

1 Empowering peer reviewers with a checklist to improve transparency

2

3 Timothy H. Parker^{1,2}

4 Simon C. Griffith²

5 Judith L. 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 W. 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 Department of Ecology and Evolutionary Biology, University of Connecticut, 75 N. Eagleville Rd U-
34 3043, Storrs, CT 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 Abstract

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

49 Two important tasks facing peer reviewers are assessing the soundness of study design and evaluating
50 the reporting of methods and results. Study soundness and reporting both bear directly on the
51 reliability of the inferences that can be drawn from the papers that are ultimately published¹. Other
52 reviewing tasks include considering the placement of the study in a broader context, the writing, and the
53 importance of the research, but these vary by journal and the expertise of the reviewer, and are often
54 more subjective. We therefore focus on only the first two reviewing tasks. Our goal here is to explain
55 particular components of this assessment process that we believe are too frequently ignored by peer
56 reviewers, ultimately to the detriment of the scientific literature. We combine these components in a
57 checklist that reviewers can use to improve transparency and reduce bias, and thus improve the
58 reliability of scientific inferences.

59
60 We present this checklist as a series of ten questions (summarized in Table 1), each accompanied by
61 suggestions for how the reviewer should proceed depending on the answer to that question. The
62 checklist is not meant to be comprehensive. A longer checklist to help reviewers in ecology and
63 evolutionary biology promote transparency was created as part of TTEE (Tools for Transparency in
64 Ecology and Evolution; <https://osf.io/y8aqx/>) in an effort to help journals in ecology and evolutionary
65 biology adopt TOP (Transparency and Openness Promotion) guidelines². TTEE checklists, for both
66 reviewers and authors, were designed to cover a broad swath of transparency issues. In contrast, the
67 short checklist we present in this paper focusses on the subset of practices that we think are critically in
68 need of improvement, and on which we think a concise checklist can achieve greatest impact. Our
69 checklist provides reviewers with an efficient tool for promoting transparency in empirical research
70 papers. Until now, no such tool has existed.

71 72 **Why a checklist?**

73
74 The use of checklists is well established among skilled practitioners working in complex systems.
75 Checklists make flying complicated aircraft safer, they free architects to devote their mental energy to
76 creativity, and they help surgeons focus on applying their skill without forgetting vital tasks^{3,4}. Good
77 checklists do not replace complex thought; they facilitate it. Of course, effective peer review requires
78 expertise and critical thinking skills that no practical checklist can provide. However, this does not mean
79 that checklists cannot be used to improve peer review, even dramatically, by calling attention to
80 essential elements that are often overlooked.

81
82 Checklists can be of use to peer reviewers in two primary ways related to creating a more transparent
83 and less biased literature: to help reviewers check (1) mundane but important details, and (2) both their
84 own and the authors' potential biases. With regard to the first point, incomplete reporting of
85 information hinders interpretation of studies and effective synthesis, and thus scientific progress^{1,5}. We
86 know from surveys of subsets of the ecology literature that approximately half of published papers omit
87 important information such as sample size or variability associated with estimates⁶⁻⁸. Nearly all papers
88 omitting this information were peer reviewed, suggesting that reviewers either overlooked these
89 details, or felt that it was someone else's job to monitor them. Whether we notice omissions as
90 reviewers depends on scrutiny that may vary unconsciously with factors such as whether we agree with
91 the study's conclusions, our perception of the expertise of the authors, or whether we have used similar
92 research designs. Regardless, the frequency of these omissions in the literature is evidence of a
93 systematic problem, but one that could be resolved with the help of an appropriate checklist.

94
95 With regard to the second point, we need to explicitly address potential bias from authors and
96 reviewers because all people, scientists included, are subject to biases that influence the information we

97 notice and how we interpret that information^{9,10}. Such biases have been shown to have major impacts
98 on the content of scientific papers¹¹⁻¹³, and so we expect that biases also influence the opinions we
99 form when reviewing scientific papers. In fact, evidence suggests that peer review often suffers from a
100 multitude of complex, systematic biases¹⁴.

101
102 We hope that reviewers will find the questions in this checklist useful for most reviews. To facilitate the
103 checklist's use, we provide some suggestions for reviewer responses to individual checklist questions,
104 although we cannot provide a set of all possible answers to each checklist question. Occasionally
105 reviewers will be uncertain about answers to one or more of these questions. Sometimes this
106 uncertainty can be resolved by asking for additional information from the authors, and sometimes the
107 reviewer should simply notify the editor so that she or he can seek additional reviewer expertise if
108 needed. Of course some questions may not apply to some papers; it will be up to the reviewer to
109 determine the relevance of each question. Determining its relevance may be aided by the explanation
110 and justification that we provide following each particular checklist item.

111 112 **The checklist**

113 114 **Questions to promote transparent reporting of methods and results**

115
116 1. Were all sample sizes fully reported, including exact values for all subsets of data (e.g., each
117 treatment group), and for all statistical analyses?

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

119
120 Knowledge of sample size is essential for understanding the power of analyses (see below) and the
121 reliability of estimates, and thus for interpreting results. It is also essential for later meta-analytic
122 synthesis⁵. Yet, researchers fail to report sample sizes with troubling frequency^{7,8}. Reporting a
123 range (e.g. "9-12 replicates per treatment") is inadequate.

124
125 2. Are the methods for carrying out the study and analysing the results reported in sufficient detail to
126 allow another researcher to gather the same data and run the identical analyses? When not in the paper
127 itself, methodological details should be included in a supplement, or in many cases, archived in a
128 publicly accessible and curated repository.

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

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

132
133 By keeping replicability in mind while reading the methods, the reviewer can determine if methods
134 have been reported in sufficient detail. Necessary details vary among studies with different
135 methods. For instance, in the case of Bayesian analyses, authors should explicitly define their priors
136 and report how their posterior distributions were derived, if applicable including Markov chain
137 Monte Carlo specifications, and method of convergence (mixing) assessment. Archiving of details
138 such as analysis code is essential if others are to understand how results were derived¹⁵ (see also
139 [http://www.britishecologicalsociety.org/wp-content/uploads/2017/12/guide-to-reproducible-
140 code.pdf](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
141 information should be stored in curated archives. Temporary and uncurated repositories, including
142 personal websites and the version-control site GitHub, are not viable for long-term storage. There
143 will occasionally be valid justifications for not reporting certain information (e.g., population

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

146
147 3. Are statistical results reported completely (considered in two parts below)?
148

149 3a. Are statistical results for each test reported in sufficient detail? What qualifies as ‘sufficient detail’
150 will differ among analyses. For most analyses, however, this will include (but not be limited to) basic
151 parameter estimates of central tendency (e.g., means) or other basic estimates (e.g., regression or
152 correlation coefficients) and variation (e.g., standard deviation) or associated estimates of uncertainty
153 (e.g., confidence/credible intervals). For null hypothesis tests, reporting P-values and test statistics by
154 themselves is almost always insufficient.

155 →If ‘no’, request that authors provide this information.

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

159
160 3b. Are results from all variables and from all models reported? Complete reporting should include
161 results related to all variables examined in preliminary models and all results from exploratory analyses.
162 It will sometimes be appropriate to include these as supplementary materials. For analysis types that
163 generate vast sets of results, it may be appropriate to place results in data archives.

164 →If ‘no’, request that authors provide this information.

165 →If you are uncertain, ask the authors to declare in the paper that all exploratory analyses are reported
166 in full. We recommend using the ‘Standard Reviewer Statement for Disclosure of Sample, Conditions,
167 Measures, and Exclusions’: "I request that the authors add a statement to the paper confirming
168 whether, for all experiments, they have reported all measures, conditions, data exclusions, and how
169 they determined their sample sizes. The authors should, of course, add any additional text to ensure the
170 statement is accurate. This is the standard reviewer disclosure request endorsed by the Center for Open
171 Science [see <http://osf.io/hadz3>]".

172
173 Insufficient reporting of results is one of the largest obstacles to an unbiased understanding of
174 empirical progress^{1,16}. Sometimes authors state that an analysis was conducted, but fail to provide
175 all the relevant statistical outcomes such as slope estimates or estimates of variability^{6-8,17}. At other
176 times, authors conduct multiple analyses but do not explicitly acknowledge that they are reporting
177 results from only a subset. Both practices may sometimes result from a direct request by the journal
178 to shorten the text because of space limits or a desire for a concise story. Regardless, they weaken
179 our ability to draw unbiased conclusions from the published literature. The failure to provide all
180 relevant details from a reported analysis is often easily recognized by reviewers. In contrast,
181 analyses that have been conducted but are completely unreported are more difficult, and
182 sometimes even impossible, to recognize. However, there can be signs of unreported analyses: for
183 instance different variables may be included in different models without obvious a priori
184 justification, a subset of potential interactions may be provided without clear justification for the
185 choice, or the authors may have failed to examine obvious predictions that are testable with
186 available data. Each of these signs were found in a sample of literature in behavioural ecology,
187 providing circumstantial evidence of unreported analyses¹⁷. Reviewers can prompt authors to
188 include missing information in supplementary materials or in searchable, curated data archives.
189 Asking authors to state whether all results from all analyses have been reported should lead authors
190 to be more transparent about their exploratory work¹⁸. If necessary, authors should be directed to
191 consult published recommendations regarding thorough reporting of results (and methodological

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

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

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

204
205 It is now well demonstrated that researchers’ observations are often influenced by what they expect
206 to see ^{12,13}. For instance, when researchers were unaware of the colony of origin of the ants they
207 were observing, they were > 3 times more likely to report aggression between colony mates than
208 were researchers who knew the ants’ colony of origin ¹². Keeping observers unaware of treatment
209 categories or expected outcomes is not always possible or reasonable, but researchers should at
210 least discuss the possibility of unconscious bias ¹⁹.

211
212 5. Did the authors explain how sample size was decided (e.g., based on a priori power analysis or
213 logistical constraints), or when an experiment with pre-set sample sizes was terminated? If sample size
214 or the end of the experiment was not decided prior to the initiation of the study, was there a decision
215 rule for when to cease data collection?
216 → If not reported, request that authors provide this information.
217 → If the stopping rule included iterative statistical tests or examination of patterns as data accumulated,
218 request that authors acknowledge the bias resulting from this process.

219
220 Cessation of data collection should never be made in response to reaching some threshold of
221 statistical significance or effect. Such a practice leads to strong bias in favour of effects inflated by
222 sampling error ^{20,21}. An explanation for the choice of stopping point should be provided (e.g., “we
223 planned to harvest samples at the end of the second growing season”).

224
225 6. Did the authors develop their analysis plan, including choices of variables, without looking at the data,
226 for instance prior to gathering data or with a dummy data set? This is most easily determined by the
227 existence of a pre-registered analysis plan. In the absence of pre-registration, a statement from the
228 authors about the development of their analysis plan is still important.
229 → If no, request that authors acknowledge the exploratory nature of their analyses and declare that
230 they are reporting the complete set of results from all exploratory analyses.
231 → If authors deviated from their analysis plan, request an explanation of why and how they deviated
232 from the plan.

233
234 Choosing the analyses to present based on the strength of the effects derived from those analyses
235 or models biases the distribution of presented results and can lead to presentation of entirely
236 spurious relationships ^{20,22}. An ideal solution is to develop an analysis plan before examining the data
237 and file it in a pre-registration archive such as offered by the Open Science Framework
238 (<https://cos.io/prereg/>). One plausible alternative is an unusually detailed and publicly available
239 grant proposal. Either way the pre-registration or proposal should be cited in the manuscript.

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

244
245 7. How suitable do you find the research methods without considering the outcome? Evaluate the
246 design and methods regardless of whether or not there was a finding of “statistical significance”, or
247 whether or not the results conform to a predicted pattern.

248 → If the methods appear to have been flawed, call attention to the problems and, if possible,
249 recommend a better design. Deciding whether the problems with the methods are sufficient to justify a
250 recommendation of rejection will require your expert judgement.

251 → If uncertain about the suitability of some aspect of the methods, state your uncertainty to the editor
252 so that she or he can seek appropriate methodological expertise as needed.

253
254 One driver of bias in the published literature is that we often evaluate the suitability of a study’s
255 methods based on the direction and strength of results²³. This is especially true in cases of smaller
256 samples or weaker study designs. In such cases, studies producing statistically significant or strong
257 effects are sometimes incorrectly viewed as more plausible than those reporting weak or
258 statistically non-significant results. There is a tendency among people we have talked with to
259 assume that if a study found statistically significant results, sample sizes were sufficient or
260 methodological weakness was not much of a problem. However, since ‘strong’ or ‘significant’ effects
261 can often arise by chance²⁴ or be selected for reporting from among other unreported results^{20,22},
262 such results cannot be taken as proof that a study’s methodological limitations were not a problem.
263 Instead, the quality of the methods must be judged independent of the results. (Of course some
264 studies include tests designed to assess a method’s effectiveness rather than to assess the biological
265 effect of primary interest, and those tests should be used to determine the quality of methods).
266 Doubts about the reliability of the methods should be given equal strength regardless of the primary
267 outcome.

268
269 8. Are the sample sizes large enough to justify the authors’ conclusions? If presenting significance tests,
270 how much power would this study have to detect statistically significant weak, moderate, and strong
271 effects? (See Table 2 for examples of how sample size and effect size combine to determine power in
272 two types of simple analyses.) Expectation of effect size can best be derived from average effect sizes
273 presented in meta-analyses of similar topics. The effect size reported in the manuscript under review
274 can be a poor estimate of the underlying effect size, especially if the sample size is small thus elevating
275 sampling uncertainty. Statistical significance is a poor indicator of the reliability of an estimate across a
276 wide range of sample sizes and common effect sizes (Table 2 provides insight into statistical power).

277 → If sample sizes are small in a system where effects are expected to be weak to moderate, request that
278 authors avoid inferences based on threshold p-values, acknowledge uncertainty in effect size estimates,
279 and acknowledge the need for further study.

280 → Do not use sample size as a criterion for recommending publication unless you do so regardless of
281 study outcome (i.e., regardless of reported effect size and regardless of the outcomes of tests for
282 significance).

283 → Do not use the failure to surpass a significance threshold as a reason to recommend rejection.

284
285 Presumably, nearly all ecologists and evolutionary biologists understand that there are problems
286 with low power. However, it is clear that most of us would benefit from a reminder that type II error
287 (false negatives) is only one of these problems. Because effect sizes are more variable with small

288 samples, inflated effect sizes are more likely, and thus large effects derived from small samples can
289 be unreliable²⁵⁻²⁷. In fact, with low power caused by some combination of a small sample and
290 relatively weak biological effect, studies are likely to reach statistical significance only if sampling
291 error drives the observed effect size much higher than the true effect^{25,27}. Unfortunately the weak
292 to moderate-strength biological effects that contribute to low power are common in ecology and
293 evolutionary biology, at least in some sub-disciplines^{27,28}. However, biological effects can be larger
294 in some types of studies and in some systems^{27,29}, and so what qualifies as a small sample in one
295 study may be sufficiently large in another. Thus, evaluating sample size and power will benefit from
296 knowledge of the typical effect sizes for the type of study in question, and this can most reliably be
297 learned by consulting meta-analyses. If the study under review appears to have low power, we
298 should not consider meeting a threshold p-value to be a reliable index of the validity of a pattern or
299 a given effect size. In general, the reviewer should treat conclusions derived from low-powered
300 studies as tentative, whether or not some significance threshold was met. However, studies with
301 low power may often be worthy of publication, as some studies face major logistical obstacles
302 regarding sample size, and it is only through publication and subsequent meta-analysis of a series of
303 studies with small samples that we build a robust understanding of the true effect size²⁷.

304
305 9. What does the size of the estimated effect (e.g., slope, correlation coefficient, difference in means)
306 suggest about its biological or practical importance, and what does uncertainty around that effect
307 estimate suggest about the estimate's precision? Depending on the biological question, weak effects
308 may be either biologically important or of limited interest; authors should justify their interpretation
309 accordingly. Uncertainty around effects can be estimated with standard error (SE), 95% confidence
310 intervals (approximately 2 x SE), or with other statistics. As sample size increases (see checklist question
311 8 above) and variance decreases, SE decreases and we gain confidence in the precision of the effect
312 estimate.

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

315
316 Evaluating results based on the size of the effect estimate and the associated uncertainty rather
317 than based on a p-value provides more direct insight into the biological phenomenon of interest³⁰.
318 Too often, interpretation of results focusses on statistical significance rather than on biological
319 significance, and thus we can be led astray regarding our understanding of their relevance.

320
321 10. How unexpected would you judge these results to be in light of prior empirically derived
322 understanding? Effects that are more surprising in light of robust prior information are those that had a
323 lower prior probability of being correct. When testing unlikely hypotheses, the chance that a statistically
324 significant result is a false positive rises dramatically (Table 3, Fig. 1). $P < 0.05$ is a poor threshold for
325 evaluating the significance of an unexpected discovery and should be presented as no more than
326 suggestive evidence for such discoveries.

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

330
331 Findings should be interpreted in light of previously published information, and the more robust the
332 body of pre-existing information, the more caution authors should exercise when interpreting the
333 implications of their contradictory results. To quote Carl Sagan, "Extraordinary claims require
334 extraordinary evidence". For instance, many researchers in biology are unaware that the strength of
335 evidence presented by a p-value depends on the prior probability of the outcome. When testing

336 moderately unlikely hypotheses (those with a 10% chance of being true) in a test with high statistical
337 power, more than 1/3 of statistically 'significant' effects below the $p < 0.05$ threshold will be false
338 positives (Table 3, Fig. 1)²¹. Thus, if robust pre-existing information makes a result unlikely, that
339 result should be held to a higher standard of evidence than would be appropriate for a hypothesis
340 that has already been empirically supported and thus has a higher prior probability³¹. For instance,
341 a finding that parental diet influenced offspring phenotype is consistent with previously published
342 findings and theory, but a finding that parental diet influenced grand-offspring phenotype more
343 strongly than it influenced offspring phenotype would be extraordinary. Extraordinary results may
344 be correct, but relative to results with a high prior probability, the extraordinary results are more
345 likely to be false positives. We are not suggesting that reviewers estimate prior probabilities.
346 However, a qualitative consideration of this issue is important for thoroughly evaluating the link
347 between evidence and inference presented in a manuscript.

348 **Conclusions**

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

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

369 Our checklist questions are practical tools. We hope that these questions, along with peer review
370 checklists that address a broader set of topics (e.g., TTEE; <https://osf.io/y8aqx/>), will improve
371 transparency in the published literature and thus reduce bias therein. More transparency and less bias
372 should mean more reliable inferences in published papers and later in the meta-analyses based on those
373 published papers¹. As we improve peer review, we improve the quality of science.

374 **References**

- 375
376
377
378
379
380
381
382 1 Parker, T. H. *et al.* Transparency in ecology and evolution: real problems, real solutions. *Trends*
383 *in Ecology & Evolution* **31**, 711-719, doi:10.1016/j.tree.2016.07.002 (2016).

384 2 TTEE_Working_Group. *Tools for Transparency in Ecology and Evolution (TTEE)*,
385 <<https://osf.io/g65cb/>> (2016).

386 3 Arriaga, A. F. *et al.* Simulation-based trial of surgical-crisis checklists. *New England Journal of*
387 *Medicine* **368**, 246-253, doi:10.1056/NEJMsa1204720 (2013).

388 4 Gawande, A. A. *The Checklist Manifesto: How to Get Things Right*. (Metropolitan Books, New
389 York, 2009).

390 5 Gerstner, K. *et al.* Will your paper be used in a meta-analysis? Make the reach of your research
391 broader and longer lasting. *Methods in Ecology and Evolution* **8**, 777-784, doi:10.1111/2041-
392 210X.12758 (2017).

393 6 Ferreira, V. *et al.* A meta-analysis of the effects of nutrient enrichment on litter decomposition
394 in streams. *Biological Reviews* **90**, 669-688, doi:10.1111/brv.12125 (2015).

395 7 Fidler, F., Burgman, M. A., Cumming, G., Buttrose, R. & Thomason, N. Impact of criticism of null-
396 hypothesis significance testing on statistical reporting practices in conservation biology.
397 *Conservation Biology* **20**, 1539-1544, doi:10.1111/j.1523-1739.2006.00525.x (2006).

398 8 Zhang, Y., Chen, H. Y. H. & Reich, P. B. Forest productivity increases with evenness, species
399 richness and trait variation: a global meta-analysis. *Journal of Ecology* **100**, 742-749,
400 doi:10.1111/j.1365-2745.2011.01944.x (2012).

401 9 Nickerson, R. S. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General*
402 *Psychology* **2**, 175-220, doi:10.1037/1089-2680.2.2.175 (1998).

403 10 Fischhoff, B. Hindsight not equal to foresight – effect of outcome knowledge on judgment under
404 uncertainty. *Journal of Experimental Psychology–Human Perception and Performance* **1**, 288–
405 299 (1975).

406 11 Kozlov, M. V., Zverev, V. & Zvereva, E. L. Confirmation bias leads to overestimation of losses of
407 woody plant foliage to insect herbivores in tropical regions. *PeerJ* **2**, e709, doi:10.7717/peerj.709
408 (2014).

409 12 van Wilgenburg, E. & Elgar, M. A. Confirmation bias in studies of nestmate recognition: a
410 cautionary note for research into the behaviour of animals. *PLoS ONE* **8**, e53548,
411 doi:10.1371/journal.pone.0053548 (2013).

412 13 Holman, L., Head, M. L., Lanfear, R. & Jennions, M. D. Evidence of experimental bias in the life
413 sciences: why we need blind data recording. *PLoS Biol* **13**, e1002190,
414 doi:10.1371/journal.pbio.1002190 (2015).

415 14 Lee, C. J., Sugimoto, C. R., Zhang, G. & Cronin, B. Bias in peer review. *Advances in Information*
416 *Science* **64**, 2-17, doi:10.1002/asi.22784 (2013).

417 15 Mislán, K. A. S., Heer, J. M. & White, E. P. Elevating the status of code in ecology. *Trends in*
418 *Ecology & Evolution* **31**, 4-7, doi:10.1016/j.tree.2015.11.006 (2016).

419 16 Fidler, F. *et al.* Metaresearch for evaluating reproducibility in ecology and evolution. *BioScience*
420 **67**, 282-289, doi:10.1093/biosci/biw159 (2017).

421 17 Parker, T. H. What do we really know about the signalling role of plumage colour in blue tits? A
422 case study of impediments to progress in evolutionary biology. *Biological Reviews* **88**, 511-536
423 (2013).

424 18 Simmons, J. P., Nelson, L. D. & Simonsohn, U. A 21 word solution. *Dialogue: Newsletter of the*
425 *SPSP* **26**, 4-7 (2012).

426 19 Kardish, M. R. *et al.* Blind trust in unblinded observation in ecology, evolution and behavior.
427 *Frontiers in Ecology and Evolution* **3**, 51, doi:10.3389/fevo.2015.00051 (2015).

428 20 Simmons, J. P., Nelson, L. D. & Simonsohn, U. False positive psychology: undisclosed flexibility in
429 data collection and analysis allows presenting anything as significant. *Psychological Science* **22**,
430 1359-1366 (2011).

431 21 Forstmeier, W., Wagenmakers, E.-J. & Parker, T. H. Detecting and avoiding likely false-positive
432 findings – a practical guide. *Biological Reviews* **92**, 1941-1968, doi:10.1111/brv.12315 (2017).
433 22 Forstmeier, W. & Schielzeth, H. Cryptic multiple hypotheses testing in linear models:
434 overestimated effect sizes and the winner's curse. *Behavioral Ecology and Sociobiology* **65**, 47-
435 55, doi:10.1007/s00265-010-1038-5 (2011).
436 23 Palmer, A. R. Quasireplication and the contract of error: lessons from sex ratios, heritabilities
437 and fluctuating asymmetry. *Annual Review of Ecology and Systematics* **31**, 441-480 (2000).
438 24 Halsey, L. G., Curran-Everett, D., Vowler, S. L. & Drummond, G. B. The fickle *P* value generates
439 irreproducible results. *Nature Methods* **12**, 179, doi:10.1038/nmeth.3288 (2015).
440 25 Gelman, A. & Weakliem, D. Of beauty, sex, and power. *American Scientist* **97**, 310-316 (2009).
441 26 Barto, E. K. & Rillig, M. C. Dissemination biases in ecology: effect sizes matter more than quality.
442 *Oikos* **121**, 228-235, doi:10.1111/j.1600-0706.2011.19401.x (2012).
443 27 Lemoine, N. P. *et al.* Underappreciated problems of low replication in ecological field studies.
444 *Ecology* **97**, 2554-2561, doi:10.1002/ecy.1506 (2016).
445 28 Møller, A. P. & Jennions, M. D. How much variance can be explained by ecologists and
446 evolutionary biologists? *Oecologia* **132**, 492-500 (2002).
447 29 Duffy, J. E., Godwin, C. M. & Cardinale, B. J. Biodiversity effects in the wild are common and as
448 strong as key drivers of productivity. *Nature* **549**, 261-264, doi:10.1038/nature23886 (2017).
449 30 Nakagawa, S. & Cuthill, I. C. Effect size, confidence interval and statistical significance: a practical
450 guide for biologists. *Biological Reviews* **82**, 591-605, doi:10.1111/j.1469-185X.2007.00027.x
451 (2007).
452 31 Benjamin, D. J. *et al.* Redefine statistical significance *Nature Human Behaviour* **2**, 6-10,
453 doi:10.1038/s41562-017-0189-z (2018).

454

455 Highlighted references:

456

457 Barto, E. K. & Rillig, M. C. Dissemination biases in ecology: effect sizes matter more than quality. *Oikos*
458 121, 228-235, doi:10.1111/j.1600-0706.2011.19401.x (2012).

459

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

462

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

465

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

468

469 Lemoine, N. P. et al. Underappreciated problems of low replication in ecological field studies. *Ecology*
470 97, 2554-2561, doi:10.1002/ecy.1506 (2016).

471

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

474

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

480 Table 2. Power to detect a true biological effect as statistically significant ($p < 0.05$) as a function of
 481 sample size and actual effect size for two types of simple analysis. High power is typically considered 0.8,
 482 or an 80% chance of detecting an effect (designated with * below) if the effect exists. Note that
 483 obtaining high power to detect small to medium effects (those most common in ecology and evolution)
 484 requires sample sizes much larger than are typical.
 485

		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*

486
 487
 488

489 Table 3. False positive report probability (the probability that a statistically significant result is a false
 490 positive – in other words, the probability that, in the case of a statistically significant rejection of the
 491 null, the null hypothesis is actually true) as a function of prior probability and statistical power. Note
 492 that for unlikely hypotheses, larger portions of statistically significant findings will be false positives. This
 493 table assumes a significance threshold of $p < 0.05$.
 494

prior	power			
	0.1	0.2	0.5	0.8
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

495
 496
 497

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

504
505
506

507 **Acknowledgements**

508

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

511

512

513 **Contributions**

514

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

517

518

519 **Competing interests**

520

521 The authors declare no competing financial interests.

522

523

524 **Corresponding author**

525

526 Correspondence to Timothy Parker.

527

