones JH,
406 Hruschka DJ, Imai K, Imbens G, Ioannidis JPA, Jeon M, Kirchler M, Laibson D, List J, Little R, Lupia
407 A, Machery E, Maxwell SE, McCarthy M, Moore D, Morgan SL, Munafó M, Nakagawa S, Nyhan B,
408 Parker TH, Pericchi L, Perugini M, Rouder J, Rousseau J, Savalei V, Schönbrodt FD, Sellke T,
409 Sinclair B, Tingley D, Van Zandt T, Vazire S, Watts DJ, Winship C, Wolpert RL, Xie Y, Young C,
410 Zinman J, Johnson VE, in press. Redefine statistical significance *Nature Human Behaviour*.
- 411 Cobo E, Cortés J, Ribera JM, Cardellach F, Selva-O'Callaghan A, Kostov B, García L, Cirugeda L, Altman
412 DG, González JA, Sánchez JA, Miras F, Urrutia A, Fonollosa V, Rey-Joly C, Vilardell M, 2011. Effect
413 of using reporting guidelines during peer review on quality of final manuscripts submitted to a
414 biomedical journal: masked randomised trial. *BMJ* 343. doi: 10.1136/bmj.d6783.
- 415 Cobo E, Selva-O'Callaghan A, Ribera J-M, Cardellach F, Dominguez R, Vilardell M, 2007. Statistical
416 reviewers improve reporting in biomedical articles: a randomized trial. *PLOS ONE* 2:e332. doi:
417 10.1371/journal.pone.0000332.
- 418 Duffy JE, Godwin CM, Cardinale BJ, 2017. Biodiversity effects in the wild are common and as strong as
419 key drivers of productivity. *Nature* 549:261-264. doi: 10.1038/nature23886.
- 420 Ferreira V, Castagnyrol B, Koricheva J, Gulis V, Chauvet E, Graça MAS, 2015. A meta-analysis of the
421 effects of nutrient enrichment on litter decomposition in streams. *Biological Reviews* 90:669-
422 688. doi: 10.1111/brv.12125.
- 423 Fidler F, Burgman MA, Cumming G, Buttrose R, Thomason N, 2006. Impact of criticism of null-hypothesis
424 significance testing on statistical reporting practices in conservation biology. *Conservation*
425 *Biology* 20:1539-1544. doi: 10.1111/j.1523-1739.2006.00525.x.
- 426 Fidler F, Chee YE, Wintle BC, Burgman MA, McCarthy MA, Gordon A, 2017. Metaresearch for evaluating
427 reproducibility in ecology and evolution. *BioScience* 67:282-289. doi: 10.1093/biosci/biw159.

428 Fischhoff B, 1975. Hindsight not equal to foresight – effect of outcome knowledge on judgment under
429 uncertainty. *Journal of Experimental Psychology–Human Perception and Performance* 1:288–
430 299.

431 Forstmeier W, Schielzeth H, 2011. Cryptic multiple hypotheses testing in linear models: overestimated
432 effect sizes and the winner's curse. *Behavioral Ecology and Sociobiology* 65:47-55. doi:
433 10.1007/s00265-010-1038-5.

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

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

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

438 Gerstner K, Moreno-Mateos D, Gurevitch J, Beckmann M, Kambach S, Jones HP, Seppelt R, 2017. Will
439 your paper be used in a meta-analysis? Make the reach of your research broader and longer
440 lasting. *Methods in Ecology and Evolution* 8:777-784. doi: 10.1111/2041-210X.12758.

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

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

444 Jefferson T, Wager E, Davidoff F, 2002. Measuring the quality of editorial peer review. *JAMA* 287:2786-
445 2790. doi: 10.1001/jama.287.21.2786.

446 Kardish MR, Mueller UG, Amador-Vargas S, Dietrich EI, Ma R, Barrett B, Fang C-C, 2015. Blind trust in
447 unblinded observation in ecology, evolution and behavior. *Frontiers in Ecology and Evolution*
448 3:51. doi: 10.3389/fevo.2015.00051.

449 Kozlov MV, Zverev V, Zvereva EL, 2014. Confirmation bias leads to overestimation of losses of woody
450 plant foliage to insect herbivores in tropical regions. *PeerJ* 2:e709. doi: 10.7717/peerj.709.

451 Lee CJ, Sugimoto CR, Zhang G, Cronin B, 2013. Bias in peer review. *Advances in Information Science*
452 64:2-17. doi: 10.1002/asi.22784.

453 Lemoine NP, Hoffman A, Felton AJ, Baur L, Chaves F, Gray J, Yu Q, Smith MD, 2016. Underappreciated
454 problems of low replication in ecological field studies. *Ecology* 97:2554-2561. doi:
455 10.1002/ecy.1506.

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

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

460 Nakagawa S, Cuthill IC, 2007. Effect size, confidence interval and statistical significance: a practical guide
461 for biologists. *Biological Reviews* 82:591-605. doi: 10.1111/j.1469-185X.2007.00027.x.

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

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

466 Parker TH, 2013. What do we really know about the signalling role of plumage colour in blue tits? A case
467 study of impediments to progress in evolutionary biology. *Biological Reviews* 88:511-536.

468 Parker TH, Forstmeier W, Koricheva J, Fidler F, Hadfield JD, Chee YE, Kelly CD, Gurevitch J, Nakagawa S,
469 2016. Transparency in ecology and evolution: real problems, real solutions. *Trends in Ecology &*
470 *Evolution* 31:711-719. doi: 10.1016/j.tree.2016.07.002.

471 Simmons JP, Nelson LD, Simonsohn U, 2011. False positive psychology: undisclosed flexibility in data
472 collection and analysis allows presenting anything as significant. *Psychological Science* 22:1359-
473 1366.

474 Simmons JP, Nelson LD, Simonsohn U, 2012. A 21 word solution. *Dialogue: Newsletter of the SPSP* 26:4-
475 7.

476 TTEE_Working_Group, 2016. Tools for Transparency in Ecology and Evolution (TTEE). Open Science
477 Framework. doi: 10.17605/OSF.IO/G65CB.
478 van Wilgenburg E, Elgar MA, 2013. Confirmation bias in studies of nestmate recognition: a cautionary
479 note for research into the behaviour of animals. PLoS ONE 8:e53548. doi:
480 10.1371/journal.pone.0053548.
481 Zhang Y, Chen HYH, Reich PB, 2012. Forest productivity increases with evenness, species richness and
482 trait variation: a global meta-analysis. Journal of Ecology 100:742-749. doi: 10.1111/j.1365-
483 2745.2011.01944.x.

484

485

486