DATE: 28 June 2023

Title

Empirical evidence of widespread exaggeration bias and selective reporting in ecology

Authors: Kaitlin Kimmel[1,2], Meghan L. Avolio[2], and Paul J. Ferraro[3,4]

1- Mad Agriculture, Boulder, CO, USA

2 - Department of Earth and Planetary Sciences, Johns Hopkins University, Baltimore, MD, USA

3- Carey Business School, Johns Hopkins University, Baltimore, MD, USA

4- Department of Environmental Health and Engineering, a joint department of the Bloomberg School of Public Health and the Whiting School of Engineering, Johns Hopkins University, Baltimore, MD, USA

Corresponding author: Ferraro, PJ (pferraro@jhu.edu)

Abstract

In many scientific disciplines, common research practices have led to unreliable and exaggerated evidence about scientific phenomena. Here, we describe some of these practices and quantify their pervasiveness in recent ecology publications in five popular journals. In an analysis of over three hundred and fifty studies published between 2018 and 2020, we detect empirical evidence of exaggeration bias and selective reporting of statistically significant results. This evidence implies that the published effect sizes in ecology journals exaggerate the importance of the ecological relationships that they aim to quantify. An exaggerated evidence base hinders the ability of empirical ecology to reliably contribute to science, policy, and management. To increase the credibility of ecology research, we describe a set of actions that ecologists should take, including changes to scientific norms about what high-quality ecology looks like and expectations about what high-quality studies can deliver.

Keywords

exaggeration bias, selective reporting, statistical power, multiple hypothesis testing, data archiving, code archiving, questionable research practices, open science

Credible evidence in ecology

Like all scientific disciplines, ecology advances, in part, through the generation of credible empirical evidence. Ecologists rely on this empirical evidence in their efforts to understand how the natural world works and to inform policy and management decisions. For example, models of climate change could drastically over- or under-predict how much carbon is sequestered by terrestrial plants without accurate estimates of effect sizes and the uncertainty about these estimates. Likewise, based on published studies, land managers may implement an intervention that promises to have relatively large effects, whereas the true effect is small or in the opposite direction.

Concerns about whether scientists have the correct incentives to generate credible empirical evidence have been raised in a wide range of scientific fields[1], including ecology[2–4]. These concerns revolve around common research practices and the professional incentives that encourage them. These practices, such as the selective reporting of results that are expected to impress reviewers and editors, undermine the credibility of empirical ecological science and have been connected to low rates of replicable findings in other fields[5–9]. A recent survey asked ecologists (N=494) and evolutionary biologists (N=313) to self-report their use of such "questionable research practices"[10]. Nearly two-thirds of respondents admitted to selective reporting at some point in their career and more than half admitted to reporting an unexpected finding as though it had been hypothesized prior to conducting the study (Hypothesizing After Results are Known or HARKing). These responses, however, do not necessarily demonstrate that these research practices are prevalent in recent ecology publications or that they have affected the empirical results reported in those publications.

Here, we report empirical analyses that indicate the prevalence of research practices that undermine the credibility of results in recent ecology publications. Our focus in these analyses is on widespread research practices that can impact the credibility and replicability of ecological science rather than on the precise meanings of "credibility" or "replicability" in ecology, which has been explored in other publications[11–13]. We hope that empirical evidence for these undesirable research practices in popular ecology journals may make ecologists take the problems they cause, and their solutions, more seriously.

We have three aims. First, we seek to provide a primer for new scientists and a refresher for experienced scientists on practices that lead to low credibility of published results. We focus on practices that can be empirically detected via analyses of published articles. Second, we quantify the extent to which these practices are prevalent in ecology publications. Specifically, (a) we assess, through the lens of statistical power, the degree to which ecologists use empirical designs that provide unreliable estimates of ecological relationships and the extent to which the magnitudes of published effect sizes are exaggerated, (b) we assess the degree to which ecologists selectively report statistically significant results (which can exacerbate the problem of exaggerated effect sizes), and (c) we assess the prevalence of multiple hypothesis testing without corrections for multiple comparisons (which can exacerbate selective reporting and exaggerated effect sizes). Our third and final aim is to summarize a set of solutions that authors, editors, reviewers, research institutions, and funders can adopt to prevent and mitigate the harms of practices that can undermine the credibility of ecological science.

To determine the extent to which these practices are prevalent in the ecology literature, we collected data from empirical studies published between January 2018 and May 2020 in five popular journals that publish general interest ecology studies and include many empirical

designs: *Ecology*, *Ecology Letters*, *Journal of Ecology*, *Nature*, and *Science*. We believe that these journals are representative of good quality ecological studies and thus we assume that the exclusion of other journals does not bias our conclusions. We included only empirical articles that reported statistically estimated parameters and errors in tables in the main or supplemental texts. Simulation, modeling, and meta-analysis articles were excluded. Because most statistical tests can be presented in table format and we have no reason to assume that certain tests or types of tests are more frequently reported in tables, we assume that including only estimates presented in tables does not bias our results. For every study, we then recorded: 1) every estimate and its associated error, 2) the sample size, 3) whether the study used multiple hypothesis testing, 4) whether there were corrections for multiple hypothesis testing, and 5) if data and code for analyses were available.

Overall, we collected data from 354 studies that reported 18,917 effect sizes and standard errors. For detailed methods, see Methods section. Our dataset and code are available at https://osf.io/9yd2b/.

Practices that lead to low credibility

Underpowered designs

The amount of information that ecologists can extract from their data depends on the variability of their data, the magnitude of the relationships they seek to estimate, and the precision with which they seek to estimate those relationships. When ecological data are highly variable and sample sizes are small relative to the true effect sizes, the estimated effect sizes are unreliable (i.e., the variability of the estimated effect sizes around the true effect will be large).

Given that most ecologists have training in frequentist statistics and engage in hypothesis testing, we explore the reliability of the estimated effects sizes in the ecology literature through the lens of statistical power. The statistical power of a test is the chance of detecting an effect, if such an effect exists[14]. Statistical power is based on the anticipated effect size, the sample size, the Type 1 error rate, and the sample variability. A conventional threshold for sufficient statistical power is 0.80, meaning that, if an effect of a given magnitude exists, a study design will detect it 80% of the time. Ecologists often seek to estimate the relationship between two variables and test whether the estimated value is different from a null hypothesis, which is usually that there is no relationship between the two variables. Consider, for example, a study that looks at how plant growth is related to phosphorus addition. A null hypothesis could be that phosphorus addition has no effect on plant growth. If a study is adequately powered, one would be likely to reject this null hypothesis if it were in fact false because the variability of the estimated effect sizes around the true effect size will be low. If, however, the study is underpowered, rejecting the null hypothesis would be unlikely because the variability of the estimated effect sizes around the true effect will be large. Thus, underpowered designs lead to greater prevalence of type II errors.

To estimate the statistical power of studies in our data, we followed the methods in [15]. First, we calculated an estimate for the magnitude of the true effect sizes that our collection of studies attempts to estimate. We estimated this effect as the weighted average of the partial correlation coefficients (PCCs) for all estimates in our study. A PCC is a measure of the strength and direction of the relationship between two variables when the influence of all other variables is held constant. Like a meta-analysis, this weighted average gives more weight to studies with more precise estimates. Our estimated "true effect" for our collection of studies was a PCC value of 0.06. Implicitly, we assume that there is no selective reporting or publication bias against

small effect size estimates in the literature (i.e., we assume ecologists report in the final publication everything that they estimated). Then we calculated the statistical power of the studies to detect this effect size (see Methods 'Power Analysis' for details). This approach does not imply that ecological effect sizes are homogenous across sites, studies, or variables in our 354 studies. Rather, the approach offers an approximation of the magnitude of the true effect size that a typical ecological study would expect to find.

Based on this approach, most tests in our collection of studies were underpowered at the conventional 0.80 threshold (Fig 1A). The median power for a test was 13.4%. Only 13.2% of all tests were powered at the 0.80 threshold or above. At a 0.60 threshold, 17.6% of all tests were adequately powered. Our results for a broad set of ecological studies are similar to those found in subfields of ecology[16–18] and in other disciplines[7,9,19]. To conclude the opposite – that the study designs are well powered – requires one to assume, among other assumptions, that ecologists have accurate expectations about the true effect sizes they seek to estimate in each study context and adjust their designs in a way that leads to less precise estimates when the true effect sizes are large (see Methods 'Power Analysis' for details). These expectations may exist, but in our collection of 354 published studies, only one mentioned performing power analyses, a finding that is similar to one reported in conservation biology where less than 10% of studies reported statistical power[20].

Whether our approach yields an accurate approximation of the statistical power of a typical ecology study also depends on another assumption. We assume that ecologists care about distinguishing small effect sizes from zero (e.g., PCC values less than our calculated weighted PCC of 0.06). Ecologists may, however, not be interested in small effect sizes. In fact, the sample sizes needed to distinguish these small effect sizes may be unattainable in single studies.

If the assumption that ecologists are interested in distinguishing from zero the typically small effect sizes reported in the literature is incorrect, we have under-estimated power in our analysis above.

Given that there is no single effect size that all ecological studies can expect or in which all ecologists would be interested, we also estimated power over a range of potential "true effect sizes." This range of PCC values includes the weighted mean of observational studies (0.05) in our sample, the unweighted median of effect sizes (0.15) in our sample, and the weighted mean of experimental studies (0.19) in our sample (see Supplemental Figure 1 for distribution of effect sizes in our data set). If we were to assume that the true effect ecologists in which ecologist are interested is large (PCC = 0.2), over half of all estimates are underpowered. For even larger effects (PCC = 0.3), over a quarter of estimates are underpowered (Fig 1B).

Exaggeration bias

The prevalence of underpowered study designs can lead to an exaggeration bias[9,21] in published studies when statistically significant results are preferred over non-significant results by editors, reviewers, and authors (i.e., publication bias [22]). Prior studies have reported evidence of publication biases in ecology[2–4,23], and these biases may be more severe in high impact journals like the ones we include in our study[24]. To illustrate how exaggeration bias arises, we consider again the example of a study that seeks to estimate the effect of phosphorus addition on plant growth. Assume that the true treatment effect is a 2% increase in aboveground biomass. In adequately powered studies, most estimated effects would be close to the 2% increase. In underpowered studies, however, the estimated values would vary widely around 2%, such that researchers are likely to report values that are much larger than the true value (type-M error) or

even opposite in sign (type-S error)[21]. Yet, in underpowered studies, only the values with exaggerated magnitudes are going to be statistically significant (i.e., with confidence intervals that exclude zero).

Prior research[21] reports that serious exaggeration problems arise when power is less than 50% (with power less than 10%, serious problems with estimates of the wrong sign also arise). If enough underpowered studies were published, researchers would be able to conduct a meta-analysis using the wide range of estimates to calculate a more accurate overall effect size[22,25]. However, where there is publication bias against results that do not pass conventional thresholds of statistical significance or have unexpected signs[9,19,26], mostly the large effect sizes with expected signs end up being published. Thus, the published effect sizes that scientists see are likely exaggerated in magnitude.

Following the methods of [7] and [15], we quantified the exaggeration bias of underpowered estimates by comparing reported effects to an average "true effect" of adequately powered estimates (see Methods 'Exaggeration Bias' for more details). As we did for the analysis of power, we also present the exaggeration bias results for a range of potential magnitudes of true effect sizes that ecologists may seek to estimate.

Our analysis implies that 63% of the estimates in underpowered studies are exaggerated over the true effect size by a factor of two or more (Fig 2A). Even if we assume a "true effect" of much greater magnitude, 1 in 4 estimates would still be exaggerated by a factor of 2 or more (Fig 2B). Our results are similar to a recent study of effect size exaggeration in three types of experimental ecological field studies. Using a different methodology, this study found that estimates were

198 exaggerated by anywhere from 0.66 times (drought experiments) to 3.29 times (warming
199 experiments) on average[17].

200 In a field where results often have real-world applications, magnitudes matter. In much of the
201 literature on "replication" and "reproducibility," the emphasis tends to be on identifying and
202 reducing false positives (e.g., [9,27]). In our view, a more important, but often overlooked, problem
203 lies in the potential for exaggeration bias in the magnitudes of reported effect sizes. This bias
204 results from a mix of the designs that researchers use and the incentives they face in trying to
205 publish their results (see next section on selective reporting).

206 Based on our empirical results, we are not asserting that most of the ecological relationships
207 reported in the literature are likely to be spurious – in fact, we doubt ecologists are studying
208 relationships for which the sharp null hypothesis of zero effect is widely true. Instead, we are
209 asserting that the magnitude of these relationships is inflated. In other words, we are asserting
210 that we have indirect empirical evidence - "fingerprints", if you will - that the published effect
211 sizes in ecology journals exaggerate the importance of many ecological relationships.

212 In our study, we use the concept of statistical power simply as a vehicle to illustrate the
213 inconvenient truth about ecological data: the outcome variables are noisy, the target effect sizes
214 are typically smaller than ecologists expect, and, given the designs ecologists are using and the
215 incentives they are facing, the estimated parameters in the literature are likely to be unreliable
216 and exaggerated. Our use of statistical power to explore the reliability of estimated effects in the
217 ecological literature is not an endorsement of Null Hypothesis Statistical Testing (NHST) or the
218 use of binary decision rules based on p-values to decide when an estimate is ecologically
219 relevant (e.g., $p<0.05$) [27–32].

Selective reporting of results

Because of publication biases in favor of statistically significant results[4,26,32], researchers may seek to find and publish such results over those that are statistically insignificant[33,34]. To obtain statistically significant results, researchers may choose methodologies or exclude data based on whether the choices yield statistically significant results. Researchers may also decide to stop collecting data based on when results are statistically significant[8,10]. Such choices are more likely when they can transform "marginally nonsignificant" results into statistically significant results (e.g., "*p*-hacking"). These choices may not be conscious and, when each is viewed in isolation, may be justifiable. Yet, the potential for these selective reporting practices to be widespread makes it difficult for readers to determine the credibility of a given analysis[35]. Selective reporting is found in most scientific disciplines[36]. Indeed, a recent survey of ecologist and evolutionary biologists reported that many researchers engaged, at least once in their careers, in selective reporting, such as not reporting response variables that did not reach a statistical significance threshold[10]. While some selective reporting practices may seem more malicious than others, all may exacerbate the reliability and exaggeration issues raised in the previous sections.

To explore the extent of selective reporting of statistically significant results in ecology, we followed the methods in [33]. We plotted the density of reported t-statistics and overlayed an Epanechnikov density kernel. We then weighted estimates by the number of estimates per table in each article (see Methods 'Selective Reporting' for more details). Without selective reporting, the density kernel should be a smooth function that declines as t-values increase. In contrast, a dip in the kernel density that creates a bimodal distribution with a second peak before the traditional 1.96 cut-off value for significance (i.e., $p = 0.05$) implies the presence of selective

reporting practices (not all selective reporting practices lead to a bimodal distribution[37], and thus its absence does not necessarily imply an absence of selective reporting practices).

When we focus on the results reported in the main article (as opposed to the supplemental material), the distribution of t-statistics has a bimodal distribution with fewer-than-expected t-statistics reported right before the traditional cut-off of 1.96 (Fig 3A). Yet when examining just the results presented in the supplemental text, we found no unusual distribution of t-statistics (Fig 3B). After combining all the results from the main text and supplemental materials, we again observe an unusual dip in the distribution of t-statistics (Fig 3C).

We hypothesize that this pattern of test statistics may arise from three sources. First, a researcher may pose a hypothesis that X influences Y and then use data on X and its covariates to test the hypothesis. The researcher may try multiple model specifications and statistical tests and then choose the combination that yields the most compelling results about the effect of X on Y to include in the main text, relegating the less compelling results to the supplemental material. Second, the same researcher may be unable to reject the null hypothesis that X has no effect on Y with any model or test. They then may search for other interesting and statistically significant effects in the data to report and revise the hypothesis they claim to be testing in the main text (HARKing). The researcher may still present all the tests that they conducted but place the nonsignificant results in the supplement instead of the main text. Third, rather than test a single hypothesis, ecology researchers often posit research questions in the form "what determines Y?" Such studies yield a range of estimated parameters, at least one estimate for each posited determinant of Y and maybe more if the researcher uses a variety of plausible models. The researcher may then selectively pick the "most interesting" estimates to report in the main text or, if they report all of the estimates, they may selectively pick the estimates from the "best"

model ("best" could be determined by statistical criteria but may also be determined by criteria that maximize the probability of publication, such as "how many statistically significant variables are obtained" or "what understudied variables deliver statistically significant results"). The perceived "less interesting" estimates or "inferior" models are relegated to the supplemental materials.

We cannot formally test these hypotheses with our data, but the responses from a recent survey of ecologists are consistent with our hypotheses[10]. Over 50% of the respondents self-reported that they did not report some variables in their analyses, did not report all the statistical tests they ran, or switched analysis strategies after seeing the results. Over one-third of ecologists admitted to collecting more data after checking to see if their initial results were statistically significant, or not reporting covariates if they failed to reach a significance threshold. Given that these responses are self-reported, they may underestimate the prevalence of these practices in ecology. They do, however, provide some evidence for why we see the bimodal distribution of t-statistics in Fig. 3A. The lack of this bimodal distribution in Fig. 3B, however, suggests that ecologists may be reporting their nonsignificant results, even if only in the supplemental materials. However, if authors are changing their hypotheses based on the results they report in the main text (i.e., HARKing), the presence of nonsignificant results in the supplemental materials provides little comfort about the credibility of the ecological evidence base (recall that over 50% of respondents in the survey by[10] self-reported HARKing in prior studies).

Multiple Hypothesis Testing

Opportunities for selective reporting grow when researchers engage in multiple hypothesis testing, where the same data are used to answer multiple research questions. The practice

includes testing the effects of one cause on multiple outcomes, testing the effect of multiple causes on one outcome, or testing heterogeneity of effects across sub-groups within the data. As more hypothesis tests are done on a given dataset, the likelihood of "false discoveries" increases simply because the error rate associated with a single hypothesis test does not account for a series (or family) of tests[38–40]. For example, a study that looks at the impact of phosphorus on total growth of the entire plant community along with growth of grass, legume, and forb species separately is testing multiple hypotheses.

In frequentist statistics, there are many procedures that allow researchers to present all of their hypothesis tests and to adjust their inferences when multiple hypotheses are tested e.g.,[39–41]; other procedures exist for the Bayesian context, e.g., [42]. However, application of these procedures is challenging because of debates about when the procedures are necessary and how best to execute them[43–45]. Further, adjusting inferences for multiple hypotheses comes with the trade-off of decreasing statistical power[46], which, as we showed above, is already low in ecology. Yet, without a full reporting of all tests that the authors performed and a justification for adjusting or not adjusting inferences based on that family of tests, the credibility of the results reported in ecology publications cannot be fully appreciated.

To shed light on the potential effects of multiple hypothesis testing on the ecological literature, we calculated the percentage of studies in our dataset that used multiple hypothesis testing and the percentage that used corrections for multiple hypothesis testing. Most studies in our dataset tested multiple hypotheses (85.0%), but very few used corrections (14% of those that tested multiple hypotheses; Fig 4). While correcting for multiple tests may not always be necessary (e.g., [41,43,47]), reporting why corrections were or were not used is necessary for readers to make judgements about the credibility of the analyses.

Together with selective reporting (for which we presented evidence in the previous section) and publication bias, multiple hypothesis testing may skew how researchers interpret the evidence base [48]. Researchers may be incentivized to report only "interesting" and statistically significant results instead of all the tests they performed on the dataset. Thus, we may not even know the extent to which multiple hypothesis testing occurs because some results may be simply excluded from publications.

Fostering a credibility culture in empirical ecology

Strengthening the reliability of ecological evidence will require changes in how ecologists produce and consume research. Ecologists must change their expectations about what high-quality ecological studies should look like and their expectations about what high-quality ecological studies can deliver. While these expectations can be shaped through better statistical knowledge[49,50], knowledge alone will be insufficient.

Changing expectations about what high-quality studies look like and can deliver will require changes in the incentives that ecologists face and in the norms that guide their empirical work. To encourage these changes across scientific fields, scholars have proposed a range of actions, including actions that individual researchers can take and actions that researchers must implement collectively[1]. A few publications describe some of these actions and some of the challenges to scaling these actions in the context of ecology[10,14,51–54]. We believe that most ecologists would readily adopt these actions but are not yet aware of them.

To help foster greater awareness, we highlight in Table 1 some promising actions that we believe will best contribute to improving the credibility and reliability of empirical ecology. Some of these actions, such as pre-registration and registered reports, are not well known in ecology.

More widely known is the importance of data and code availability for computational reproducibility[55,56] (a study is computationally reproducible if the same results can be achieved with the data and code used for the original analyses[12,13]). Best practices have been laid out for data and code archiving in ecology [57–61], and several journals (e.g., *Journal of Ecology*, Ecological Society of America publications (https://www.esa.org/publications/data-policy/)) and institutions (e.g., the NSF funded LTER network) require public data archiving. Yet, despite these attempts to make data and code more accessible (e.g., [62]), obtaining data and code can still be challenging[60,63–68]. For example, researchers were only able to obtain data from 19 of 74 articles in wildlife management. Using the data from these 19 publications, the researchers could reproduce the results in only 6 publications, even though code was provided for 9 studies[56]. Therefore, availability does not equate to quality of data or code[61]; most ecology and evolution publicly available datasets in a recent analysis were not reusable (a measure of ease with which data can be reused by third parties) and only slightly over half were complete[66]. In our data set of 354 studies, we found that most studies (78.5%) did make the data available, but only 18% of studies provided code for their analysis (and the code provided did not necessarily show the data cleaning steps; Fig 5). These percentages are similar to those reported using a sample of 346 articles from ecology journals that had mandatory or encouraged code sharing policies. In that study, 79% of studies provided data, 27% provided code, and 21% had both data and code[60].

Even with broader implementation of actions like pre-registration and the provision of both data and analysis code, many important decisions will remain in the hands of researchers and thus unobservable to outsiders. Thus, to fully address the issues raised in our article, we need a cultural shift, a shift where we assign more value to important questions and best practices and less value to exciting stories and statistically significant results[51,69]. Given the complexity of

ecological systems, we should not expect high-quality empirical studies to provide "airtight" conclusions or discontinuous jumps in our understanding of ecological processes. Instead, we should expect single studies to incrementally build on prior studies, to have substantial uncertainty arising from many sources (not just sampling variability), and to even present conflicting inferences, implying that we do not fully understand the underlying ecological processes.

One important step in the direction of a cultural shift is the recently created Society for Open, Reliable, and Transparent Ecology and Evolutionary biology (SORTEE: http://sortee.org). SORTEE aims to bring about cultural and institutional changes that can improve reliability and transparency in ecology, evolutionary biology, and related fields. The more the practices that SORTEE promotes are taught to new scientists, reinforced by senior researchers, and institutionalized by journals, funders and departments, the more reliable ecology research will be in the future.

We acknowledge that this cultural shift will not be swift because it requires structural changes in the incentives and norms in academia and other research settings. Yet, the continued scientific and policy relevance of ecology depends on our collective action to change these incentives and norms as soon as possible.

Methods

Data collection

Our methods follow those of [7]. We collected data from articles published between January 2018 and May 2020 in five popular journals for ecology publications. We collected data from every empirical article in three ecology journals (*Ecology*, *Ecology Letters*, and *Journal of Ecology*)

and every empirical ecology article in two general interest journals (*Nature* and *Science*) [n = 1,568 papers total]. Only empirical articles that statistically estimated parameters from data were included. These articles needed to have reported estimates and errors (standard errors or 95% confidence intervals) in tables either in the main text or supplemental materials. We focused on results reported in tables so that estimates and associated errors were easy to identify by the research team and to make sure that we were able to collect enough estimates for our analyses. Simulation or modeling articles were excluded. Meta-analyses were also excluded because we sought primary empirical data and did not want to double count any estimates that were found in both an original study and a meta-analysis.

Two people looked at every article to make sure that it fit our criteria. Dr. Kimmel initially pulled ecology subject papers from *Nature* and *Science* because these are for general audiences and publish on a wide range of topics. Papers were automatically excluded if they did not include tables. Those papers that did include tables were categorized into those that were empirical and those that were not.

We then recorded: 1) every estimate and its associated error, 2) the sample size, 3) whether the study used multiple hypothesis testing, 4) whether there were corrections for multiple hypothesis testing, and 5) if data and code for analyses in the study were available.

From the 1,568 papers in the five journals between our target years, we excluded 1,038 that did not report statistical tests in tables. We excluded 136 that were either meta-analyses or not empirical. 15 papers were removed that did not report errors and another 3 were removed that reported 0 for a standard error. One paper was removed because it was duplicated in 2019 and one was removed because the supplemental materials where tables may have been located did not open. 17 complete papers were removed because we could not discern sample sizes for any

of the tests. When checking our sampled data, one paper was removed because it should not have been classified as an ecology topic from *Science.* During data processing, we removed one publication that had over 6,000 estimates and one was removed when we discarded the top percentile of t-statistics. Thus, our final sample size was 354 publications.

When confidence intervals were reported instead of standard errors, we calculated upper confidence interval minus the estimate and lower confidence interval minus the estimate. We then recorded the smaller of the two if the interval was uneven. Thus, we are assuming less error about an estimate and potentially biasing our results towards a more favorable assessment of the literature than is warranted. These values were divided by 1.96 to obtain an equivalent standard error. Our use of 1.96 may not be correct for small sample sizes, but assuming that 1.96 is the benchmark will attribute less error about the point estimate. Thus, we will be overestimating the power of the tests. In other words, it makes our estimates of power more conservative.

When sample sizes were not directly reported in the tables, we inferred sample size from the methods. If we could not determine the sample size based on information given in the tables and methods, we made note that the sample size was unclear and dropped these papers from our analyses (n = 5,412 estimates from 29 publications).

To determine if a study used multiple hypothesis testing, we read the methods and looked at results presented in the main text of the manuscript. We categorized a study as using multiple hypothesis testing if the authors investigated multiple outcomes (dependent variables) associated with one cause (independent variable), investigated multiple causes (independent variables) associated with one outcome (dependent variables), or investigated sub-groups within their dataset. We were not concerned with one multiple regression being run (which could fall under multiple causes associated with one outcome), but instead several multiple regressions being run

on the same dataset. We tried not to include robustness checks as multiple hypothesis testing. We identified robustness checks by reading how the analysis was referenced and where possible reading figure or table captions. In most cases, robustness checks were easily identified – but the text was not always clear.

Further, to determine if there were corrections done, we did a keyword search for the following phrases: false discovery rate, family-wise error rate, Benjamini-Hochberg, Benjamini-Yekutieli, Bonferroni, Sidak, Dunn-Sidak, Holm, Hochberg, per-comparison error rate, and Dunnett's test.

We also categorized each study as experimental or observational and each results table as presenting "main" or "non-main" results, as in [7,33]. "Main" results were tables that were explicitly mentioned in the results text or figure legends. "Non-main" results were all other tables – usually those which were only reported in the methods or supplemental sections.

Software used

All data manipulation were done in R version 4.0.0 [70], and we utilized the 'here' package (version 1.0.1) for replicability [71]. Throughout our script, we used dplyr (version 1.0.7)[72] and tidyr (version 1.1.4) [73] to manipulate our data. We also relied on ggplot2 (version 3.3.5) [74], ggpubr (version 0.4.0) [75], patchwork (version 1.1.1) [76], and scales (version 1.1.1)[77] for making figures.

Data cleaning

Prior to the analyses, we cleaned and trimmed our data. First, we dropped 5,484 estimates from 34 studies where we could not determine the sample size for the analyses presented in tables. Then, we removed all estimates with a standard error of 0 (n = 810 estimates) and all coefficients that were not reported as integers (n = 7 estimates).

We "derounded" our estimates and standard errors, as in [33], to account for differences in how test statistics were rounded when reported. To deround, we picked a random value from the uniform distribution with the range of where n is the reported value and x is the number of decimal places the in the original value. For example, if the original estimate was 0.007, we picked a value from the range of [0.0065, 0.0075) using a random draw from the uniform distribution in this interval.

We then calculated t-stats based on the derounded estimates and their standard errors. The top percentile of the absolute value of the t-stats was then trimmed from the data (n = 257). This trimming ensures that a few data points do not disproportionately distort our estimate of power. We also excluded a study with more than 6,600 estimates (~26% of our total data before removed) so that our results would not be skewed by this one study. Our final sample size comprised 18,909 estimates from 353 unique publications.

Power analysis

To estimate the statistical power of studies in our data set and the extent of exaggeration bias, we followed the methods in [15]. Power calculations are conditional on some assumption of the size of the effect that the researchers are seeking to estimate. Here, we expressed power in the form of the minimum detectable effect (MDE). The MDE of a study design is the smallest effect that, if true, has an X% chance of producing an impact estimate that is statistically significant at the Y% level [78]. X is the level of statistical power (denoted as $(1-\beta)$ and commonly set to 80%) and Y is the Type I error rate (denoted as α and commonly set to 5%). The MDE can be written in terms of the standard error [79]:

$$MDE = \left(t_{1-\frac{\alpha}{2}} + t_{1-\beta}\right)\varepsilon \qquad (1)$$

where, $t_{1-\alpha/2}$ is the t-distribution with 1-α/2 degrees of freedom, $t_{1-\beta}$ is the t-distribution with 1-β degrees of freedom, and ε is the standard error of the estimated effect. Using conventional values of α=0.05 and β=20% for power of 80%) in (1) yields:

$$MDE = (1.96 + 0.84)\varepsilon \qquad (2)$$

$$= 2.8\varepsilon$$

Thus, when the standard error of an estimate is less than or equal to the MDE divided by 2.8, the test is adequately powered at the 80% threshold.

To calculate the MDE across our sample of studies, we must convert the estimates to a unitless measure with a common scale. This conversion allows us to compare estimates across studies. Here, we used the partial correlation coefficient (PCC), calculated as [80]:

$$PCC = \frac{t}{\sqrt{t \quad + df}} \qquad (3)$$

where t is the associated t-statistic of the estimate and df is the degrees of freedom. The standard error of the PCC was then estimated using [80]:

$$SEpcc = \frac{PCC}{t} = \frac{1}{\sqrt{t^2 + df}} \qquad (4)$$

Using the absolute values of PCC, we calculated the weighted average PCC for our entire dataset. The PCC values were weighted by the estimates' precision (e.g., the standard error about the estimate), so that estimates with higher precision (smaller standard errors) were assigned a larger weight. This weighted average PCC value served as our estimate of the true effect (the MDE in equation 2) that ecological studies are attempting to estimate. We then divided the weighted average of the PCC values by 2.8 to get the threshold to which we compared the *SEpcc*

values. When the *SEpcc* of an estimate was less than or equal to the threshold, the estimate had adequate power, otherwise it was under powered. We repeated these analyses for 75% and 60% power also where the weighted PCC was divided by 2.63 or 2.21 respectively to obtain the threshold values. See lines 110-142 in RepCode.R for how these analyses were done.

Most published studies did not provide the information required to calculate the degrees of freedom (df) for each model. To be consistent across studies, we approximate df using the sample size, N. Thus, we are often overestimating the *df* of a model, even more so when the estimates come from a mixed effects model (42% of the estimates in our dataset are from some sort of mixed effects model). Therefore, most of our calculated PCC values are smaller than they would be if we used *df*. Because we are using N, we are also likely underestimating SE of the PCC values (which are smaller to a greater degree than the PCC values are smaller). This will reduce our SE of the PCC values which we compare to the MDE threshold. Thus, overall, we are likely overestimating the power of most tests in our sample of studies.

We recognize that each empirical study in ecology seeks to estimate a different effect, whose true value may vary across studies. Given that the true effect size is not known, we also explored how our conclusions changed with changes in the assumed true effect size (Fig. 1B). For a range of "true effect" values, we computed how many PCC estimates had a standard error greater than the threshold value based on hypothetical true effect sizes divided by 2.8. Our range went up to PCC values of 0.20 (in terms of standard deviations of the outcome variable, this effect size would be analogous to an effect size of roughly 0.5 SD). Our estimated weighted PCC value (our MDE in equation 2) from our entire dataset was 0.06. For only observational studies in our dataset, it was 0.05. For only experimental studies in our dataset, it was 0.19. These values make sense if we assume that experimental studies tend to push the system further than observational

studies and, consequently, have larger effects to report. Further, this range spans most of the PCC values recorded from our dataset (Supplemental Figure 1) and covers the unweighted median PCC value of our sample. Thus, the values we present in Fig. 1B represent a reasonable range of PCC values that we may expect in ecological studies.

Because several reviewers of our original manuscript raised concerns about using a single effect size to estimate power, we wanted to present the assumptions about the data generating process to come to an opposite conclusion; i.e., to conclude that the study designs are, in fact, well powered or, more generally, able to easily isolate signal from noise.

Step 1: First, recall how we concluded that the typical true effect size in ecological studies is small in magnitude. In our data, the smaller the estimated effect, the more precise the estimate. Thus, our meta-regression estimator, which weights the estimates by their precision, yields a relatively small effect size, which we claim serves as a benchmark for thinking about the typical true effect size in ecological studies.

Step 2: Let us consider how the conclusion from Step 1 could be wrong (i.e., our conclusion that true effect sizes tend to be small and thus most ecological studies are underpowered to detect the true effect sizes). One would have to make two assumptions: (i) ecologists, before designing their studies, think about the true effect sizes they are targeting and the underlying sampling variability, and they are roughly accurate in their expectations; and (ii) ecologists who target larger true effect sizes choose designs with relatively smaller sample sizes or contexts in which the variance in the outcome measure is relatively higher (i.e., ecologists who seek to estimate larger effect sizes do not maintain the same relative level of precision as those who seek to estimate smaller effect sizes). In other words, ecologists are adjusting their designs to match the true heterogeneous effect sizes that they target and are adjusting their designs in a way that

reduces the relative precision of the estimate as the true effect size increases in magnitude. If those two conditions hold, then our conclusions about unreliable estimates could be wrong.

Step 3: Let us consider more deeply the two assumptions required to come to the opposite conclusion from the one described in our manuscript. Assumption (i) would require that ecologists think very carefully about the noise in their data and the magnitude of the target effect size prior to collecting data. Although we acknowledge that statistical power calculations or simulations are not the only way to think about such design attributes, they are likely to be one of the most popular ways of doing so among ecologists. Yet if ecologists conduct power analyses with regularity, they do not report them in their publications: only one study of the 353 publications in our dataset reported conducting a power analysis.

Even if ecologists do carefully think about the noise in their data and the magnitude of the target effect size prior to collecting data, assumption (ii) would require one of two additional conditions. First, when the expected treatment effect sizes are large, the costs of data collection or selecting study units are also large. This pattern of costs could imply that, in comparison to ecologists seeking to estimate small true effect sizes, ecologists seeking to estimate large effects cannot as easily reduce the influence of noise by increasing sample size or by selecting a subset of the target population that has lower outcome variance. If this first condition about differences in relative costs were not satisfied, an alternative condition could support assumption (ii). In comparison to ecologists who work on studies seeking to estimate small effect sizes, ecologists who seek to estimate large true effect sizes must be more cognizant that peer reviewers and editors are unlikely to care about the precision of their estimates as long as the confidence interval doesn't cross the null hypothesis value.

Lastly, we computed the median power for our sample of tests as in [81]. The median power is calculated as one minus the cumulative normal probability of the difference between 1.96 and the absolute value of the weighted average PCC estimate divided by the median standard error. We calculated this value for six sets of the data: the entire dataset, the set of "main" estimates, the set of estimates in the main text, the set of estimates in the supplemental text, the set of estimates from observational studies, and set of the estimates from experimental studies (see RepCode.R lines 304-333 for these calculations).

Exaggeration Bias

We calculated the exaggeration bias as in [7,15]. First, we calculate the weighted average of PCC values for the subset of tests that are adequately powered. We refer to this value as the weighted average of the adequately powered estimators (WAAP). The WAAP that we calculated for our dataset was 0.05. According to Ioannidis (2017), the WAAP is a conservative benchmark for the "true" effect. To calculate how exaggerated estimates from underpowered designs were, we calculated the ratio between the absolute value of the PCC for each estimate and the WAAP. If this ratio was less than 1, estimates were deflated (e.g., smaller than expected). If this ratio was greater than 1, estimates were inflated. Specifically, we categorized estimates that were inflated by 0-100% (ratio greater than or equal to 1, but less than 2), by 100-300% (ratio greater than or equal to 2, but less than 4), and by 300% or more (ratio greater than or equal to 4).

Again, because we acknowledge that the WAAP estimate may be different for different types of studies, we then explore how our conclusions may change given different WAAP values (Fig. 2B). For a range of WAAP values from 0.01 to 0.2, we calculated how many estimates would be inflated by 100% or more. To do this, we compared the WAAP values in this range to the absolute value of the PCC values for underpowered estimates. Any PCC value divided by the

WAAP that was greater than 2 was considered inflated by 100% or more. See RepCode.R lines 335-417 for these calculations and creation of figures.

Selective Reporting

To explore the extent of selective reporting of statistically significant results, we followed the methods in [33]. We plotted the density of t-statistics and overlayed an Epanechnikov density kernel. Estimates were weighted by the number of estimates per table in each article. Without selective reporting, the density kernel should be a smooth function declining at higher t-values. A dip that creates a bimodal distribution with a second peak near the 1.96 cut-off for significance (i.e., $p = 0.05$) suggests selective reporting.

Multiple hypothesis testing, data & code availability

We calculated the percentage of studies in our dataset that used multiple hypothesis testing and the percentage that used corrections for multiple hypothesis testing (see definitions in Data Collection section above). To quantify the extent to which the data and analysis code from our studies are available for replication, we calculated the percentage of studies that made the data or analysis code, or both, available.

Data Availability

Our dataset is available at https://osf.io/9yd2b/.

Code Availability

Our analysis code is available at https://osf.io/9yd2b/.

Acknowledgments

We thank the Glenadore and Howard L. Pim Postdoctoral Fellowship in Global Change for funding KK. We thank Tim Parker for his helpful comments on revising the manuscript. We thank Morgan Buchanan, Patrick Dye, Zachary Ellis, Yuchen Li, Lydia Wang, and Lauren Williams for helping in the data collection for this paper. We thank Pallavi Shukla for providing sample code for the analyses.

Author Contributions

PJF and KK designed the study. KK analyzed the data. MLA, PJF, and KK wrote the paper.

Competing interests

The authors declare no competing interests.

608 **Table 1. Changes in research practices to help increase the reliability of ecological research.***

Recommendation	Details	Purpose	References
Checklists	Used at multiple stages of the publication process: for example, they can be used before submitting, during review, and by editors	- ensure researchers include necessary information for evaluating the study - highlight key features of study design for reviewers - educate authors and reviewers on best practices	Simmons *et al.* 2011; Nosek *et al.* 2012; Parker *et al.* 2016, 2018
Data and Code Archiving	Publicly available except where data privacy is necessary.	- increase the transparency of study workflows and conclusions - facilitate computational reproducibility and evidence synthesis	Nakagawa & Parker 2015; Nosek *et al.* 2015; Parker *et al.* 2016; Munafò *et al.* 2017; Culina *et al.* 2020
Pre-registration & pre-analysis plans	Pre-analysis plans: describe the research questions, the design, and the methods that will be used in a study; completed before data analysis begins (ideally, before all data have been collected). Pre-registration: process of registering, before the study or data analysis begins, a researcher's intent to undertake a study and the study's pre-analysis plan.	- help authors to be transparent in their research decisions - reduce, or at least make more transparent, the practices of HARKing, selective reporting of results, and presentations of presentations of exploratory analyses as if they were confirmatory analyses planned from the outset - help scholars quantify the "file drawer" problem: studies that were completed, but never published	Kaplan & Irvin 2015; Forstmeier *et al* 2017; Nosek *et al.* 2018; Parker *et al.* 2019
Registered Reports	Two-stage peer review. Prior to data collection and analysis, authors submit study motivation, design, and methods. Reviewers judge submission based on quality of question and design. Second stage reviews assess how closely study follows original plan.	- reduce selective reporting of results - reviewers focus on importance of question and quality of design, not the sign, magnitude, and statistical significance of results	https://www.cos.io/initiatives/registered-reports Button *et al.* 2016; Allen & Mehler 2019; Nosek *et al.* 2019; Scheel *et al.* 2021; Soderberg *et al.* 2021
Results-Blind Reviews	Full manuscript submitted for review, but results are not included.	- reviewers focus on importance of question and quality of design, not the sign, magnitude, and statistical significance of results - no mechanism to reduce selective reporting because no pre-analysis plan is required	Smulders 2013; Button *et al.* 2016
Incentives	Institutions that matter - namely, employers, funders, and publishers –move away from incentivizing "exciting" results and towards incentivizing best practices.	- align personal values of many researchers to create and disseminate credible science - value replication studies along with "ground-breaking" research	Anderson *et al.* 2007; Nosek *et al.* 2012; O'Dea *et al.* 2021 http://sortee.org https://sfdora.org/

609

610 *** See SM text 'Promising actions …' for more details on practices.**

Figure Legends/Captions

Fig. 1. Percentage of statistical tests that meet and do not meet the conventional 0.8 threshold for statistical power. In A, we show a histogram of the standard error of the partial correlation coefficients (PCC) from ecological studies. All estimates to the right of the red line are under-powered at an 80% power threshold. n = 18,917 estimates from 354 studies. In B, we show what percentage of the 18,917 estimates would be underpowered for a range of PCC values.

Fig. 2. The percentage of underpowered estimates from ecological studies that are exaggerated. In A, we show the percentage of estimates from underpowered studies that are exaggerated based on the weighted averages of adequately powered estimates in our sample of studies. Deflation refers to any estimate that is smaller than the hypothesized true effect, while the other categories represent exaggeration. n = 16,407 estimates from 330 studies with underpowered estimates. In B, we show what percentage of the 16,407 estimates would be exaggerated by 100% or more given a range of WAAP values.

Fig. 3. Evidence of selective reporting of statistically significant results in (A) main text tables (n = 2,286 estimates), (B) supplemental text tables (n = 14,680 estimates), and (C) all tables (n = 16,966) in ecology publications. The solid black line is a fitted density kernel of the distribution of t-statistics. Each gray bar represents the density of studies with that t-statistic value. We present t-statistics up to 10 although there are higher values in our data. The red arrow points to the value at the conventional threshold of statistical significance ($p<0.05$). At that point, we would expect a smoothly decreasing line in the absence of selective reporting.

Fig. 4. The percentage of ecology studies that use multiple hypothesis testing. The gray section represents the percentage of studies that used multiple hypothesis corrections. n = 354 studies.

Fig. 5. The percentage of ecology studies which (A) have data available, and (B) provide code for their analyses. Bars are colored by journal. E – *Ecology*, EL – *Ecology Letters*, JOE – *Journal of Ecology*, N – *Nature*, S – *Science*. n = 354 studies.

References

1. Nosek, B. A., Spies, J. R. & Motyl, M. Scientific Utopia: II. Restructuring Incentives and Practices to Promote Truth Over Publishability. *Perspect. Psychol. Sci.* **7**, 615–631 (2012).
2. Leimu, R. & Koricheva, J. Cumulative meta-analysis: A new tool for detection of temporal trends and publication bias in ecology. *Proc. R. Soc. B Biol. Sci.* **271**, 1961–1966 (2004).
3. Møller, A. P. & Jennions, M. D. Testing and adjusting for publication bias. *Trends Ecol. Evol.* **16**, 580–586 (2001).
4. Barto, E. K. & Rillig, M. C. Dissemination biases in ecology: Effect sizes matter more than quality. *Oikos* **121**, 228–235 (2012).
5. Christensen, G. & Miguel, E. Transparency, reproducibility, and the credibility of economics Research. *J. Econ. Lit.* **56**, 920–980 (2018).
6. Collaboration, O. S. Estimating the reproducibility of psychological science. *Science* **349**, aac4716 (2015).
7. Ferraro, P. J. & Shukla, P. Is a Replicability Crisis on the Horizon for Environmental and Resource Economics? *Rev. Environ. Econ. Policy* **14**, 339–351 (2020).
8. Martinson, B. C., Anderson, M. S. & de Vries. Scientists behaving badly. *Nature* **435**, 737–738 (2005).
9. Ioannidis, J. P. A. Why most published research findings are false. *PLoS Med.* **2**, 696–701 (2005).
10. Fraser, H., Parker, T., Nakagawa, S., Barnett, A. & Fidler, F. Questionable research practices in ecology and evolution. *PLoS One* **13**, (2018).

11. Fraser, H., Barnett, A., Parker, T. H. & Fidler, F. The role of replication studies in ecology. *Ecol. Evol.* **10**, 5197–5207 (2020).

12. Fidler, F. *et al.* Metaresearch for evaluating reproducibility in ecology and evolution. *Bioscience* **67**, 282–289 (2017).

13. Cassey, P. & Blackburn, T. M. Reproducibility and repeatability in Ecology. *Bioscience* **56**, 958–959 (2006).

14. Parker, T. H. *et al.* Transparency in Ecology and Evolution: Real Problems, Real Solutions. *Trends Ecol. Evol.* **31**, 711–719 (2016).

15. Ioannidis, J. P. A., Stanley, T. D. & Doucouliagos, H. The Power of Bias in Economics Research. *Econ. J.* **127**, F236–F265 (2017).

16. Jennions, M. D. & Møller, A. P. A survey of the statistical power of research in behavioral ecology and animal behavior. *Behav. Ecol.* **14**, 438–445 (2003).

17. Lemoine, N. P. *et al.* Underappreciated problems of low replication in ecological field studies. *Ecology* **97**, 2562–2569 (2016).

18. Yang, Y. *et al.* Publication bias impacts on effect size, statistical power, and magnitude (Type M) and sign (Type S) errors in ecology and evolutionary biology. *BMC Bio.* (2022).

19. Button, K. S. *et al.* Power failure: Why small sample size undermines the reliability of neuroscience. *Nat. Rev. Neurosci.* **14**, 365–376 (2013).

20. Fidler, F., Burgman, M. A., Cumming, G., Buttrose, R. & Thomason, N. Impact of criticism of null-hypothesis significance testing on statistical reporting practices in conservation biology. *Conserv. Biol.* **20**, 1539–1544 (2006).

21. Gelman, A. & Carlin, J. Beyond Power Calculations: Assessing Type S (Sign) and Type M (Magnitude) Errors. *Perspect. Psychol. Sci.* **9**, 641–651 (2014).

22. Nichols, J. D., Oli, M. K., Kendall, W. L. & Scott Boomer, G. A better approach for dealing with reproducibility and replicability in science. *Proc. Natl. Acad. Sci. U. S. A.* **118**, 1–5 (2021).

23. Koricheva, J. Non-significant results in ecology: A burden or a blessing in disguise? *Oikos* **102**, 397–401 (2003).

24. Ceausu, I. *et al.* High impact journals in ecology cover proportionally more statistically significant findings. *bioRxiv* (2018) doi:10.1093/sw/38.6.771.

25. Nichols, J. D., Kendall, W. L. & Boomer, G. S. Accumulating evidence in ecology: Once is not enough. *Ecol. Evol.* **9**, 13991–14004 (2019).

26. Fanelli, D. Negative results are disappearing from most disciplines and countries. *Scientometrics* **90**, 891–904 (2012).

27. Fanelli, D. Is science really facing a reproducibility crisis, and do we need it to? *Proc. Natl. Acad. Sci. U. S. A.* **115**, 2628–2631 (2018).

28. Yoccoz, N. G. Use, Overuse, and Misuse of Significance Tests in Evolutionary Biology and Ecology. *Bull. Ecol. Soc. Am.* **72**, 106–111 (1991).

29. Fidler, F., Fraser, H., McCarthy, M. A. & Game, E. T. Improving the transparency of statistical reporting in Conservation Letters. *Conserv. Lett.* **11**, 1–5 (2018).

30. Murtaugh, P. A. In defense of P values. *Ecology* **95**, 611–617 (2014).

31. Anderson, D. R., Burnham, K. P. & Thompson, W. L. Null hypothesis testing: Problems, prevalence, and an alternative. *J. Wildl. Manage.* **64**, 912–923 (2000).

32. Callaham, M., Wears, R. L. & Weber, E. Journal prestige, publication bias, and other characteristics associated with citation of published studies in peer-reviewed journals. *J. Am. Med. Assoc.* **287**, 2847–2850 (2002).

33. Brodeur, A., Lé, M., Sangnier, M. & Zylberberg, Y. Star Wars: The empirics strike back. *Am. Econ. J. Appl. Econ.* **8**, 1–32 (2016).
34. Gopalakrishna, G. *et al.* Prevalence of questionable research practices, research misconduct and their potential explanatory factors: A survey among academic researchers in the Netherlands. *PLoS One* **17**, 1–16 (2022).
35. Simmons, J. P., Nelson, L. D. & Simonsohn, U. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychol. Sci.* **22**, 1359–1366 (2011).
36. Head, M. L., Holman, L., Lanfear, R., Kahn, A. T. & Jennions, M. D. The Extent and Consequences of P-Hacking in Science. *PLoS Biol.* **13**, 1–15 (2015).
37. Hartgerink, C. H. J., Van Aert, R. C. M., Nuijten, M. B., Wicherts, J. M. & Van Assen, M. A. L. M. Distributions of p-values smaller than .05 in psychology: What is going on? *PeerJ* **2016**, (2016).
38. Shaffer, J. P. Multiple hypothesis testing. *Annu. Rev. Psychol.* **46**, 561–84 (1995).
39. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B* **57**, 289–300 (1995).
40. Dunnett, C. W. A Multiple Comparison Procedure for Comparing Several Treatments with a Control. *J. Am. Stat. Assoc.* **50**, 1096–1121 (1955).
41. Yekutieli, D. & Benjamini, Y. Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics. *J. Stat. Plan. Inference* **82**, 171–196 (1999).
42. Berry, D. A. & Hochberg, Y. Bayesian perspectives on multiple comparisons. *J. Stat. Plan. Inference* **82**, 215–227 (1999).
43. Gelman, A., Hill, J. & Yajima, M. Why We (Usually) Don't Have to Worry About

Multiple Comparisons. *J. Res. Educ. Eff.* **5**, 189–211 (2012).

44. Rubin, M. Do p values lose their meaning in exploratory analyses? It depends how you define the familywise error rate. *Rev. Gen. Psychol.* **21**, 269–275 (2017).

45. Rubin, M. When does HARKing hurt? Identifying when different types of undisclosed post hoc hypothesizing harm scientific progress. *Rev. Gen. Psychol.* **21**, 308–320 (2017).

46. Nakagawa, S. A farewell to Bonferroni: The problems of low statistical power and publication bias. *Behav. Ecol.* **15**, 1044–1045 (2004).

47. Berry, D. A. & Hochberg, Y. Bayesian perspectives on multiple comparisons. *J. Stat. Plan. Inference* **82**, 215–227 (1999).

48. Forstmeier, W., Wagenmakers, E. J. & Parker, T. H. Detecting and avoiding likely false-positive findings – a practical guide. *Biol. Rev.* **92**, 1941–1968 (2017).

49. Baker, M. & Penny, D. Is there a reproducibility crisis? *Nature* **533**, 452–454 (2016).

50. Gelman, A. & Loken, E. The statistical crisis in science. *Am. Sci.* **102**, 460–465 (2014).

51. O'Dea, R. E. *et al.* Towards open, reliable, and transparent ecology and evolutionary biology. *BMC Biol.* **19**, 1–5 (2021).

52. Parker, T. H., Nakagawa, S. & Gurevitch, J. Promoting transparency in evolutionary biology and ecology. *Ecol. Lett.* **19**, 726–728 (2016).

53. Parker, T., Fraser, H. & Nakagawa, S. Making conservation science more reliable with preregistration and registered reports. *Conserv. Biol.* **33**, 747–750 (2019).

54. Buxton, R. T. *et al.* Avoiding wasted research resources in conservation science. *Conserv. Sci. Pract.* **3**, 1–11 (2021).

55. Powers, S. M. & Hampton, S. E. Open science, reproducibility, and transparency in ecology. *Ecol. Appl.* **29**, 1–8 (2019).

56. Archmiller, A. A. *et al.* Computational Reproducibility in The Wildlife Society's Flagship Journals. *J. Wildl. Manage.* **84**, 1012–1017 (2020).

57. Whitlock, M. C., McPeek, M. A., Rausher, M. D., Rieseberg, L. & Moore, A. J. Data archiving. *Am. Nat.* **175**, 145–146 (2010).

58. Whitlock, M. C. Data archiving in ecology and evolution: Best practices. *Trends Ecol. Evol.* **26**, 61–65 (2011).

59. Mislan, K. A. S., Heer, J. M. & White, E. P. Elevating The Status of Code in Ecology. *Trends Ecol. Evol.* **31**, 4–7 (2016).

60. Culina, A., van den Berg, I., Evans, S. & Sánchez-Tójar, A. Low availability of code in ecology: A call for urgent action. *PLoS Biol.* **18**, 1–9 (2020).

61. Wilkinson, M. D. *et al.* Comment: The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* **3**, 1–9 (2016).

62. Gopalakrishna, G. *et al.* Prevalence of responsible research practices among academics in The Netherlands. *F1000Research* **11**, 1–34 (2022).

63. Hardwicke, T. E. *et al.* Data availability, reusability, and analytic reproducibility: Evaluating the impact of a mandatory open data policy at the journal Cognition. *R. Soc. Open Sci.* **5**, (2018).

64. Stodden, V., Seiler, J. & Ma, Z. An empirical analysis of journal policy effectiveness for computational reproducibility. *Proc. Natl. Acad. Sci. U. S. A.* **115**, 2584–2589 (2018).

65. Roche, D. G., Kruuk, L. E. B., Lanfear, R. & Binning, S. A. Public Data Archiving in Ecology and Evolution: How Well Are We Doing? *PLoS Biol.* **13**, 1–12 (2015).

66. Roche, D. G. *et al.* Slow improvement to the archiving quality of open datasets shared by researchers in ecology and evolution. *Proc. R. Soc. B Biol. Sci.* **289**, (2022).

67. Lindsey, P. A. *et al.* The bushmeat trade in African savannas: Impacts, drivers, and possible solutions. *Biol. Conserv.* **160**, 80–96 (2013).

68. Roche, D. G. *et al.* Paths towards greater consensus building in experimental biology. *J. Exp. Biol.* **225**, (2022).

69. Smaldino, P. E. & McElreath, R. The natural selection of bad science. *R. Soc. Open Sci.* **3**, (2016).

70. R Core Team. R: A language and environment for statistical computing. (2019).

71. Müller, K. here: A Simpler Way to Find Your Files. R. (2017).

72. Wickham, H., Francois, R., Henry, L. & Muller, K. dplyr: A Grammar of Data Manipulation. (2020).

73. Wickham, H. & Henry, L. tidyr: Tidy Messy Data. (2020).

74. Wickham, H. ggplot2: Elegant Graphics for Data Analysis. (2016).

75. Kassambara, A. ggpubr: 'ggplot2' Based Publication Ready Plots. (2020).

76. Pedersen, T. L. patchwork: The Composer of Plots. (2021).

77. Wickham, H. & Seidel, D. scales: Scale Functions for Visualization. (2020).

78. Bloom, H. S. Minimum Detectable Effects: A Simple Way to Report the Statistical Power of Experimental Designs. *Eval. Rev.* **19**, 547–556 (1995).

79. Djimeu, E. W. & Houndolo, D. G. Power calculation for causal inference in social science: sample size and minimum detectable effect determination. *J. Dev. Eff.* **8**, 508–527 (2016).

80. Havranek, T., Horvath, R. & Zeynalov, A. Natural Resources and Economic Growth: A Meta-Analysis. *World Dev.* **88**, 134–151 (2016).

81. Stanley, T.D., Carter, E. C., & Doucouliagos, H. What Meta-Analyses Reveal about the

Replicability of Psychological Research. *Psychol. Bull.* **144**, 1325–1346 (2018).

82. Parker, T. H. *et al.* Empowering peer reviewers with a checklist to improve transparency. *Nat. Ecol. Evol.* **2**, 929–935 (2018).

83. Munafò, M. R. *et al.* A manifesto for reproducible science. *Nat. Hum. Behav.* **1**, 1–9 (2017).

84. Nosek, B. A. *et al.* Promoting an open research culture. *Science (80-.).* **348**, 1422–1425 (2015).

85. Nakagawa, S. & Parker, T. H. Replicating research in ecology and evolution: Feasibility, incentives, and the cost-benefit conundrum. *BMC Biol.* **13**, 1–6 (2015).

86. Kaplan, R. M. & Irvin, V. L. Likelihood of null effects of large NHLBI clinical trials has increased over time. *PLoS One* **10**, 1–12 (2015).

87. Nosek, B. A., Ebersole, C. R., DeHaven, A. C. & Mellor, D. T. The preregistration revolution. *Proc. Natl. Acad. Sci. U. S. A.* **115**, 2600–2606 (2018).

88. Allen, C. & Mehler, D. M. A. Open science challenges, benefits and tips in early career and beyond. *PLoS Biol.* **17**, 1–14 (2019).

89. Scheel, A. M., Schijen, M. R. M. J. & Lakens, D. An Excess of Positive Results: Comparing the Standard Psychology Literature With Registered Reports. *Adv. Methods Pract. Psychol. Sci.* **4**, (2021).

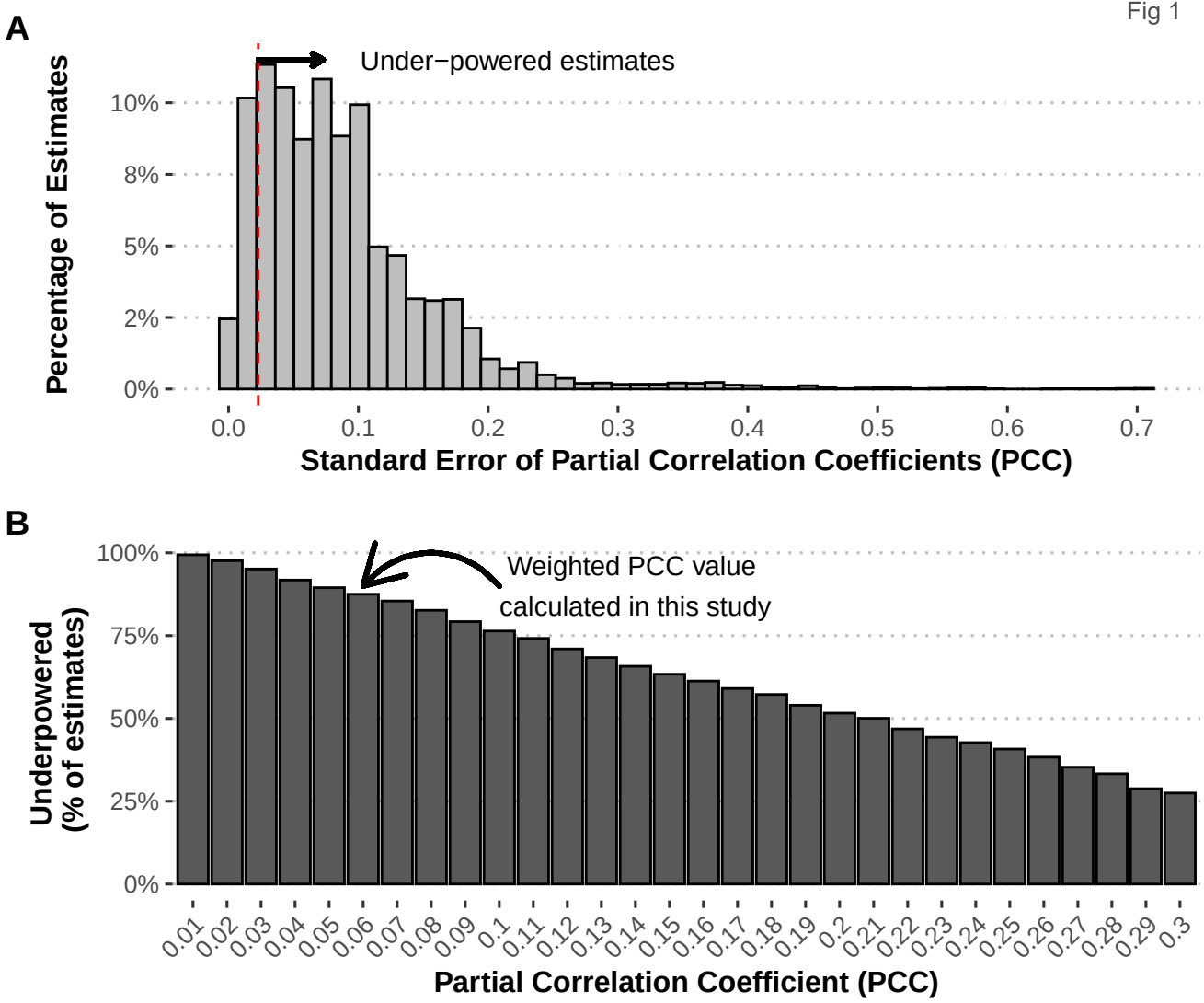
90. Nosek, B. A. *et al.* Preregistration Is Hard, And Worthwhile. *Trends Cogn. Sci.* **23**, 815–818 (2019).

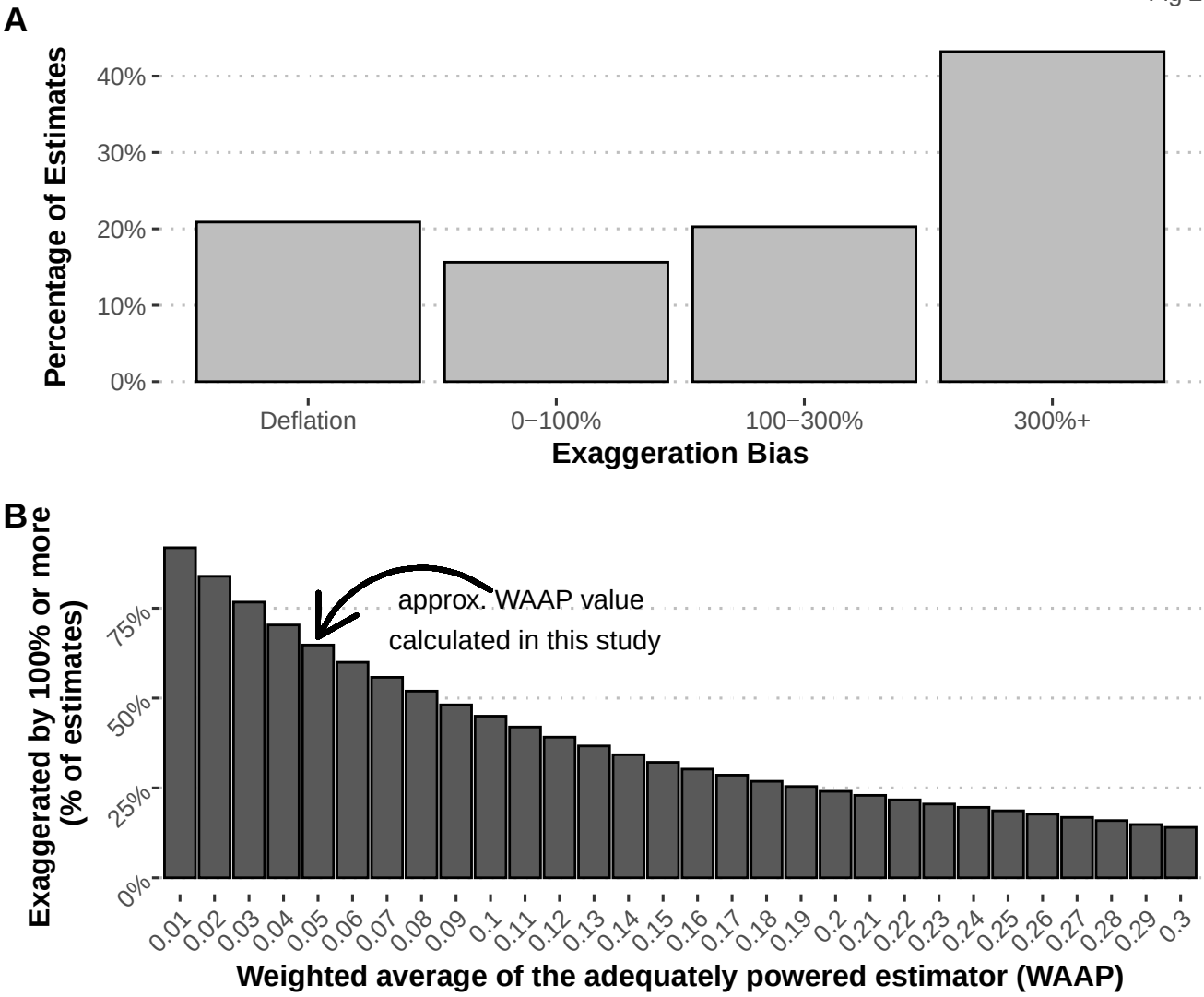
91. Button, K. S., Bal, L., Clark, A. & Shipley, T. Preventing the ends from justifying the means: Withholding results to address publication bias in peer-review. *BMC Psychol.* **4**, 1–7 (2016).

92. Soderberg, C. K. *et al.* Initial evidence of research quality of registered reports compared with the standard publishing model. *Nat. Hum. Behav.* **5**, 990–997 (2021).

93. Smulders, Y. M. A two-step manuscript submission process can reduce publication bias. *J. Clin. Epidemiol.* **66**, 946–947 (2013).

94. Anderson, M. S., Martinson, B. C. & De Vries, R. Normative dissonance in science: Results from a national survey of U.S. scientists. *J. Empir. Res. Hum. Res. Ethics* 3–14 (2007).


A
Under−powered estimates
Percentage of Estimates
10%
8%
5%
2%
0%
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
Standard Error of Partial Correlation Coefficients (PCC)
B
Weighted PCC value
calculated in this study
Underpowered
(% of estimates)
100%
75%
50%
25%
0%
0.01
0.02
0.03
0.04
0.05
0.06
0.07
0.08
0.09
0.1
0.11
0.12
0.13
0.14
0.15
0.16
0.17
0.18
0.19
0.2
0.21
0.22
0.23
0.24
0.25
0.26
0.27
0.28
0.29
0.3
Partial Correlation Coefficient (PCC)


A
Percentage of Estimates
40%
30%
20%
10%
0%
Deflation
0−100%
100−300%
300%+
Exaggeration Bias
B
Exaggerated by 100% or more
(% of estimates)
75%
50%
25%
0%
approx. WAAP value
calculated in this study
0.01
0.02
0.03
0.04
0.05
0.06
0.07
0.08
0.09
0.1
0.11
0.12
0.13
0.14
0.15
0.16
0.17
0.18
0.19
0.2
0.21
0.22
0.23
0.24
0.25
0.26
0.27
0.28
0.29
0.3
Weighted average of the adequately powered estimator (WAAP)

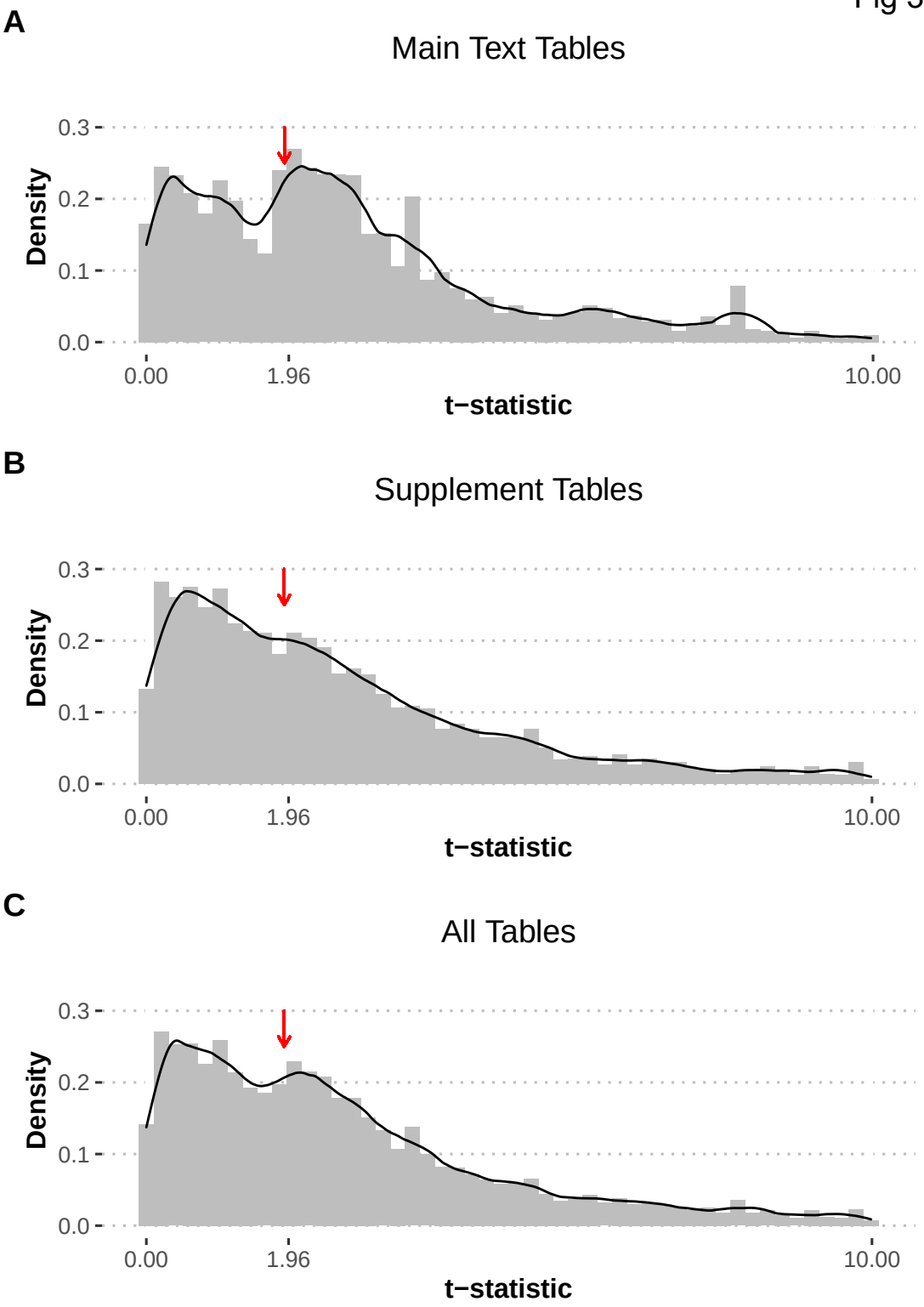

A
Main Text Tables
Density
0.3
0.2
0.1
0.0
0.00
1.96
10.00
t−statistic
B
Supplement Tables
Density
0.3
0.2
0.1
0.0
0.00
1.96
10.00
t−statistic
C
All Tables
Density
0.3
0.2
0.1
0.0
0.00
1.96
10.00
t−statistic

Fig 4

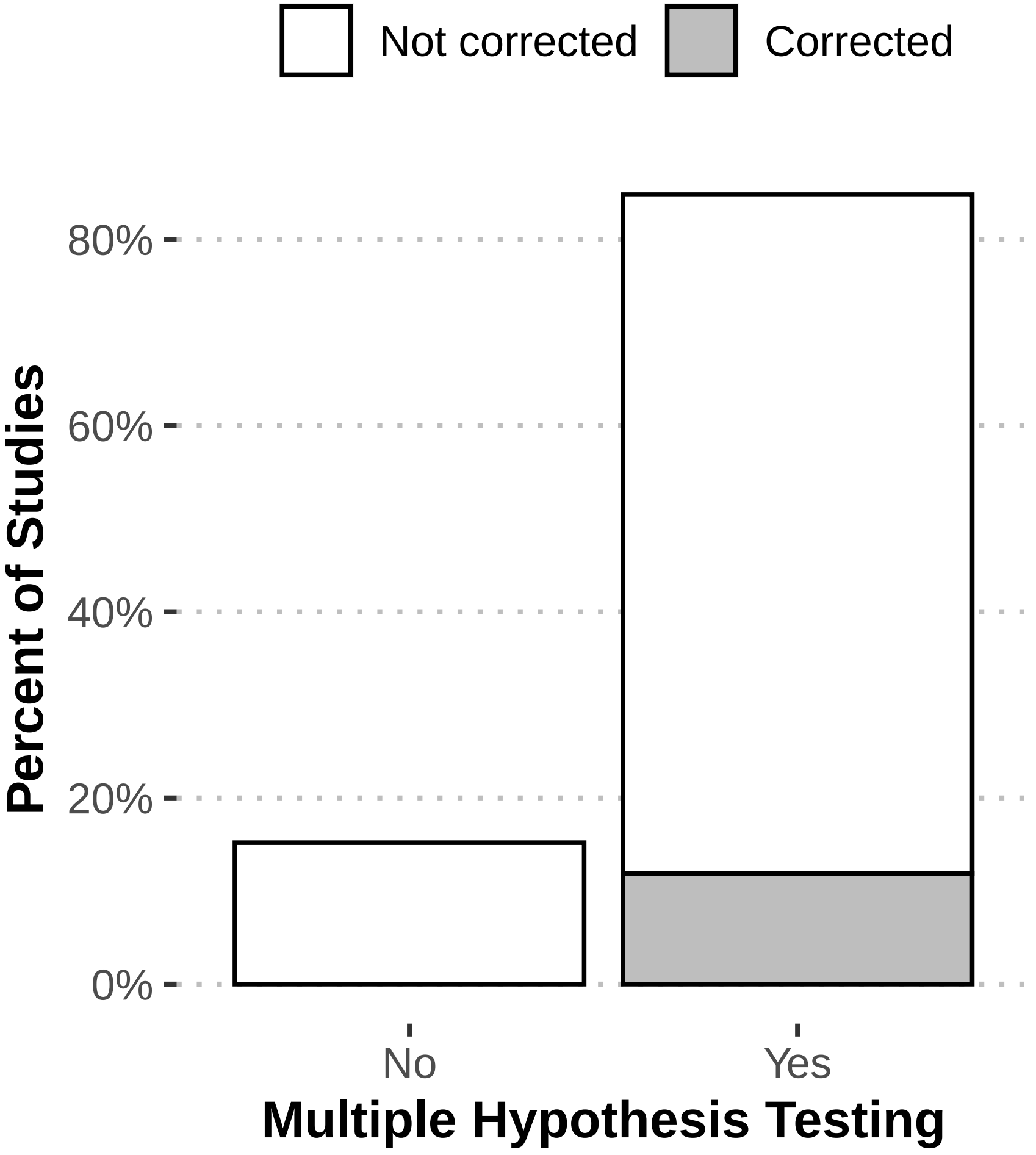

Not corrected
Corrected
80%
60%
40%
20%
0%
Percent of Studies
No
Yes
Multiple Hypothesis Testing

Fig 5

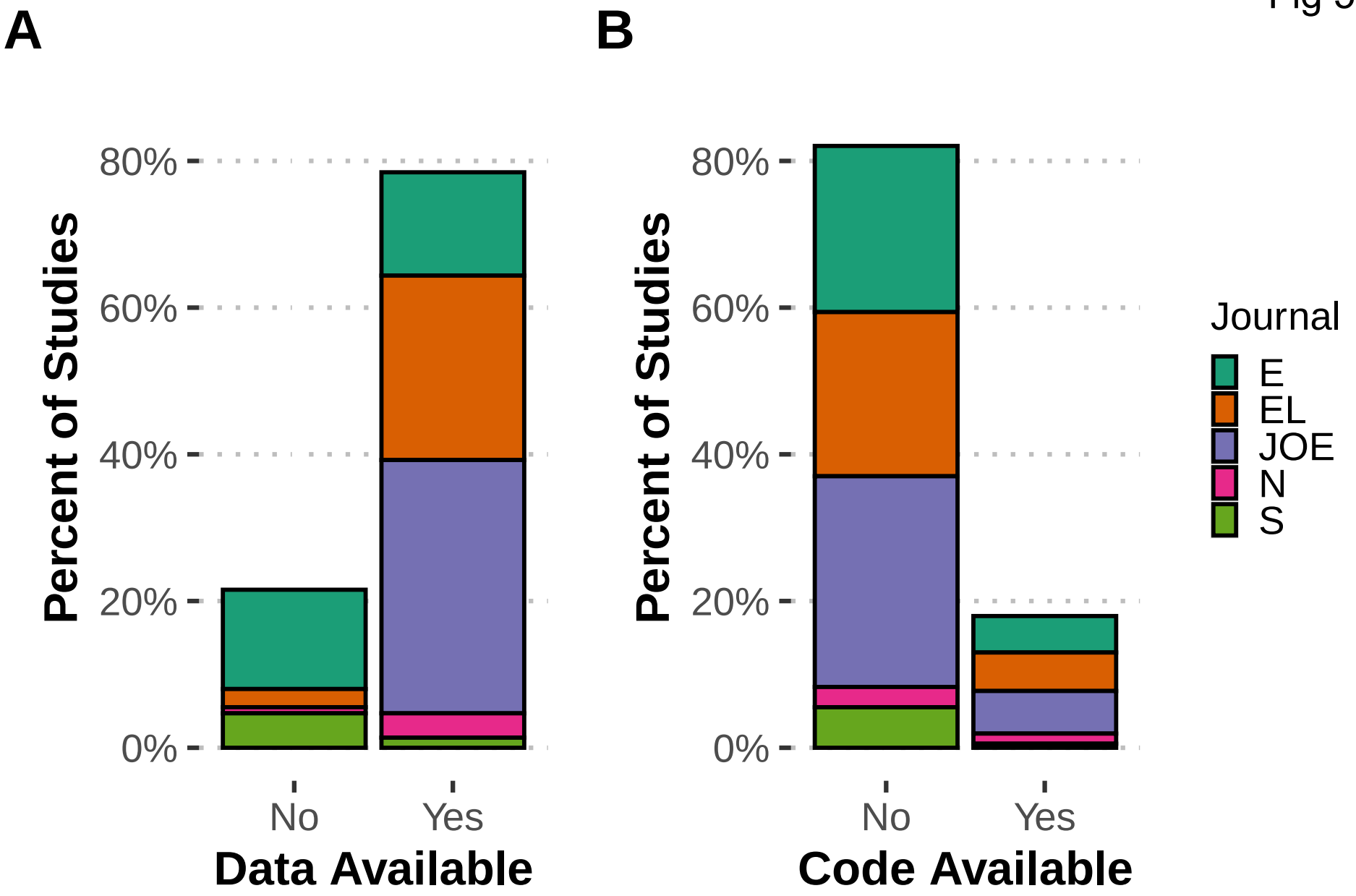
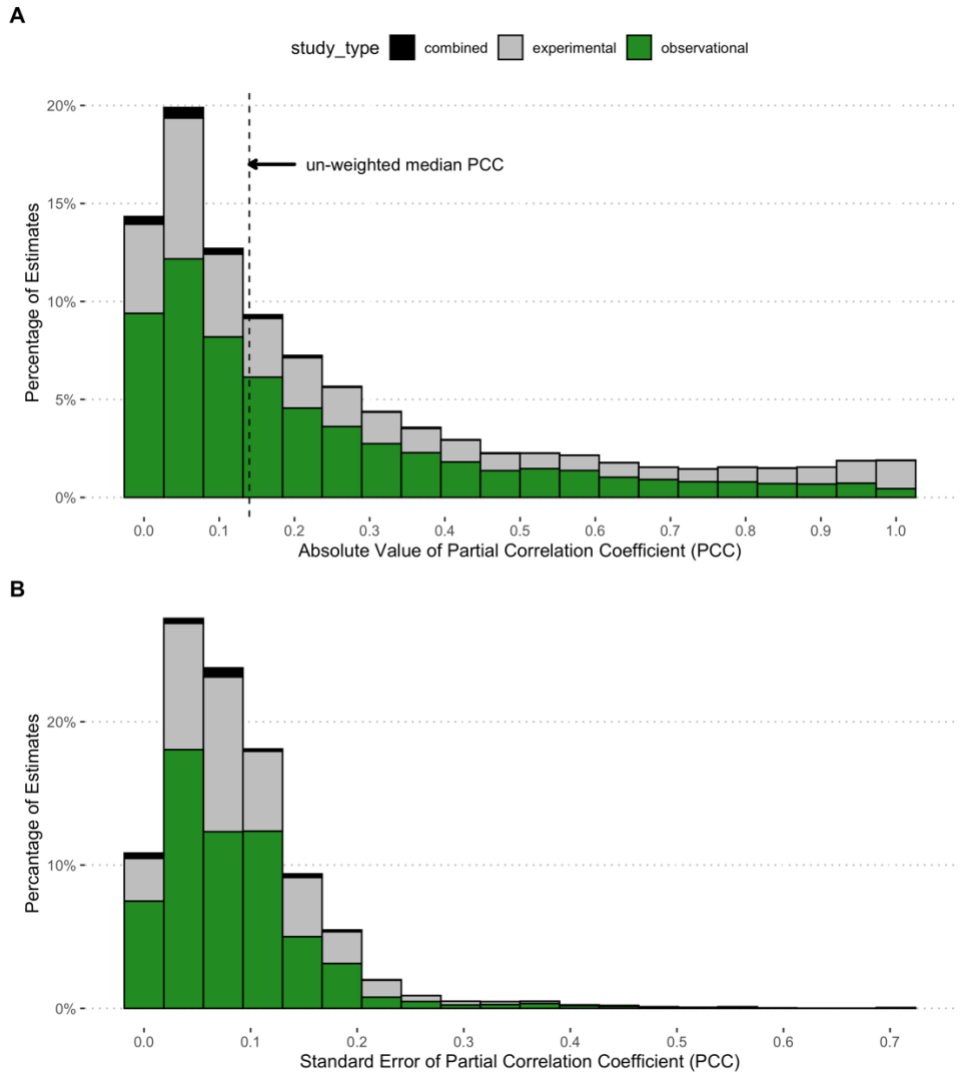

A
B
80%
60%
40%
20%
0%
Percent of Studies
No
Yes
Data Available
Code Available
Journal
E
EL
JOE
N
S

TABLE OF CONTENTS

Supplementary Text



Supplemental Figure 1. Unweighted distribution of (A) partial correlation coefficients (PCC) and (B) standard errors of the partial correlation coefficients calculated in our collection of studies. The weighted mean PCC value was 0.06, and the unweighted median is shown at the dashed line in A (~0.15). In both graphs, bars are colored according to the proportion of observational (green), experimental (gray), and combined (black) studies in that bin (n = 18,909 estimates from 353 papers).

Promising actions that contribute to the larger-scale systemic changes that are needed

Checklists & Data and Code Sharing Requirements

We faced multiple challenges in aggregating the data from our set of published articles because studies often did not report key information. For example, determining sample sizes was not always straightforward. In some cases, we had to make assumptions about the total sample size when the authors ran different analyses but did not report changes in sample size across the analyses. We had to exclude 5,484 estimates from 34 studies because we could not determine the sample size that the researchers were using (see "Data Cleaning"). While it is likely that most ecologists do not intentionally leave out important information, leaving this information out makes it difficult to interpret the results or aggregate them into meta-analyses.

So that readers may adequately judge the methods, analysis, and results in a study, ecologists should make sure to report all necessary information. Necessary information includes sample sizes and degrees of freedom for each analysis, estimates of error or uncertainty, and descriptions of the originally planned analyses and any deviations from those plans [1].

Checklists at multiple stages of the publication process can help researchers and reviewers include necessary information [2–4]. Checklists are used to reduce mistakes in other professions like surgery [5] and airplane piloting [6]. Individual labs, departments, or professional societies can provide checklists to researchers for standardized information to report in all publications [2]. More impactfully, journals can provide checklists that authors must fill out before submitting their manuscripts, similar to *Nature* (https://www.nature.com/documents/nr-reporting-summary-flat.pdf). Further, reviewers can be provided checklists as well to standardize what they should be looking for when accessing the soundness of methods, analysis, and reported results [4]. Checklists at the review stage may also reduce bias against negative results, which tend to be

scrutinized more than positive results [4,7]. Overall, checklists should provide an easy way to increase the transparency of ecological publications and make it easier for readers to find the necessary information to synthesize effect sizes and uncertainty in those estimates.

Researchers should also be required to provide data and code as a condition for manuscript publication (and the code should run with little or no manipulation). Exceptions can be allowed for some proprietary data. Many journals are moving towards encouraging data and code sharing, but few require archiving of both data and code [8]. Such requirements do, however, seem to increase the likelihood of providing data and code. For example, in our dataset, every paper in *Journal of Ecology* had data available, which highlights the effectiveness of journals requiring data archiving once papers are accepted (https://besjournals.onlinelibrary.wiley.com/hub/editorial-policies#archiving). Indeed, providing data and analytic code increases the transparency of workflows and conclusions reported in studies [1,9–11]. Journals may even consider having a reviewer check code files to see if the study results are reproducible with the code and data that they authors provide (see, for example, the data editor positions at the *American Naturalist* [http://comments.amnat.org/2021/01/note-since-fall-2020-robert-montgomerie.html], the *Journal of Evolutionary Biology* [https://jevbio.net/data-editing-at-jeb/ and http://comments.amnat.org/2021/01/note-since-fall-2020-robert-montgomerie.html], and the *American Economic Review* [12]). This extra step will further ensure the computational replicability of results, even at the potential monetary cost of this extra step.

Pre-registration and Pre-Analysis Plans

A pre-analysis plan describes the research questions, the study design, and the methods that will be used in a study. As its name suggests, the plan is completed before data analysis begins

(ideally, before all the data have been collected). Pre-registration is the process of registering, before the study or data analysis begins, a researcher's intent to undertake a study and the study's pre-analysis plan [13]. Ideally, the pre-registration is digitally time-stamped and publicly available, so that third parties can confirm which questions and analyses were anticipated in advance and which were devised only after collecting, and perhaps analyzing, the data. To prevent competing researchers from "scooping" a study prior to its publication, pre-registration platforms typically allow researchers to keep their pre-registration private while the research is completed, although sometimes the length of this embargo is limited to several years [14].

Preregistered analysis plans provide two main benefits. First, they help scholars quantify the "file drawer" problem: studies that were proposed, and perhaps completed, but never published. Studies may not be published for many reasons, but one reason is that the authors believed, or observed, that the results would not be acceptable to editors and peer (e.g., null results or statistically significant, but small estimated effects). Without pre-registration, scholars have no idea how many studies have been proposed and perhaps completed, but never published. That lack of knowledge can be costly for science; costly in terms of unnecessary repetition of studies and, when only serendipitously impressive results get published, exaggerated scientific claims. Knowing the full set of studies that may have been completed is also critical for ensuring that meta-analyses provide an accurate picture of what scientists have discovered [15].

Second, pre-registered plans help scientists to be transparent in all their research decisions. Science benefits when scholars are limited in their ability to selectively report or frame their results after seeing the impact of their decisions on their results. For example, pre-registered plans help to clearly demarcate confirmatory analyses from exploratory analyses [13,14,16]. Confirmatory analyses seek to test a specific hypothesis or estimate a specific parameter,

whereas exploratory analyses probe the data to look for interesting patterns. For example, a confirmatory analysis may seek to estimate the effect size of phosphorus addition on plant productivity, whereas an exploratory analysis may use the same data to see whether phosphorous addition is correlated with any other ecosystem functions that are measured in the data set. Exploratory analyses are important because they help scientists generate hypotheses that can then be tested with different data. Yet when exploratory analyses are repackaged in publications as confirmatory analyses, science suffers. Indeed, these repackages exploratory analyses never have the chance to be falsified and may need complex hypothesis to accommodate the results[17]. A related problem is when an author, after seeing the results from an analysis, changes the hypothesis to better match the results (i.e., HARKing). Ideally, the author would report in the article that the published hypothesis was not the original hypothesis and thus readers should treat the analysis as more exploratory than confirmatory. With pre-registration, even if the author does not report this deviation from the original plan, a reviewer or reader of the article could easily check the study's pre-registration document to confirm whether the hypotheses reported were the original hypotheses of the study [14,18].

Although pre-registration and pre-analysis plans are commonly associated with experimental designs, they can, and ought to be, used for all study designs. In fact, given that observational designs typically offer many more degrees of researcher freedom than experimental analyses, pre-registered plans may be even more important in observational designs than experimental designs.

Although journals and funders in ecology could require researchers to pre-register their studies and analysis plans [13,16], we believe the widespread adoption of pre-registration in ecology will take time because ecologists will need to become accustomed to working out details that often

were left for the post-data collection phase. When starting the preregistration process, it may be difficult for researchers to anticipate all the choices they will have to make during the analysis phase [16]. For example, a researcher may not have decided what to do with outliers or how to transform skewed data. These additions to, and deviations from, the original plan can be incorporated into amendments to the pre-analysis plan and can be reported in the final publication. The point of preregistration is not to punish researchers for failing to anticipate an obstacle, but to promote transparency during all steps of the research process [13], especially when researchers may forget what the original plan was and what deviations were made. Ideally, all pre-analysis plans would be registered before a study begins, but what does pre-registration mean for ongoing studies? In cases in which data collection is ongoing, researchers should try to preregister their subsequent analyses before new data are collected. As new ideas arise for old datasets, pre-analysis should also be submitted even though some of the data may be known to the researchers [14].

In ecology, pre-analysis plans ought to include detailed methodology that relates to several of the issues we describe above. For example, ecologists should include some reasoning about why they chose a specific sample size, including any design calculations that justify the sample size or elucidate the uncertainty within a study design (e.g., power analyses for frequentist methodologies, assurance analyses for Bayesian methodologies [19], or other design calculations [20]). In many cases, these design calculations will likely show that the number of replicates needed to credibly isolate signal from noise (e.g., power greater than 0.8) is logistically infeasible in terms of space, time, or money. Such conclusions do not mean that the studies should not be undertaken [21], but rather highlight the need for more coordination across study teams and a greater reliance on meta-analyses rather than single studies in ecology [15]. Pre-

analysis plans should also include rationale with respect to correcting or not for multiple hypothesis testing. As noted above, studies testing multiple hypotheses in ecology are common, but few papers correct for these comparisons or state why they chose not to use corrections. In some cases, a simple solution is to differentiate, in the pre-analysis plan, the "primary" hypothesis from the "secondary" hypotheses. This differentiation implicitly frames some planned analyses as confirmatory (primary hypothesis) and others as exploratory (secondary hypotheses).

In sum, pre-registration and pre-analysis plans reduce, or at least make more transparent, the practices of HARKing, selective reporting of results, and presenting *ex post* exploratory analyses as if they were part of the original design [14]. Some authors argue that pre-registration and pre-analysis plans are unnecessary if scientists are transparent in all their decisions in their manuscripts and that they create an unnecessary barrier to conducting science [22]. However, when clinical trials in heart, blood and lung treatments were required to be preregistered, the pattern of reported results changed dramatically: in comparison to findings reported before preregistration was required, the magnitudes of the reported treatment effects decreased substantially with a corresponding increase in the number of negative and null findings [23].

Pre-registered plans do not limit science. Rather, they limit the ways scientific results can be reported. Ecologists should be encouraged to explore their data or frame the results in ways that were not originally envisioned – but ecologists should also be required to report those deviations and the scientific community should have a way to confirm that those deviations are reported. Pre-registration and pre-analysis plans help to achieve this goal.

Registered Reports & Results-Blind Reviews

Another step towards increased transparency is Registered Reports – a two stage review process (https://www.cos.io/initiatives/registered-reports). During the Registered Report process, an introduction and methods section outlining the study design and analysis are submitted for peer review. The merit of the study is judged based on the question being asked and the methods used to address that question, rather than the sign, magnitude, or statistical significance of the results. After a study is accepted in the first phase of the review process, reviewers in the second phase judge how closely the study follows the original plan and whether any deviations are substantial enough to affect the study quality [24].

Registered Reports should reduce selective reporting of results. Studies have shown that registered reports decrease the amount of positive findings compared to conventional publication practices [25,26]. Registered reports should also help reviewers focus on the importance of the questions asked and quality of the study design, rather than the sign, magnitude, and statistical significance of the results. Indeed, a study found that researchers rated Registered Reports as being more rigorous in methodology and analysis, while not reducing novelty or creativity compared to non-Registered Report publications [27]. By emphasizing research questions and designs, registered reports make it more likely that ecologists can abandon NHST based on simple binary rules to decide when an estimate is ecologically relevant (e.g., if $p<0.05$ or Bayes Factor > 3), a practice that warps the presentation and interpretation of empirical results [28–32].

While Registered Reports are growing in popularity, few ecology publications are in this format. Currently, 12 ecology-related or general interest journals offer a Registered Reports option for submitting manuscripts (https://www.cos.io/initiatives/registered-reports; Supplemental Table 1). While the option for submitting Registered Reports has been around for several years at some journals, it seems that few researchers are aware of or using the process. For example,

Conservation Biology has published three Registered Reports, *Ecology and Evolution* has published only one, and none have been received at the *Journal of Plant Nutrition and Soil Science*. These journals are leading the way on Registered Reports, but there may need to be other incentives to have this publication format become more popular. For example, funding agencies could require this format, journals could spotlight these types of publications, or departments could require or up-weight publications in this format for career advancement. A preliminary written dissertation plan, where students' ideas and methods are critiqued by faculty, is already almost in the Registered Report format [16]. Thus, moving from the status quo towards greater use of Registered Reports is feasible and could be easily adopted for both early and later-career researchers.

Supplemental Table 1. Ecology or general interest journals that offer Registered Report format as of January 16, 2023.

Journal Name	Website
BMC Biology	https://bmcbiol.biomedcentral.com/
Ecology & Evolution	https://onlinelibrary.wiley.com/journal/20457758
Ecological Solutions & Evidence	https://besjournals.onlinelibrary.wiley.com/journal/26888319
Environment International	https://www.journals.elsevier.com/environment-international/
Frontiers in Plant Science	https://www.frontiersin.org/journals/plant-science#
Journal of Plant Nutrition and Soil Science	https://onlinelibrary.wiley.com/journal/15222624
Nature Communications	https://www.nature.com/ncomms/
PeerJ Life and Environment	https://peerj.com/life-environment/

PLoS Biology	https://journals.plos.org/plosbiology/
PLoS One	https://journals.plos.org/plosone
Royal Society Open Science	http://rsos.royalsocietypublishing.org/
Scientific Reports	https://www.nature.com/srep/

Similar to Registered Reports, results-blind reviews are another option to reduce publication bias against negative results [33]. In fact, results-blind reviews may be a good first step because they are closest to the current review process. Unlike Registered Reports where the study only starts after the first review, researchers submitting a results-blind review may have completed the study and written a complete manuscript – they simply do not include the results as part of the submitted manuscript. Like Registered Reports, results-blind review can reduce reviewer bias against negative results and can mitigate the pressure to engage in NHST guided by binary decision rules. Unlike Registered Reports, however, it has no mechanism in place to reduce selective reporting of results by the authors [14,24,33].

Changing Incentives

In the "publish or perish" environment in which many researchers operate, the benefits of engaging in these best practices are unlikely to exceed the costs without buy-in from the institutions that matter - namely, employers, funders, and publishers. For example, funding agencies could prioritize studies that use Registered Reports, such that high-profile grant programs reinforce best practices in ecology. Employers should explicitly encourage examples of credible, reproducible research and could require the practices outlined above for career advancement in a way that, as a metric of success, puts best practices on par with number of publications and impact factors of journals.

Among the practices that should be encouraged by employers, funders and publishers are replications of prior studies. Despite prior publications on the importance of replications [11,34], one study found replications were rare in ecology [35]. Employers should value researchers who replicate studies just as much as researchers who find novel results. High impact journals can help make replications more professionally rewarding by publishing replications alongside of ground-breaking research.

Without a change in researcher incentives it is difficult to imagine that a change in research practices will happen on its own – despite how much scientists value credibility within their discipline [36]. Unfortunately, researchers' professional incentives to publish novel and exciting studies are often at odds with their personal values of creating and disseminating credible science [2,36,37]. In fact, an ecology researcher who unilaterally adopts these practices may find herself at a disadvantage in the competition to place studies in high impact journals.

Supplemental References

1. Parker, T. H., Nakagawa, S. & Gurevitch, J. Promoting transparency in evolutionary biology and ecology. *Ecol. Lett.* **19**, 726–728 (2016).
2. Nosek, B. A., Spies, J. R. & Motyl, M. Scientific Utopia: II. Restructuring Incentives and Practices to Promote Truth Over Publishability. *Perspect. Psychol. Sci.* **7**, 615–631 (2012).
3. Simmons, J. P., Nelson, L. D. & Simonsohn, U. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychol. Sci.* **22**, 1359–1366 (2011).
4. Parker, T. H. *et al.* Empowering peer reviewers with a checklist to improve transparency. *Nat. Ecol. Evol.* **2**, 929–935 (2018).
5. Treadwell, J. R., Lucas, S. & Tsou, A. Y. Surgical checklists: A systematic review of impacts and implementation. *BMJ Qual. Saf.* **23**, 299–318 (2014).
6. Degani, A. & Wiener, E. L. Cockpit checklists: Concepts, design, and use. *Hum. Factors* **35**, 345–359 (1993).
7. Emerson, G. B. *et al.* Testing for the presence of positive-outcome bias is peer review. *Arch. Intern. Med.* **170**, 1934–1939 (2010).
8. Culina, A., van den Berg, I., Evans, S. & Sánchez-Tójar, A. Low availability of code in ecology: A call for urgent action. *PLoS Biol.* **18**, 1–9 (2020).
9. Munafò, M. R. *et al.* A manifesto for reproducible science. *Nat. Hum. Behav.* **1**, 1–9 (2017).
10. Nosek, B. A. *et al.* Promoting an open research culture. *Science (80-.).* **348**, 1422–1425 (2015).

11. Nakagawa, S. & Parker, T. H. Replicating research in ecology and evolution: Feasibility, incentives, and the cost-benefit conundrum. *BMC Biol.* **13**, 1–6 (2015).
12. Vilhuber, L. Report by the AEA Data Editor. *AEA Pap. Proc.* **109**, 718–729 (2019).
13. Nosek, B. A., Ebersole, C. R., DeHaven, A. C. & Mellor, D. T. The preregistration revolution. *Proc. Natl. Acad. Sci. U. S. A.* **115**, 2600–2606 (2018).
14. Parker, T., Fraser, H. & Nakagawa, S. Making conservation science more reliable with preregistration and registered reports. *Conserv. Biol.* **33**, 747–750 (2019).
15. Nichols, J. D., Kendall, W. L. & Boomer, G. S. Accumulating evidence in ecology: Once is not enough. *Ecol. Evol.* **9**, 13991–14004 (2019).
16. Nosek, B. A. *et al.* Preregistration Is Hard, And Worthwhile. *Trends Cogn. Sci.* **23**, 815–818 (2019).
17. Rubin, M. When does HARKing hurt? Identifying when different types of undisclosed post hoc hypothesizing harm scientific progress. *Rev. Gen. Psychol.* **21**, 308–320 (2017).
18. National Academy of Sciences publishes questionable, potentially chance findings on the crime effects of restoring vacant urban lots. *Straight Talk on Evidence* https://www.straighttalkonevidence.org/2018/09/11/national-academy-of-sciences-publishes-overstated-findings-on-the-crime-effects-of-restoring-vacant-lots-that-could-have-appeared-by-chance/ (2018).
19. Chen, D. G. (Din) & Ho, S. From statistical power to statistical assurance: It's time for a paradigm change in clinical trial design. *Commun. Stat. Simul. Comput.* **46**, 7957–7971 (2017).
20. Gelman, A. & Carlin, J. Beyond Power Calculations: Assessing Type S (Sign) and Type M (Magnitude) Errors. *Perspect. Psychol. Sci.* **9**, 641–651 (2014).

21. Lemoine, N. P. *et al.* Underappreciated problems of low replication in ecological field studies. *Ecology* **97**, 2562–2569 (2016).

22. Rubin, M. Does preregistration improve the credibility of research findings? *Quant. Methods Psychol.* **16**, 376–390 (2020).

23. Kaplan, R. M. & Irvin, V. L. Likelihood of null effects of large NHLBI clinical trials has increased over time. *PLoS One* **10**, 1–12 (2015).

24. Button, K. S., Bal, L., Clark, A. & Shipley, T. Preventing the ends from justifying the means: Withholding results to address publication bias in peer-review. *BMC Psychol.* **4**, 1–7 (2016).

25. Allen, C. & Mehler, D. M. A. Open science challenges, benefits and tips in early career and beyond. *PLoS Biol.* **17**, 1–14 (2019).

26. Scheel, A. M., Schijen, M. R. M. J. & Lakens, D. An Excess of Positive Results: Comparing the Standard Psychology Literature With Registered Reports. *Adv. Methods Pract. Psychol. Sci.* **4**, (2021).

27. Soderberg, C. K. *et al.* Initial evidence of research quality of registered reports compared with the standard publishing model. *Nat. Hum. Behav.* **5**, 990–997 (2021).

28. Yoccoz, N. G. Use, Overuse, and Misuse of Significance Tests in Evolutionary Biology and Ecology. *Bull. Ecol. Soc. Am.* **72**, 106–111 (1991).

29. Fidler, F., Fraser, H., McCarthy, M. A. & Game, E. T. Improving the transparency of statistical reporting in Conservation Letters. *Conserv. Lett.* **11**, 1–5 (2018).

30. Amrhein, V. & Greenland, S. Retire statistical significance. *Nature* **567**, 305–307 (2019).

31. Wasserstein, R. L., Schirm, A. L. & Lazar, N. A. Moving to a World Beyond "p < 0.05". *Am. Stat.* **73**, 1–19 (2019).

32. Anderson, D. R., Burnham, K. P. & Thompson, W. L. Null hypothesis testing: Problems, prevalence, and an alternative. *J. Wildl. Manage.* **64**, 912–923 (2000).

33. Smulders, Y. M. A two-step manuscript submission process can reduce publication bias. *J. Clin. Epidemiol.* **66**, 946–947 (2013).

34. Fraser, H., Barnett, A., Parker, T. H. & Fidler, F. The role of replication studies in ecology. *Ecol. Evol.* **10**, 5197–5207 (2020).

35. Kelly, C. D. Rate and success of study replication in ecology and evolution. *PeerJ* **2019**, (2019).

36. Anderson, M. S., Martinson, B. C. & De Vries, R. Normative dissonance in science: Results from a national survey of U.S. scientists. *J. Empir. Res. Hum. Res. Ethics* 3–14 (2007).

37. O'Dea, R. E. *et al.* Towards open, reliable, and transparent ecology and evolutionary biology. *BMC Biol.* **19**, 1–5 (2021).