



THE UNIVERSITY *of* EDINBURGH

Edinburgh Research Explorer

Transparency in ecology and evolution: real problems, real solutions

Citation for published version:

Hadfield, J, Parker, TH, Forstmeier, W, Koricheva, J, Fidler, F, En Chee, Y, Kelly, C, Gurevitch, J & Nakagawa, S 2016, 'Transparency in ecology and evolution: real problems, real solutions', *Trends in Ecology & Evolution*, vol. 31, no. 9, pp. 711-719. <https://doi.org/10.1016/j.tree.2016.07.002>

Digital Object Identifier (DOI):

[10.1016/j.tree.2016.07.002](https://doi.org/10.1016/j.tree.2016.07.002)

Link:

[Link to publication record in Edinburgh Research Explorer](#)

Document Version:

Peer reviewed version

Published In:

Trends in Ecology & Evolution

General rights

Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy

The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.



1 Transparency in ecology and evolution: real problems, real solutions

2

3 Timothy H. Parker, Whitman College, Walla Walla, USA

4 Wolfgang Forstmeier, Max Planck Institute for Ornithology, Seewiesen, Germany

5 Julia Koricheva, Royal Holloway University of London, Egham, UK

6 Fiona Fidler, University of Melbourne, Melbourne, Australia

7 Jarrod Hadfield, University of Edinburgh, Edinburgh, UK

8 Yung En Chee, University of Melbourne, Melbourne, Australia

9 Clint Kelly, University of Quebec, Montreal, Canada

10 Jessica Gurevitch, Stony Brook University, New York, USA

11 Shinichi Nakagawa, University of New South Wales, Sydney, Australia

12

13 Keywords:

14 confirmation bias; inflated effect size; *p*-hacking; pre-registration; replication; selective reporting

15

16 Highlights (separate from the manuscript):

17

18 1. Evidence suggests that insufficient transparency is a common problem across much of ecology and
19 evolution. Results and methods are often reported with insufficient detail thereby hampering
20 interpretation and meta-analysis, and many results go entirely unreported. Further, these unreported
21 results are often a biased subset. Thus the conclusions we can draw from the published literature, both
22 from individual papers and from aggregated results, are themselves often biased.

23

24 2. Journals and other institutions, such as funding agencies, influence the decisions researchers make
25 about disseminating their results. Various existing policies of these institutions promote or facilitate
26 practices that are not transparent. However, there is a movement across empirical disciplines, and now
27 within ecology and evolution, to shape editorial policies to better promote transparency. This can be
28 done by requiring or encouraging more disclosure, as with the now-familiar data archiving, or by
29 developing an incentive structure promoting disclosure, such as pre-registration of studies and analysis
30 plans.

31

32

33 Abstract

34

35 To make progress scientists need to know what other researchers have found and how they found it.
36 Unfortunately, transparency is often insufficient across much of ecology and evolution. Researchers
37 often fail to report results and methods with detail sufficient to permit interpretation and meta-analysis,
38 and many results go entirely unreported. Further, these unreported results are often a biased subset.
39 Thus the conclusions we can draw from the published literature are themselves often biased and
40 sometimes may be entirely incorrect. Fortunately there is a movement across empirical disciplines, and
41 now within ecology and evolution, to shape editorial policies to better promote transparency. This can
42 be done by either requiring more disclosure by scientists or by developing incentives to encourage
43 disclosure.

44

45

46 Science is a uniquely effective tool for understanding the world, and ecologists and evolutionary
47 biologists have built a robust body of scientific knowledge over the past century. However, several
48 common practices are limiting progress in these fields. For science to progress, results and clear
49 explanations of methods must be shared with other scientists. Although this fundamental principle is
50 widely understood, practices that cloud transparency of methods and results, such as selective reporting
51 (see glossary), appear far more common than they should be. This is unlikely to be an issue of deliberate
52 dishonesty, which we assume is rare in ecology and evolution. Instead, we believe that the unintended
53 negative consequences of insufficient transparency are often unrecognized by many members of the
54 scientific community. In addition, the institutions that shape our choices often inadvertently encourage
55 or reward choices that obstruct transparency [1]. Without sufficient transparency, we are hindered in
56 our ability to interpret published findings, conclusions based on published literature may be biased or
57 wrong, and meta-analytical syntheses are weakened [2]. Although these challenges to transparency vary
58 across disciplines and sub-disciplines, evidence suggests they are often common and present very real
59 problems for the advancement of ecology and evolutionary biology. In this paper, we first review
60 evidence of insufficient transparency in ecology and evolutionary biology, and then discuss new efforts
61 in these fields and in empirical science in general to improve transparency and thus improve scientific
62 progress.

63
64

65 The problems

66

67 Once researchers have collected and analyzed data, they commonly publish only a portion of the results
68 derived from these data (Fig. 1). Such selective reporting may lead to publication bias (see glossary) if
69 researchers preferentially publish certain types of results, such as those with the strongest or the most
70 surprising patterns. However, selective reporting is not limited to the classic ‘file-drawer’ problem in
71 which a study that does not produce the hoped-for result goes unpublished (e.g., [3]). For instance,
72 researchers may conduct multiple alternative forms of an analysis and report only the one with the
73 strongest relationships or lowest p -values. This practice has become known as ‘ p -hacking’ (see glossary)
74 [4, 5]. P -hacking and other forms of selective reporting can be masked by ‘HARKing’, or Hypothesizing
75 After Results are Known (see glossary)[6]. We may convince ourselves of the validity of selective
76 reporting in various ways. For instance, human cognitive tendencies, such as confirmation bias (see
77 glossary) (Box 1)[7], can lead researchers to select evidence that lends the clearest support for a pre-
78 existing hypothesis. Alternatively, selective reporting may not seem problematic as researchers often
79 tend to be more interested in the existence of patterns than in their absence. However, ignoring weak,
80 negative, or absent patterns is a major hindrance to our understanding of the biological world. First, the
81 absence of an effect or the presence of only a weak effect is itself important as we sort through
82 explanations of how biological systems work. Second, any observed statistical relationship is an estimate
83 of a true biological relationship, and as an estimate, it is inherently uncertain. Sampling variance results
84 in some estimates being higher than the true value, and some lower (Type M errors; see glossary), and
85 some being even opposite in sign (Type S error; see glossary) [8]. If we systematically eliminate the
86 smaller or contradictory effect sizes (see glossary) from publication, we get a biased picture of the size
87 of the true underlying effect, and under some circumstances this bias can be extreme [2]. Methods exist
88 for estimating the effect of publication bias in meta-analysis, but these methods are imperfect because
89 most are indirect and thus must make major assumptions about missing unpublished results whose true
90 values we can never know [9]. Therefore, the clearest path towards a reliable average is minimizing bias
91 in the original sample of statistical effects [2]. The selective reporting behind much publication bias
92 clearly varies among sub-disciplines and with the type of data reported, but evidence suggests it is
93 common in many areas of ecology and evolution, as in many other scientific disciplines. Most authors of
94 this manuscript have engaged in selective reporting at one or more points in their pasts, sometimes at
95 the request of reviewers or editors, and anecdotal evidence from conversations with others suggest it
96 may be widespread and frequent. However, it is not just our personal experience that suggests selective

97 reporting is common. There is considerable published empirical evidence for publication bias in ecology
98 and evolutionary biology.

99

100 Under-reporting (see glossary) is the easiest form of selective reporting to document because we know
101 the analysis was completed; the paper just fails to provide all the details of results or statistical methods.
102 For instance, studies may include means with no indication of uncertainty around those means, p -values
103 with no indication of the direction of the trend, or statistical results without the sample size for the
104 particular subset of data examined. These practices all limit readers' abilities to build an unbiased
105 understanding of a system and severely limit the usefulness of data for meta-analysis. A long and
106 growing list of surveys and meta-analyses has documented widespread under-reporting across many of
107 our sub-disciplines. Studies in fields including conservation [10], plant ecology [11], behavioral ecology
108 [12], ecosystem ecology [13, 14], community ecology [15], and others [16, 17] often find that around
109 half of published articles lack at least one key piece of information regarding statistical relationships
110 (Table 1). Further, where it has been examined these under-reported results were more likely to come
111 from non-significant comparisons or patterns contradictory to the primary hypothesis [18]. Finally, even
112 if authors report statistical results, they often do not report how the analyses were conducted in
113 sufficient detail, which makes it impossible for readers to critique the statistical methodology and to
114 replicate the analyses.

115

116 Estimating the rate at which results go completely unreported is more challenging. Results could remain
117 hidden from comparisons that authors decided were uninteresting. Unreported results might also come
118 from alternative versions of analyses conducted with, for instance, different covariates, interactions, or
119 subsets of data, as we might expect from p -hacking. One proposed method for identifying p -hacking is
120 ' p -curve' analysis, which predicts a clumping of p -values just below 0.05 if p -hacking is common [5].
121 Recently p -curve analysis was used to argue that p -hacking was having only modest impacts on biology
122 [4]. Regrettably, this reassuring conclusion is unwarranted. First, when researchers can include or
123 exclude covariates depending on their effects on p -values, p -values much smaller than 0.05 can often be
124 generated in the absence of a real effect [19, 20]. Thus, p -curve analysis focused on a 0.05 threshold can
125 dramatically underestimate p -hacking in fields where multiple covariates are common [19], such as
126 much of ecology and evolutionary biology. In fact, p -values have been shown to clump under lower
127 thresholds (0.01, 0.001, etc.) as well [21], as would be expected if p -hacking often ended with
128 calculation of a "highly significant" p -value. However, the second problem with these analyses is that
129 assumptions about the expected distribution of a collection of published p -values are almost certainly
130 incorrect, and thus inferring bias from the ' p -curve' is untenable under most conditions [22].

131

132 There are, however, other ways to estimate the magnitude of selective reporting. We can compare rates
133 of publication of statistically significant results with the observed distribution of statistical power (see
134 glossary) and estimates of average strength of effect. Rates of publication of statistically significant
135 effects are very high. In "Environment/Ecology" and "Plant and Animal Sciences", 74% of 150 and 78% of
136 200 statistical tests, each from a different randomly selected paper, were statistically significant and
137 supported the researchers' putative *a priori* hypotheses [23]. Similarly, in a cross-section of biological
138 journals, many from the disciplines of ecology and evolution, only 8.6% presented non-significant tests
139 of the main hypothesis [24]. Part of the explanation for these numbers is likely to be HARKing, in which
140 authors choose their strongest patterns and build the paper around those results, either de-emphasizing
141 or leaving out other results. While in some sub-fields of ecology and evolution researchers may often
142 test hypotheses that are likely to be true, this is probably not the case across all of ecology and
143 evolution. Further, even if most of our hypotheses were true, the proportion of statistically significant
144 results should be much lower since many of our studies have low statistical power. This low power
145 results from sample sizes that are often small, and average effect sizes that are also relatively small ($|r|$
146 = 0.19 [25], which should actually be an overestimate [26]) and thus difficult to detect (Box 2). The
147 resulting statistical power to detect effects of this observed average magnitude in the behavior, ecology,
148 and evolution literature is in the neighborhood of 20% [27, 28] (Box 2). If we thus conclude that typical

149 power is about 20% and we assume that 74% of tested hypotheses are true, we would still expect only
150 16% of findings to be statistically significant (Box 3) rather than 74%. This is a strong indication of
151 HARKing and selective reporting. Further, we discuss evidence below which suggests that published
152 statistically significant results may often be false or inflated relative to the true effect.

153
154 The proportion of significant results that are false positives is, somewhat counter-intuitively, increased
155 in studies with small samples and low power [29]. This increase happens because the probability of
156 detecting a true positive declines as power is reduced but the probability of detecting a false positive
157 remains fixed (typically at 0.05). As a consequence a greater proportion of positives will be false as
158 power decreases (Box 3). This means that reports of significant findings with low sample size should be
159 disproportionately likely to be incorrect [30], and of course such underpowered studies are common in
160 much of ecology and evolutionary biology [27].

161
162 Insufficient statistical power also hinders detection of real effects, and Type II errors (see glossary)
163 should thus also be common in ecology and evolution [31]. In fact, we predict that Type II error, when
164 they occur, will often go hand and hand with Type I error, as *p*-hacking extracts false positives from data
165 while true relationships go undetected. As described above, the rarity of negative results in the
166 literature suggests that Type II error is often concealed by HARKing, selective reporting, or both.

167
168 Much of our focus in this paper is on null hypothesis tests because these tests remain the most common
169 type of statistical analyses in ecology and evolution. However, it is important to note that most of the
170 choices related to sample size and selective reporting that can bias null hypothesis tests can bias other
171 threshold tests (e.g., Akaike information criterion: $\Delta AIC > 2$ [32]) and can also generate misleading and
172 inflated effect sizes. For instance, large effects reported from studies with small samples are likely to
173 often be inflated, or even of the wrong sign [30]. Examination of 3867 ecological studies from 52
174 previously published meta-analyses showed that studies with the largest effect sizes tended to have the
175 lowest samples sizes [33]. Further, 'p-hacking' could also be considered 'effect-size hacking' since the
176 same practices produce inflated effect sizes, and if combined with selective reporting, produce a
177 distribution of published effects that is biased upwards.

178
179 Given that studies with larger effects may be more likely to end up in journals with higher impact scores
180 [34], perhaps high impact journals are often publishing studies with large effects despite their small
181 samples and unreliability. Although there is evidence that in some subsets of the published literature
182 sample size and journal impact factor are negatively correlated, this trend appears to vary across study
183 types, and when averaged across a large number of studies ($n = 3867$), impact factor was uncorrelated
184 with sample size [33]. While this lack of correlation is certainly better than a consistent negative
185 correlation, given that studies with larger samples produce more reliable results, it would actually be
186 preferable to see a positive relationship between sample size and journal impact factor. Further, it is
187 effect size, not sample size, that predicts the number of citations a study receives [33]. So, not only are
188 published studies with small sample sizes more likely to report inflated effects (i.e. more prone to Type
189 M errors), the unreliability of these studies does not dependably deter their publication in high impact
190 journals or their accumulation of citations.

191
192 It has long been established that as the number of statistical comparisons increases, the probability of
193 observing patterns that result only from chance (i.e., false positives) also increases [35]. This happens
194 both with multiple separate tests or if, instead of alternative tests, we combine multiple possible
195 predictors in the same model [36]. Within a single model we might include a set of different equally
196 plausible predictors of the variable of interest, or we might include multiple alternative interaction
197 terms between our predictor of interest and different covariates. In a survey of 50 randomly selected
198 studies from ecology and evolution, 28 studies (56%) used GLMs with two or more predictors [36], and
199 none of these 28 considered any type of correction for multiple comparisons to counter the risk of
200 inflated significance. We could not locate other attempts to quantify failures to correct for multiple

201 comparisons, but uncorrected multiple comparisons appear common in at least some portions of the
202 literature [12]. Although false positives from multiple comparisons in exploratory analyses need not be a
203 major problem if we recognize the provisional nature of the results [35], two current practices in our
204 disciplines make uncorrected multiple comparisons a severe issue. First, multiple comparisons are often
205 hidden, with researchers conducting multiple tests but only reporting a subset of them. Thus the
206 likelihood that a result is a false positive is concealed and the scientific community is misled about the
207 probability that the result is true. Second, calls for tolerating a high false positive rate (to reduce Type II
208 errors) emphasize the importance of validating findings with replication studies [35], but replications or
209 other types of independent evaluation are currently far too rare to sort out the false from the true
210 positives [37, 38].

211
212 The problems outlined above are heavily influenced by the institutions that shape the decisions of
213 researchers, including journals, funding bodies, and employers. Calls for individual scientists to improve
214 transparency are not uncommon [e.g., 39, 40, 41], and scientists sometimes respond to these calls.
215 However, individual scientists, like all people, make decisions in response to the institutions in which
216 they operate [1]. Funders reward novelty, typically to the complete exclusion of replication, and journals
217 preferentially publish statistically significant findings, especially if those findings are surprising. These
218 factors alone would influence researchers' decisions, but these incentives are even more influential
219 because universities and research institutes often hire and promote scientists based on their record of
220 acquiring grant money and the number and impact factors of their publications [1]. Thus to increase
221 transparency, we should identify components of this incentive structure amenable to improvement.

222
223 Some solutions

224
225 There is growing recognition of the problems hindering empirical progress and of the role that
226 institutions must play in shaping science in ecology, evolutionary biology, and beyond [42-44]. In
227 November 2015, representatives (mostly editors-in-chief) from nearly 30 journals in ecology and
228 evolution joined funding agency panelists and other researchers to identify ways to improve
229 transparency in these disciplines. At this workshop, strong support emerged for the recently introduced
230 Transparency and Openness Promotion (TOP) framework (<https://cos.io/top/>)[45]. TOP currently
231 consists of eight guidelines that can be implemented by journals and funding agencies. Institutions can
232 adopt whichever of the eight guidelines they choose, and they can implement these guidelines along a
233 gradient of stringency. The rapid and extensive spread of support for TOP (>500 journals in < 1 year)
234 across scientific disciplines appears to herald a revolution in transparency standards.

235
236 Several TOP guidelines simply request or require more thorough reporting of methods, results, data, or
237 analysis code. Ecologists and evolutionary biologists made important progress in this regard several
238 years ago when a growing number of journals began requiring the archiving of data [46]. Calls for more
239 expanded archiving are growing in ecology and evolution [47], and the TOP guidelines can facilitate the
240 expansion of these types of disclosures. Interestingly, an incentive to archive in the form of a badge may
241 be similarly effective [48] as requiring archiving [49] and could therefore eliminate much of the
242 controversy regarding archiving [e.g., 50]. The TOP guideline titled 'analysis and design transparency'
243 calls for discipline-specific guidance regarding what information should be disclosed in publications, and
244 to that end, the workshop produced a document 'Tools for Transparency in Ecology and Evolution'
245 (TTEE; <https://osf.io/g65cb/>) that provides checklist questions that journals can provide to authors,
246 reviewers, and editors to facilitate transparent reporting. Promoting more thorough and consistent
247 reporting of results and methods through TOP and TTEE should dramatically improve transparency, but
248 here we also highlight two other TOP components that could have transformative impacts on our field.

249
250 Pre-registration (see glossary), in which researchers register their study and data analysis plan prior to
251 collecting data, can greatly improve transparency. Although requiring pre-registration (as in clinical trial
252 research)[51] might thwart publication of valuable exploratory and serendipitous findings in ecology and

253 evolution, encouraging pre-registration where appropriate has large potential benefits. Most obviously,
254 it makes unpublished results more discoverable [45], thus helping to reduce publication bias. Potentially
255 more important, however, pre-registration of analysis plans ensures that we can identify genuine *a*
256 *priori* planned tests, helping to improve confidence in results because they are unlikely to result from
257 hidden multiple hypothesis testing and selective reporting. As pre-registration becomes more common,
258 results that do not come from pre-registered analysis plans become viewed as exploratory, and thus
259 provisional and less convincing than pre-registered results [52], providing a strong incentive to pre-
260 register studies. We acknowledge that exploratory work is hugely important in ecology and evolutionary
261 biology and we do not wish to impede it, but it should be more consistently identifiable and it should be
262 follow-up with planned, ideally pre-registered, tests [35]. A common concern is that pre-registration
263 ignores the inevitable tweaking of methods that occurs as field projects evolve. However, alterations to
264 methods or analysis plans can be justified in the published study [e.g., 48]. Reviewers and editors can
265 decide if the reported methods adhered closely enough to the pre-registration to earn a pre-registration
266 badge (<https://osf.io/tvyxz/wiki/home/>). Further, pre-registered analyses and exploratory results can be
267 published in the same paper when the distinction between them is made clear. In an effort to further
268 jump start the pre-registration process, the Center for Open Science recently announced the Pre-
269 registration Challenge, in which the first thousand researchers to publish pre-registered research will be
270 awarded US\$1000 each (<https://cos.io/prereg/>). Independently, institutions promoting systematic
271 reviews in ecology and conservation have also been encouraging pre-registration
272 (<http://www.environmentalevidence.org/>; <http://cebc.bangor.ac.uk/>).

273
274 The final TOP guideline promotes replications (see glossary) of previously published studies. Replication
275 to assess validity and generality of prior results is a core practice of science. Exact replication is not
276 possible, especially in field studies, but various forms of replication, especially when combined with
277 meta-analysis, are powerful tools for establishing the applicability of hypotheses [37]. Unfortunately,
278 institutional incentive structures often work strongly against replication in ecology and evolution,
279 especially replications that seek to closely match methods as part of the process of assessing validity
280 [37]. Journals and funding bodies explicitly favor novelty. Of course progress requires novelty, but
281 progress also requires rigorous evaluation of prior findings. Not all studies are of high priority for
282 replication. The more interesting or important a finding, however, the more important it is to replicate
283 that study. Allocating funding to replication would certainly increase its frequency, as would journals
284 adopting policies that explicitly encourage submission of replications (e.g.,
285 <http://biotropica.org/reproducibility-repeatability/>). As with any other articles, journals can obviously
286 reject less valuable replication studies. For instance, journals might require sample sizes larger than in
287 the original study, review of methods prior to conducting the research (i.e., ‘registered reports’; see
288 glossary) [53], or replications only of original studies that cross some threshold of impact or interest.
289 Replication is an essential part of doing science in other fields, as, for example, anyone who remembers
290 the ‘cold fusion in a jar’ debacle of 1989 can attest [54].

291
292 As institutions in ecology and evolutionary biology more vigorously promote transparency, we will
293 become better able to evaluate the results we read, the average result will be more reliable, and there
294 will be clearer paths for empirical progress (Fig. 1). We need to deliberately shape the institutions in
295 which we operate to best facilitate scientific progress. Not all institutions will be equally responsive to
296 attempts at reform. However, we already know that journals can take deliberate steps to increase
297 transparency [46], and in response to the TEE workshop mentioned above, nearly 30 ecology and
298 evolution journals are engaged in ongoing discussions about adopting TOP guidelines or have already
299 adopted these guidelines. Funding agencies have also implemented data archiving policies [46] and
300 could promote transparency in multiple other ways as guided by TOP. The proposals we review here are
301 only a subset of possible solutions to insufficient transparency. We hope to stimulate a continuing
302 exploration of these issues. This is an historic crossroads for the practice of science in ecology and
303 evolutionary biology, and for empirical disciplines in general [45].

304 Acknowledgements

305

306 We thank Mark Elgar for requesting an aggregation of evidence regarding the current state of
307 transparency in ecology and evolution. This request was made at the November 2015 workshop titled
308 “Improving Inference in Evolutionary Biology and Ecology.” Other participants at this workshop
309 (complete list: <https://osf.io/dhp3t/>) were also vital contributors to discussions that inspired this paper.
310 Financial support for the workshop was provided by the US National Science Foundation (DEB: 1548207)
311 and The Laura and John Arnold Foundation, and logistical support was provided by the Center for Open
312 Science. ARC Future Fellowships supported S. N. (FT130100268) and F. F. (FT150100297). We also thank
313 Losia Lagisz for helping to make Figure 1. Comments from an anonymous reviewer significantly
314 improved the manuscript.

315 Glossary

316

317 **Blind observation:** The observer (person making measurements) is unaware of the group membership
318 (e.g., treatment condition) of the subject being measured

319

320 **Confirmation bias:** The widespread human tendency to interpret observations as consistent with one's
321 belief about how the world works or to preferentially search for and recall such observations

322

323 **Effect size:** A measure of study outcome that indicates the magnitude and direction of the outcome of
324 each study. Effect sizes can be based on the magnitude of difference between groups or the strength of
325 the correlation between variables. Effect sizes can be unstandardized (e.g., mean difference or
326 covariance) or standardized (e.g., Cohen's *d* or correlation coefficient).

327

328 **Exploratory analysis:** conducting many graphical and/or statistical comparisons in an effort to identify
329 previously unidentified relationships among variables in a data set

330

331 **False positive:** In null hypothesis testing, a rejection of the null hypothesis when the null hypothesis is
332 actually true (Type I error)

333

334 **HARKing:** Hypothesizing After Results are Known – presenting a *post hoc* explanation for an exploratory
335 result as though it were an *a priori* hypothesis. Many of us were taught to HARK and to write papers as
336 though we were testing *a priori* hypotheses even if we were conducting exploratory analyses. Although
337 philosophers debate the importance of distinguishing between *a priori* and *post hoc* hypotheses,
338 HARKing is problematic even if one discounts this distinction. This is because HARKing often serves to
339 conceal selective reporting of exploratory analyses (often without a deliberate attempt to deceive), and
340 thus skews the distribution of reported results.

341

342 **Inflated effect size:** An estimated effect size that is larger than the actual effect size, for instance
343 because the researcher selected the covariate that led to the largest effect in the target relationship
344 after testing multiple covariates

345

346 **Meta-analysis:** The quantitative synthesis of the outcomes of different studies, based on combining
347 effect sizes, to determine overall results across studies and sources of heterogeneity in outcomes among
348 studies. Generally study outcomes are weighted by the precision with which the effects are estimated.
349 Meta-regression is a variant of meta-analysis in which the effects of covariates are modeled statistically.

350

351 ***p*-hacking:** A variety of practices that increase the odds of finding a statistically significant result by, for
352 instance, conducting multiple versions of an analysis with different covariates, interactions, or subsets of
353 data. Some processes that contribute to *p*-hacking, such as conducting multiple versions of an analysis
354 with different interaction terms, may be pursued out of a sincere desire to discover the story the data
355 have to tell. However, each additional version of the analysis increases the risk of a false positive or of
356 an inflated effect, and unless we disclose all results from all versions of analyses and all decisions
357 regarding data gathering and analyses, we will contribute to the biased distribution of effects in the
358 literature.

359

360 **Pre-registration:** A process by which planned studies, including methods and an analysis plan, are
361 registered in a secure and accessible platform (e.g. website such as Open Science Framework;
362 <https://osf.io/>) before commencement of the research. Once a pre-registration has been submitted, it
363 cannot be altered. Pre-registrations can be embargoed to protect ideas prior to publication.

364

365 **Publication bias:** A bias in the distribution of published effect sizes resulting from any number of factors,
366 including selective reporting by authors and rejection of non-significant results by editors

367
368 **Registered report:** A study in which the rationale, methods, and analysis plan are submitted to a journal
369 for review, and possible revision, with the objective of achieving in-principle acceptance based on the
370 importance of the question and the quality of the study design, not the outcome, prior to initiation of
371 the study.
372
373 **Replication:** a study designed to replicate a previously published result, either by closely following the
374 original methods in an effort to assess validity ('direct' or 'close' replication) or by designing a study
375 inspired by the original concept in an effort to assess generality ('conceptual replication')
376
377 **Selective reporting:** Reporting only a subset of analyses conducted. In medicine, a similar concept is
378 often referred to as reporting bias.
379
380 **Statistical power:** The probability of detecting a statistically significant effect if that effect actually exists.
381 This probability is a function of the significance threshold, sample size, and strength of statistical effect.
382
383 **Type I error:** Rejection of a null hypothesis when the null hypothesis is true (a 'false positive').
384
385 **Type II error:** a failure to reject a null hypothesis when the null hypothesis is false (a 'false negative')
386
387 **Type M error:** an error in estimating the magnitude of an effect
388
389 **Type S error:** an error in estimating the sign of an effect
390
391 **Under-reporting:** Reporting an analysis without sufficient details of analytical methods or results to
392 allow for interpretation
393
394
395
396

397 Text boxes

398

399 Text Box 1

400

401 Confirmation bias

402

403 People have a strong tendency to interpret observations as supporting their existing worldview and to
404 seek out evidence in support of this worldview [7]. This can play out in various forms of selective
405 reporting as we convince ourselves that we are simply focusing our reporting on the real phenomena.
406 Confirmation bias can thus help rationalize p -hacking and selective reporting, often by preventing us
407 from recognizing our own subtle HARKing. Confirmation bias can also influence data gathering. Studies
408 in ecology and evolution in which individuals gathering data were not blind to the treatment condition
409 or the predicted outcomes showed stronger effects and higher rates of significance than studies with
410 blinded observers [55, 56]. Blind observation (see Glossary) is quite rare in ecology and evolutionary
411 biology [57] in part because in some studies blinding is nearly impossible. However, in a large sample of
412 recent studies, 56% that could have benefited from blinding could also have implemented it with little
413 difficulty (e.g., no additional personnel), and an additional 22% could have adopted blinding by
414 employing an observer naïve to certain details of the study [57].

415

416

417 Text Box 2

418

419 Evidence of low power

420

421 In a sample of 1362 statistical tests from 697 papers published in 2000 in 10 behavior, evolution, and
422 ecology journals, the average power to detect a small effect ($|r| = 0.1$) was only 13-16% [27]. In other
423 words, studies would only be expected to reject a false null hypothesis 13-16% of the time in the case of
424 weak effects. Power to detect medium ($|r| = 0.3$) and large ($|r| = 0.5$) effects, though of course higher
425 (40-47% and 65-72%, respectively), was still typically well below the commonly recommended threshold
426 of 80%. Examined another way, the proportion of studies reaching this 80% power threshold to detect
427 weak effects was 2-3%, 13-21% for medium effects, and 37-50% for strong effects [27]. Other analyses
428 of power find similar results. For example, an analysis of studies published in *Animal Behaviour* in 1996,
429 2003, and 2009 found, across all three years, an average power of just 23-26% for detection of medium
430 effects and 1-2% for weak effects [28]. It thus appears that studies in ecology and evolution often lack
431 power to detect small and medium effects, and this is particularly problematic because effects in
432 ecology and evolution tend to be weak. Average effects across 43 meta-analyses in ecology and
433 evolutionary biology were found to be weak to moderate ($|r| = 0.18-0.19$) [25]. Further, these rather
434 low values are actually overestimates because averages of estimated absolute values of effect size are
435 upwardly biased [26]. To detect these relatively small effects requires large samples (e.g., $n = 207$ to
436 obtain an 80% probability of detecting a true effect of $r = 0.193$) [25], but obtaining sufficient power
437 through large samples is rare [27].

438

439 Text Box 3

440 False-positive report probability (FPRP)

441

442 In many sub-fields of evolution and ecology it remains common to use a significance threshold of 5%.
 443 This means that if our null hypothesis were true we would incorrectly reject it 5% of the time. However,
 444 we often incorrectly attribute a frequency of 5% to a different phenomenon: the chance that a
 445 significant finding is a false positive. This is incorrect because the probability that a positive result is a
 446 false positive depends on three factors (1) the proportion of our hypotheses that are in fact true (π , the
 447 probability that a hypothesis is true), (2) the significance threshold (α), and (3) statistical power ($1 - \beta$,
 448 where β is the probability of making a type II error; Table 1): $FPRP = (\alpha(1 - \pi)/[\alpha(1 - \pi) + (1 - \beta)\pi]$. With
 449 50% of our hypotheses true and statistical power of 20% (a power typical in ecology and evolution [25]),
 450 the chance that a significant finding is a false positive is 20%. This value is known as the false positive
 451 report probability [58]. This number is notably larger than 5%, but it becomes dramatically larger when,
 452 in pursuit of novelty, we turn our interest towards testing relatively unlikely hypotheses, those that in
 453 the Bayesian sense could be said to have a low prior probability. For instance, when only 10% of tested
 454 hypotheses are in fact true, the expected false positive report probability rises to 69% $((0.05(1 -$
 455 $0.1)/((0.05(1 - 0.1) + (0.2)0.1))$ [58]! In fact, false positives could be even more prevalent. The above
 456 calculations assume complete and transparent reporting of the full set of analyses conducted, as
 457 promoted by pre-registration and other recently proposed transparency tools. If, in contrast,
 458 researchers make their choices of analysis strategy conditional on the outcome as with *p*-hacking (i.e.
 459 preferring test variants that yield significance or stronger effects) then the false-report probability
 460 increases further.

461

462 I. Four possible outcomes from a null hypothesis statistical test together with the probabilities of
 463 each outcome depending on whether the null-hypothesis is true

| | Null Hypothesis True | Alternate Hypothesis True |
|-------------------------|-----------------------------|----------------------------|
| Significant Finding | False Positive: α | True Positive: $1 - \beta$ |
| Non-Significant Finding | True Negative: $1 - \alpha$ | False Negative: β |

464

465

466

467

468 Tables

469

470 Table 1. A sample of studies in ecology and evolution that quantify rates of under-reporting of important
471 details of methods or results in the published literature.

472

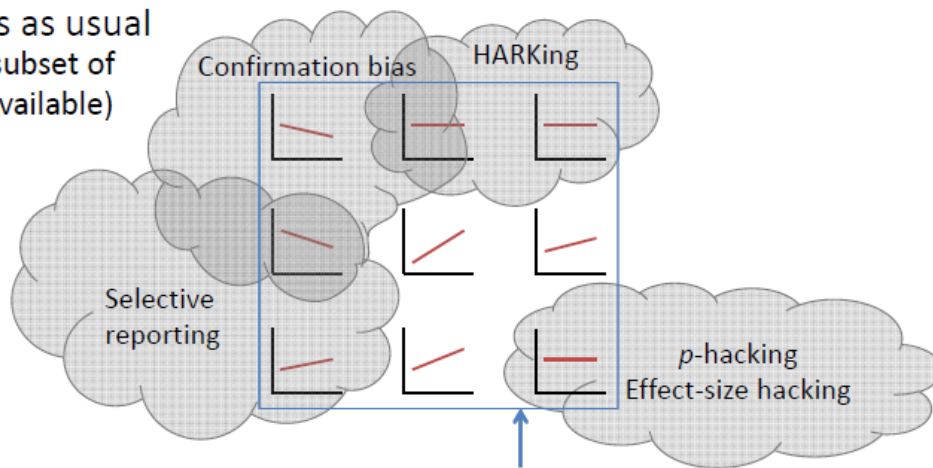
| Citation | Studies reviewed | finding |
|------------------------|---|---|
| Ferreira et al. (2015) | 99 studies of litter decomposition in streams as an effect of nutrient enrichment | Estimates of decomposition rate presented without estimate of uncertainty in 54% of studies (even after requesting details directly from authors) |
| Fidler et al. (2006) | 78 articles published in 2005 in Conservation Biology and Biological Conservation | 58% missing at least one effect size 51% missing at least one sample size 85% missing at least one SE or SD |
| Parker (2013) | 48 studies of plumage color in a well-studied European songbird species | 409 of 997 main-effect relationships lacked information to estimate the strength and/or direction of the effect |
| Zhang et al. (2012) | 54 studies of forest productivity as a function of tree diversity | 29 studies failed to provide either estimates of variance associated with means or corresponding sample sizes |

473

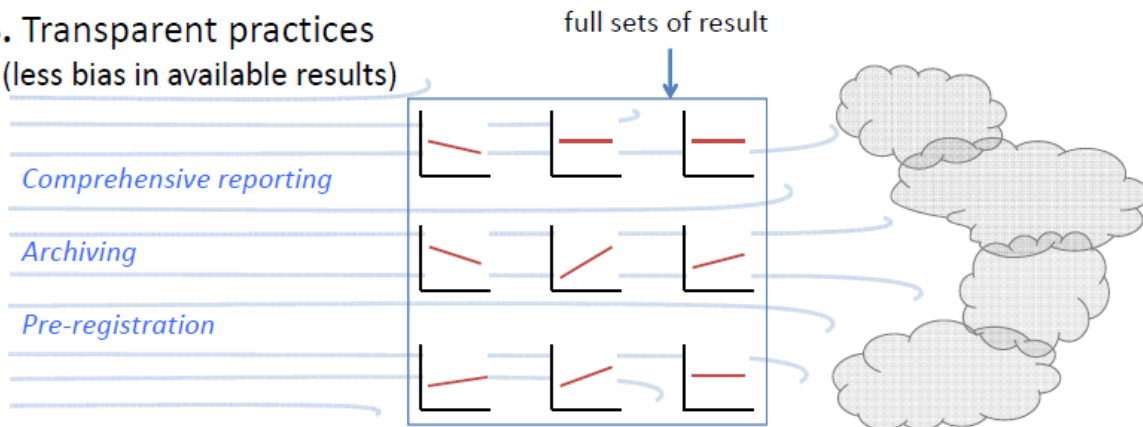
474

475

A. Business as usual
(biased subset of results available)



B. Transparent practices
(less bias in available results)



478
479 Figure 1. ‘Business as usual’ in ecology and evolution allows and often promotes practices that keep
480 many analyses hidden and this leads to biases in the published literature. For example, current practices
481 (A) could result in only the three ‘unclouded’ graphs making it to publication, leaving the impression that
482 all results were consistently positive. However, full transparency (B) will sometimes leave a very
483 different impression of results. In this illustration, we see results that are more complicated and less
484 consistent, and suggest a much smaller average effect, if any.
485

486 Literature Cited

487

- 488 1. Smaldino, P.E. and McElreath, R. (2016) The natural selection of bad science. *arXiv*,
489 1605.19511v19511.
- 490 2. Møller, A.P. and Jennions, M.D. (2001) Testing and adjusting for publication bias. *Trends in*
491 *Ecology & Evolution* 16, 580-586.
- 492 3. Godefroid, S., *et al.* (2011) How successful are plant species reintroductions? *Biological*
493 *Conservation* 144, 672-682.
- 494 4. Head, M.L., *et al.* (2015) The extent and consequences of *P*-Hacking in science. *PLoS Biol* 13,
495 e1002106.
- 496 5. Simonsohn, U., *et al.* (2014) *P*-curve: a key to the file drawer. *Journal of Experimental*
497 *Psychology: General* 143, 534-547.
- 498 6. Kerr, N.L. (1998) HARKing: hypothesizing after the results are known. *Personality and Social*
499 *Psychology Review* 2, 196-217.
- 500 7. Nickerson, R.S. (1998) Confirmation bias: A ubiquitous phenomenon in many guises. *Review of*
501 *General Psychology* 2, 175-220.
- 502 8. Gelman, A. (2015) Working through some issues. *Significance* 12, 33-35.
- 503 9. Rothstein, H.R., *et al.*, eds (2005) *Publication bias in meta-analysis: prevention, assessment and*
504 *adjustments*. John Wiley & Sons, Lt.
- 505 10. Fidler, F., *et al.* (2006) Impact of criticism of null-hypothesis significance testing on statistical
506 reporting practices in conservation biology. *Conservation Biology* 20, 1539-1544.
- 507 11. Koricheva, J. and Gurevitch, J. (2014) Uses and misuses of meta-analysis in plant ecology. *Journal*
508 *of Ecology* 102, 828-844.
- 509 12. Parker, T.H. (2013) What do we really know about the signalling role of plumage colour in blue
510 tits? A case study of impediments to progress in evolutionary biology. *Biological Reviews* 88,
511 511-536.
- 512 13. Ferreira, V., *et al.* (2015) A meta-analysis of the effects of nutrient enrichment on litter
513 decomposition in streams. *Biological Reviews* 90, 669-688.
- 514 14. Menge, D.N.L. and Field, C.B. (2007) Simulated global changes alter phosphorus demand in
515 annual grassland. *Global Change Biology* 13, 2582-2591.
- 516 15. Zhang, Y., *et al.* (2012) Forest productivity increases with evenness, species richness and trait
517 variation: a global meta-analysis. *Journal of Ecology* 100, 742-749.
- 518 16. Leisner, C.P. and Ainsworth, E.A. (2012) Quantifying the effects of ozone on plant reproductive
519 growth and development. *Global Change Biology* 18, 606-616.
- 520 17. Moles, A.T., *et al.* (2011) Assessing the evidence for latitudinal gradients in plant defence and
521 herbivory. *Functional Ecology* 25, 380-388.
- 522 18. Cassey, P., *et al.* (2004) A survey of publication bias within evolutionary ecology. *Proceedings of*
523 *the Royal Society of London B: Biological Sciences* 271, S451-S454.
- 524 19. Bruns, S.B. and Ioannidis, J.P.A. (2016) *p*-Curve and *p*-Hacking in observational research. *PLoS*
525 *ONE* 11, e0149144.
- 526 20. Bishop, D.V.M. and Thompson, P.A. (2016) Problems in using *p*-curve analysis and text-mining to
527 detect rate of *p*-hacking and evidential value. *PeerJ* 4, e1715.
- 528 21. Ridley, J., *et al.* (2007) An unexpected influence of widely used significance thresholds on the
529 distribution of reported *P*-values. *J. Evol. Biol.* 20, 1082-1089.
- 530 22. Gelman, A. and O'Rourke, K. (2014) Discussion: Difficulties in making inferences about scientific
531 truth from distributions of published *p*-values. *Biostatistics* 15, 18-23.
- 532 23. Fanelli, D. (2010) "Positive" results increase down the hierarchy of the sciences. *PLoS ONE* 5,
533 e10068.
- 534 24. Csada, R.D., *et al.* (1996) The "file drawer problem" of non-significant results: does it apply to
535 biological research? *Oikos* 76, 591-593.

- 536 25. Møller, A.P. and Jennions, M.D. (2002) How much variance can be explained by ecologists and
537 evolutionary biologists? *Oecologia* 132, 492-500.
- 538 26. Hereford, J., *et al.* (2004) Comparing strengths of directional selection: how strong is strong?
539 *Evolution* 58, 2133-2143.
- 540 27. Jennions, M.D. and Møller, A.P. (2003) A survey of the statistical power of research in behavioral
541 ecology and animal behavior. *Behav. Ecol.* 14, 438-445.
- 542 28. Smith, D.R., *et al.* (2011) Power rangers: no improvement in the statistical power of analyses
543 published in *Animal Behaviour*. *Animal Behaviour* 81, 347-352.
- 544 29. Button, K.S., *et al.* (2013) Power failure: why small sample size undermines the reliability of
545 neuroscience. *Nat Rev Neurosci* 14, 365-376.
- 546 30. Gelman, A. and Weakliem, D. (2009) Of beauty, sex, and power. *American Scientist* 97, 310-316.
- 547 31. Eberhardt, L.L. and Thomas, J.M. (1991) Designing environmental field studies. *Ecological*
548 *Monographs* 61, 53-73.
- 549 32. Murtaugh, P.A. (2014) In defense of *P* values. *Ecology* 95, 611-617.
- 550 33. Barto, E.K. and Rillig, M.C. (2012) Dissemination biases in ecology: effect sizes matter more than
551 quality. *Oikos* 121, 228-235.
- 552 34. Murtaugh, P.A. (2002) Journal quality, effect size, and publication bias in meta-analysis. *Ecology*
553 83, 1162-1166.
- 554 35. Pike, N. (2011) Using false discovery rates for multiple comparisons in ecology and evolution.
555 *Methods in Ecology and Evolution* 2, 278-282.
- 556 36. Forstmeier, W. and Schielzeth, H. (2011) Cryptic multiple hypotheses testing in linear models:
557 overestimated effect sizes and the winner's curse. *Behavioral Ecology and Sociobiology* 65, 47-
558 55.
- 559 37. Nakagawa, S. and Parker, T.H. (2015) Replicating research in ecology and evolution: feasibility,
560 incentives, and the cost-benefit conundrum. *BMC Biology* 13, 88.
- 561 38. Kelly, C.D. (2006) Replicating empirical research in behavioral ecology: how and why it should be
562 done but rarely ever is. *Q. Rev. Biol.* 81, 221-236.
- 563 39. Birkhead, T.R. (2002) Of Moths and Men (book review). *International Society for Behavioral*
564 *Ecology Newsletter* 14, 15-16.
- 565 40. Nakagawa, S. and Cuthill, I.C. (2007) Effect size, confidence interval and statistical significance: a
566 practical guide for biologists. *Biological Reviews* 82, 591-605.
- 567 41. Belovsky, G.E., *et al.* (2004) Ten suggestions to strengthen the science of ecology. *BioScience* 54,
568 345-351.
- 569 42. Baker, M. (2016) 1,500 scientists lift the lid on reproducibility. *Nature* 533, 452-454.
- 570 43. Parker, T.H. and Nakagawa, S. (2014) Mitigating the epidemic of type I error: ecology and
571 evolution can learn from other disciplines. *Frontiers in Ecology and Evolution* 2.
- 572 44. Open_Science_Collaboration (2015) Estimating the reproducibility of psychological science.
573 *Science* 349.
- 574 45. Nosek, B.A., *et al.* (2015) Promoting an open research culture. *Science* 348, 1422-1425.
- 575 46. Whitlock, M.C. (2011) Data archiving in ecology and evolution: best practices. *Trends in Ecology*
576 *& Evolution* 26, 61-65.
- 577 47. Mislán, K.A.S., *et al.* (2016) Elevating the status of code in ecology. *Trends in Ecology & Evolution*
578 31, 4-7.
- 579 48. Kidwell, M.C., *et al.* (2016) Badges to acknowledge open practices: a simple, low cost, effective
580 method for increasing transparency. *PLOS Biology* 14, e1002456.
- 581 49. Roche, D.G., *et al.* (2015) Public data archiving in ecology and evolution: how well are we doing?
582 *PLoS Biology* 13, e1002295.
- 583 50. Mills, J.A., *et al.* (2015) Archiving primary data: solutions for long-term studies. *Trends in Ecology*
584 *& Evolution* 30, 581-589.
- 585 51. Ross, J.S., *et al.* (2009) Trial publication after registration in ClinicalTrials.gov: a cross-sectional
586 analysis. *PLoS Med* 6, e1000144.

- 587 52. Wagenmakers, E.-J., *et al.* (2012) An agenda for purely confirmatory research. *Perspectives on*
588 *Psychological Science* 7, 632-638.
- 589 53. Chambers, C., D. (2013) *Registered Reports*: a new publishing initiative at *Cortex*. *Cortex* 49, 609-
590 610.
- 591 54. Huizenga, J.R. (1994) *Cold Fusion: The Scientific Fiasco of the Century*. Oxford University Press.
- 592 55. van Wilgenburg, E. and Elgar, M.A. (2013) Confirmation bias in studies of nestmate recognition:
593 a cautionary note for research into the behaviour of animals. *PLoS ONE* 8, e53548.
- 594 56. Holman, L., *et al.* (2015) Evidence of experimental bias in the life sciences: why we need blind
595 data recording. *PLoS Biol* 13, e1002190.
- 596 57. Kardish, M.R., *et al.* (2015) Blind trust in unblinded observation in ecology, evolution and
597 behavior. *Frontiers in Ecology and Evolution* 3, 51.
- 598 58. Wacholder, S., *et al.* (2004) Assessing the probability that a positive report is false: an approach
599 for molecular epidemiology studies. *Journal of the National Cancer Institute* 96, 434-442.

600