Transparency in ecology and evolution: real problems, real solutions

Transparency in ecology and evolution: real problems, real solutions

Timothy H. Parker, Whitman College, Walla Walla, USA
Wolfgang Forstmeier, Max Planck Institute for Ornithology, Seewiessen, Germany
Julia Koricheva, Royal Holloway University of London, Egham, UK
Fiona Fidler, University of Melbourne, Melbourne, Australia
Jarrod Hadfield, University of Edinburgh, Edinburgh, UK
Yung En Chee, University of Melbourne, Melbourne, Australia
Clint Kelly, University of Quebec, Montreal, Canada
Jessica Gurevitch, Stony Brook University, New York, USA
Shinichi Nakagawa, University of New South Wales, Sydney, Australia

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

Highlights (separate from the manuscript):

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

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

Abstract

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

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

The problems

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

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

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

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

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

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

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

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

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

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

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

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

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

Some solutions

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

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

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

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

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

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

Acknowledgements

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

Glossary

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

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

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

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

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

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

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

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

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

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

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

Registered report: A study in which the rationale, methods, and analysis plan are submitted to a journal for review, and possible revision, with the objective of achieving in-principle acceptance based on the importance of the question and the quality of the study design, not the outcome, prior to initiation of the study.

Replication: a study designed to replicate a previously published result, either by closely following the original methods in an effort to assess validity ('direct' or 'close' replication) or by designing a study inspired by the original concept in an effort to assess generality ('conceptual replication')

Selective reporting: Reporting only a subset of analyses conducted. In medicine, a similar concept is often referred to as reporting bias.

Statistical power: The probability of detecting a statistically significant effect if that effect actually exists. This probability is a function of the significance threshold, sample size, and strength of statistical effect.

Type I error: Rejection of a null hypothesis when the null hypothesis is true (a 'false positive').

Type II error: a failure to reject a null hypothesis when the null hypothesis is false (a 'false negative')

Type M error: an error in estimating the magnitude of an effect

Type S error: an error in estimating the sign of an effect

Under-reporting: Reporting an analysis without sufficient details of analytical methods or results to allow for interpretation

Text boxes

Text Box 1

Confirmation bias

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

Text Box 2

Evidence of low power

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

Text Box 3

False-positive report probability (FPRP)

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

I. Four possible outcomes from a null hypothesis statistical test together with the probabilities of 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

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

Figures

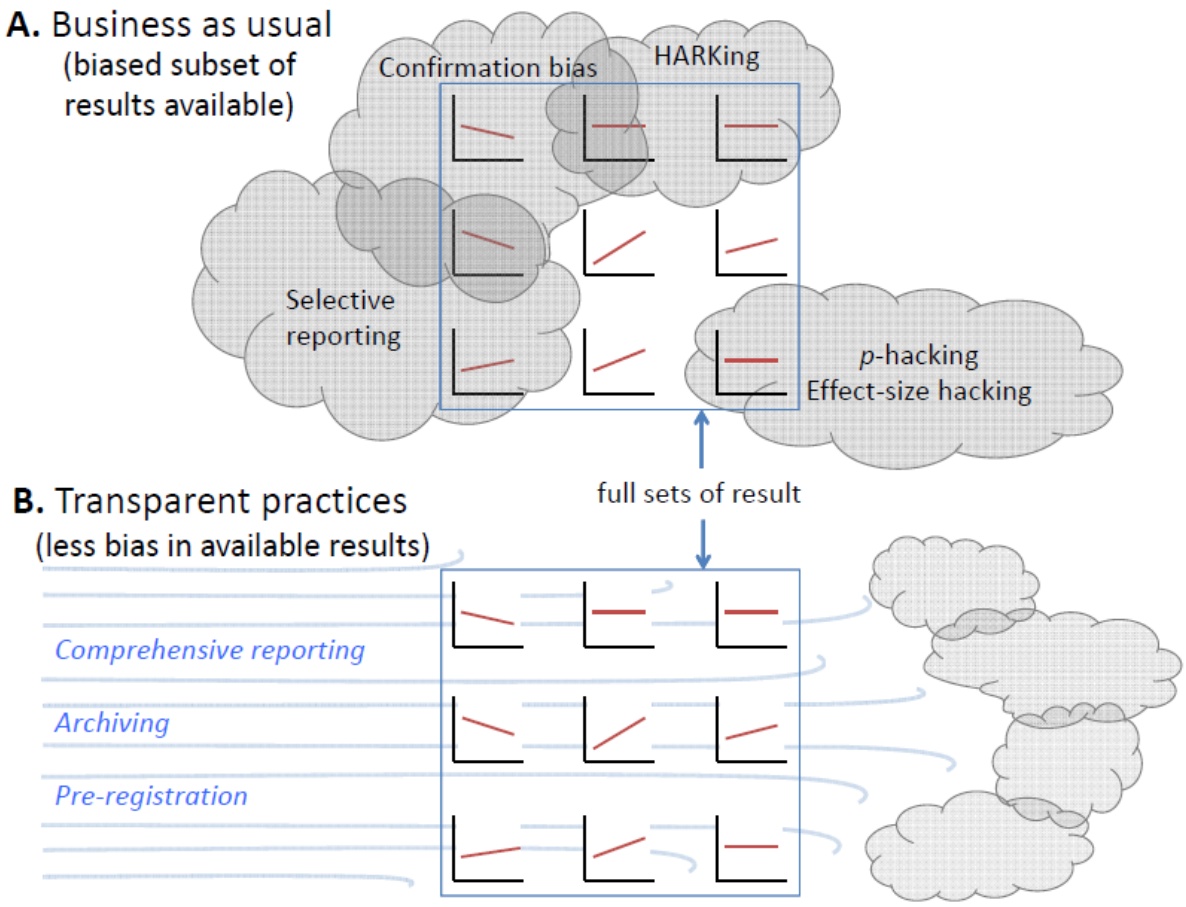

Figure 1. 'Business as usual' in ecology and evolution allows and often promotes practices that keep many analyses hidden and this leads to biases in the published literature. For example, current practices (A) could result in only the three 'unclouded' graphs making it to publication, leaving the impression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

Literature Cited

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

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

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