nature portfolio

Peer Review Information

Journal: Nature Ecology & Evolution
Manuscript Title: Empirical evidence of widespread exaggeration bias and selective reporting in ecology
Corresponding author name(s): Paul J. Ferraro

Editorial Notes:

nature portfolio

Reviewer Comments & Decisions:

Decision Letter, initial version:

20th May 2022

Dear Dr. Ferraro

Thank you very much for your enquiry about submitting a manuscript to Nature Ecology & Evolution.

I've now had a chance to discuss your work with my colleagues, and although we think that it sounds very interesting, we are still uncertain as to the degree to which the study would be suitable for the journal. For example, we are interested in the justification for the using only a small number of selected journals.

Therefore, we would like to invite you to submit the full manuscript to Nature Ecology & Evolution so that we can examine the data -- especially details of the survey -- before deciding whether to send the paper out to review.

If this is acceptable to you, you can submit the complete manuscript using the link below:

[REDACTED]

If you have any questions, please feel free to contact me.

[REDACTED]

Decision Letter, first revision:

11th November 2022

*Please ensure you delete the link to your author homepage in this e-mail if you wish to forward it to your co-authors.

Dear Dr Ferraro,

Thanks for you patience during the review process, which I appreciate was longer that i'm sure you would have hoped. I can confirm that your manuscript entitled "Reliability of empirical evidence in ecology and a proposal for action" has now been seen by 4 reviewers, whose comments are attached. The reviewers have raised a number of concerns which will need to be addressed before we can offer publication in Nature Ecology & Evolution. We will therefore need to see your responses to the

criticisms raised and to some editorial concerns, along with a revised manuscript, before we can reach a final decision regarding publication.

In particular, Referees #1, #2 and #3 all find the study interesting and timely, but have some relatively minor suggestions for changes. For example, Referee #1 feels that some toning down is needed regarding the utility of the average effect size, as well as the sections on multiple comparisons and power analysis. Referee #2 also feels that clearer justification is needed for the inclusions of the specific journals in the study, as well as why these particular practices were chosen for evaluation. Referee #3 has requested that metadata by made available.

You will also see that Referee #4 has looked closely at the survey methods, and feels that they do not meet the criteria for a rigorous, and therefore useful, survey. As such, we feel that we cannot continue to consider this aspect of the work in the Analysis, and suggest that it be removed from the next revision.

We therefore invite you to revise your manuscript taking into account all reviewer and editor comments. Please highlight all changes in the manuscript text file [OPTIONAL: in Microsoft Word format].

We are committed to providing a fair and constructive peer-review process. Do not hesitate to contact us if there are specific requests from the reviewers that you believe are technically impossible or unlikely to yield a meaningful outcome.

When revising your manuscript:

* Include a "Response to reviewers" document detailing, point-by-point, how you addressed each reviewer comment. If no action was taken to address a point, you must provide a compelling argument. This response will be sent back to the reviewers along with the revised manuscript.

* If you have not done so already please begin to revise your manuscript so that it conforms to our Analysis format instructions at http://www.nature.com/natecolevol/info/final-submission. Refer also to any guidelines provided in this letter.

* Include a revised version of any required reporting checklist. It will be available to referees (and, potentially, statisticians) to aid in their evaluation if the manuscript goes back for peer review. A revised checklist is essential for re-review of the paper.

Please use the link below to submit your revised manuscript and related files:

[REDACTED]

Note: This URL links to your confidential home page and associated information about manuscripts you may have submitted, or that you are reviewing for us. If you wish to forward this email to co-authors, please delete the link to your homepage.

We hope to receive your revised manuscript within four to eight weeks. If you cannot send it within this time, please let us know. We will be happy to consider your revision so long as nothing similar has been accepted for publication at Nature Ecology & Evolution or published elsewhere.

Nature Ecology & Evolution is committed to improving transparency in authorship. As part of our efforts in this direction, we are now requesting that all authors identified as 'corresponding author' on published papers create and link their Open Researcher and Contributor Identifier (ORCID) with their account on the Manuscript Tracking System (MTS), prior to acceptance. ORCID helps the scientific community achieve unambiguous attribution of all scholarly contributions. You can create and link your ORCID from the home page of the MTS by clicking on 'Modify my Springer Nature account'. For more information please visit please visit www.springernature.com/orcid.

Please do not hesitate to contact me if you have any questions or would like to discuss these revisions further.

We look forward to seeing the revised manuscript and thank you for the opportunity to review your work.

[REDACTED]

Reviewer expertise:

Reviewer #1: Open science, ecology, replication

Reviewer #2: Open science, ecology

Reviewer #3: Open science, ecology, replication

Reviewer #4: Expert elicitation

Reviewers' comments:

Reviewer #1 (Remarks to the Author):

This manuscript presents empirical evidence (based on a large sample of the recent ecology literature and a survey of ecologists) with the aim of assessing the degree of bias in the ecology literature. The evidence in this paper is thorough and diverse, and taken together, will be interesting to many ecologists as well as to meta-scientists because it represents an empirical advancement in our understanding of the issue of bias in the ecology literature.

The core evidence presented in this manuscript is a set of many thousands of statistical effects

extracted from ecology articles published over 2018, 2019, and part of 2020 in five high profile journals. Based on these effect sizes, the authors calculate an average effect size and then calculate the average power of the studies in this sample to reach nominal statistical significance if the average effect size across studies is assumed to be the true effect size. The authors are explicit in their acknowledgement that they do not view the average effect size as the true effect size. Instead, they present it as a useful benchmark for comparison. I think it is valuable to present this analysis in this manuscript, but as I discuss in more detail in several places below, I think the authors have overstated the utility of this approach. This average effect size, though credible, is not as useful as the authors argue (because it is probably actually an underestimate of the sizes of the effects that researchers want power to distinguish from zero). So, many of my comments below are related to my request that the authors moderate their statements about the conclusions that can be drawn from these data and acknowledge additional shortcomings of these data. All that said, the manuscript does present evidence that allows us to evaluate the hypothesis that, under a range of plausible conditions, many effects published in ecology are inflated, and this is important and should be of broad interest.

The authors also present valuable evidence to evaluate the rate of selective reporting in published papers and their supplements, useful evidence regarding rates of data and code archiving, and useful evidence regarding multiple hypothesis testing (and corrections thereof) within individual papers, as well as results of an interesting survey of ecologists regarding statistical power.

I make specific suggestions regarding most of these sections below, organized by line number.

7-8: "We show that most ecology studies are underpowered" – this statement rests on an important assumption that I don't think is sufficiently supported and therefore I think this statement should be moderated somewhat. I explain my concerns about this assumption below.

8: "that interacts with publication biases" – I think this statement should also be moderated to something like "can interact with" or even "may be likely to interact with"

25-26: ideas, hypotheses, models etc. are also important. Maybe you could say something like "an essential component of advancement"

26: maybe "and often to inform" (not all ecologists are doing these things)

33: references 2-4 are a surprising choice to support your statement about debate about incentives. These references all address the issue of publication bias, but they do not really explore or give contrasting perspectives regarding the incentives that might drive these biases.

42: you might consider replacing "imply" with "demonstrate"

52: it seems to me that this paragraph would be best if it addressed just your second aim. The first aim is simply the background info that would normally go into an introduction to justify the study and the third aim is simply expected content of a discussion section on such a topic.

76-80: this is an important point that merits more explanation. A naïve reader would benefit from

some more detailed explanations of how and why these statements are true

94: your "Thus" statement here is not sufficiently clearly linked to the prior statement. Again, think of the naïve reader

106: I have long had a concern about this method of estimating the average effect size for the sake of estimating power. Given that some hypotheses are wrong, in other words, the hypothesized relationships are unimportant/trivial/tiny, we don't want power to detect those. Instead, we want high power to detect (distinguish from zero) relationships that are large enough to be meaningful (this threshold between meaningful and not meaningful is obviously both subjective and contextual, but it will often exist). So, if an estimate of the average effect size we want to detect is based in part on many effects that are trivial and which we therefore do not want to detect (as yours surely is), we will dramatically underestimate our power. (see below for two further discussions of the relevance of hypothesis testing vs estimation)
That said, I appreciate that you also presented analyses based on other (more plausible) potential target effect sizes (though I would like to see presentation of wider range [extending larger than 0.2]). I would further appreciate an explicit de-emphasis of your calculated mean effect size (explained to the reader in the context of the reasoning I just presented above), and a corresponding additional emphasis on larger potential target effect sizes.

(I want to acknowledge that in some cases, ecologists will want a precise estimate of a very small effect [but see more on this below]. However, my sense is that most ecologists are happy to consider very small effect sizes as approximately zero, and thus do not require precise estimates of very small effects. I also want to point out that, unfortunately, many researchers in ecology remain more focused on dichotomous hypothesis testing than on estimating sizes of effects)

107: for reasons stated above, I think you probably actually underestimate power (despite your confident statement of overestimation).

125, also Fig 1: see comments above

129: effects of 0.2 are often described as moderate or between small and moderate (usually not large). Of course this is all relative to expectation, but effects can certainly be much larger than 0.2 in ecology studies.

130-135: these arguments rest on the assumption that you have used accurate estimates of effect sizes that researchers want to distinguish from zero.
I want to be clear that I am not saying I think that you are wrong, only that I don't think we understand as much about average effect size as you assume.

148-9: and authors!

163: instead of "only the large effects", maybe you should say something like "mostly the large effects" or "disproportionately the large effects" (or something to that effect)

171-174, Fig 2: These estimates are again based on the assumption that our goal should be power to detect or precisely estimate even trivially small effects.
The range of WAAP values you selected to display in Fig 2B is skewed towards small effects, and the calculated value of WAAP is probably an underestimate given that it presumable incorporates many trivial effects.
Similarly, the values shown in Fig 2a are also based on a calculation that incorporates trivial effects (into the estimate of the 'true' effect size).

178: some ecology is applied, but it is not universally an applied field. Maybe instead say something like "In a field where results often have application" or "In a field where results often have real-world application" or "In a field where results often have management implications" …

178-182: I very much agree with this statement that magnitude matters and that a major problem with publication bias is the bias in magnitude of published effects. However, much of your empirical work in this paper misrepresents the extent of the problem to some unknown amount because you focus on power, and 'power' is not primarily about magnitude, it is about crossing a threshold of 'statistical significance'. As true effect sizes get smaller and smaller, we need greater and greater power to distinguish these effect sizes from zero. However, as they get smaller and smaller, distinguishing these effect sizes from zero may often become less and less important to the researcher. In other words, if we focus on SE relative to the size of the estimate, we will need to make the SE smaller and smaller as the effect size gets smaller to make sure we cross the significance threshold. However, it may instead be the case that we want a certain amount of confidence in a point estimate (a certain absolute SE) regardless of the distance of the estimate from zero. In other words, we may sometimes want the same degree of absolute precision in our estimates regardless of the magnitude of the effect. If that is the case, then focusing on power to distinguish tiny/trivial effects from zero (as your study currently does inadvertently), is misguided.

I want to emphasize here that I like this study and I don't want to see very many substantial changes to your methods. Instead, I want to see more acknowledgement and discussion of the limits of your inference.

197: Amrhein (citation 30) is also an ecologist, but maybe you mean that the paper was published in a general science journal, in which case, this is fine

220: I think "necessarily" and "not" are reversed

262: This section of the study is the one I am least excited about.
There is legitimate debate about the circumstances that require correcting for multiple hypothesis tests and the methods that should be used for these corrections.

267-268: You state that "Even in the absence of selective reporting, multiple hypothesis testing without statistical corrections can inflate the prevalence of spurious results reported in the ecology literature".
I'm not sure what you mean here. If you mean that the absolute number of spurious results in the literature will increase as the number of tests within individual studies increases, then I agree.

However, I do not see the problem with that. As the number of different studies conducted on a topic increases, the number of spurious results will increase also, but this is not a bad thing. That is because, as the number of different studies increases, the number of effects contributing to our meta-analytic mean also increases, and so the precision of our mean increases. More effects will mean better precision of average effects whether those different effects come from the same study or from a different study. Thus, more effects calculated = better.
Instead, on the scale of the literature at large, multiple hypothesis testing is primarily a problem if it is combined with selective reporting or other deceptive practices.
On the scale of an individual study, multiple hypothesis testing is a problem for interpretation of the paper in question if the authors do not acknowledge the increase in the likelihood of false positives or strong relationships due to chance. This is clearly important when considering studies in isolation, but the much greater concern to me is multiple testing combined with some form of publication or reporting bias since that biases that literature at large.

273-274: this is an example of a debatable statement on this topic (and it is not backed up by citation). Some authors argue (and I apologize for not going back and finding any citations for you here) that corrections for multiple comparisons are only necessary when testing different predictions of the same hypothesis.

279: OK, good. Thanks for including this. However, this debate is not reflected in your earlier statements.

308: you may wish to add this recent citation on data archiving:
https://doi.org/10.1098/rspb.2021.2780

Supplement

56-59: please provide version numbers for packages

170: I am sympathetic to the argument you are making here, and I suspect that you are correct that many ecologists are not adjusting their sampling effort in accordance with anticipated effect size. However, I'm confident that such adjustments happen (I've seen them happen in some ecological disciplines), and so the real question should be, how often do they happen. I do not think we know the answer to this question, and so I think you should express a bit less certainty in your argument here. In other words, your statement on line 174-175 is overly confident.

230: could you please provide more information about the survey? For instance, what introductory text preceded the questions you present here? What sample of ecological tests were they told to hypothesize about?

337: please include a citation in this paragraph

444: where does this "may not be a top criterion" come from? Is there evidence of this in the Nosek

paper? This is certainly a top criterion at the hiring stage for many (most?) academic research institutions

Signed,
Tim Parker

Reviewer #2 (Remarks to the Author):

The manuscript deals with an important topic in ecology (and other sciences) - that of credibility and reliability, which has been on the radar for the past years. Generally, methods seem sound (though see my comments below) and findings are in line with previous work, calling for urgent action. I congratulate the authors on all the work conducted. I provide my comments below, structured following as Reviewers' guidelines.

- Key results: Please summarise what you consider to be the outstanding features of the work.

This work estimates the prevalence of several suboptimal practices in ecology based on publications from five popular journals. Results correspond to some previous estimates for ecology and definitely support calls for action on these issues. The dataset encompasses a large and comprehensive set of articles published in 5 journals.

- Validity: Does the manuscript have flaws which should prohibit its publication? If so, please provide details.

I find that MS is valid, with no major flaws, but I do have some issues with how the topic was presented, as I note below (Data & methods)

- Originality and significance: If the conclusions are not original, please provide relevant references. On a more subjective note, do you feel that the results presented are of immediate interest to many people in your own discipline, and/or to people from several disciplines?

I am unclear as to what the overlap between data collected here and those collected in previous studies on these topics is. Good references can be found in the supplementary materials of Purgar et al. 2022, https://doi.org/10.5281/ zenodo.6566100) - They list several studies that estimate power in ecology and publication bias (these studies are based on a larger corpus of primary studies) and also a study by Culina et al on reproducibility.
I think that the results are of wide interest, however with some adjustments (as I explain in my comments).

- Data & methodology: Please comment on the validity of the approach, quality of the data and quality of presentation. Please note that we expect our reviewers to review all data, including any extended data and supplementary information. Is the reporting of data and methodology sufficiently detailed and transparent to enable reproducing the results?

The methodology seems generally good, however, I have several comments.

My main comment is about the conceptual approach: It is unclear why were these particular practices (power, selective reporting, and multiple hypothesis testing) chosen, among all others that are known to happen (and could also be assessed). There are many other practices that contribute to the low reliability of scientific results. E.g. use of flowed data collection designs? While there is a very limited explanation on why (line 47 'widespread') research shows that other suboptimal practices are equally widespread. Even in the specification of research aims, it is stated that 'degree to which ecologists use empirical designs ….' [line 55] but later the analysis concentrates only on power, which is only one of the aspects of the empirical design.

Second, I would argue that while the first three practices are suboptimal (questionable) and can lead to unreliable results, the fourth practice (availability of codes and data) is of a different kind as it does not lead to unreliable but to unverifiable evidence, and also to the lower impact of work. My suggestion would thus be to make a distinction between low power, selective reporting, and multiple hypothesis testing vs availability of data and code. Indeed, the latter allow for identifying of the QRPs, and their correction (as recognized by authors in Table 1).

Other comments:
1. Why were these specific 5 journals selected? Can the results be representative for other journals publishing ecological research? While I do not think that other journals have to be considered, some discussion on the implication of this particular selection on the results should be there.

2. Given that only articles that represent data in tables were considered, can this introduce any bias in the results (e.g. I know that results of particular types of analysis are more likely to be presented in a table format).

3. It is unclear how were articles scored as appropriate for the research, i.e. being empirical (reading abstract? Or full texts? How many people independently scored these to avoid misclassification?). How many articles passed this stage?

4. Were data on statistical tests extracted manually? How many articles did have results reported in tables?

5. The final sample size differs from the main MS (353) and the supplement (366-29 dropped as of no sample size) articles. Maybe Im missing something.

6. How does the assumption that there is no selective reporting or publication bias against small effect sizes (109-110) affect the reliability of the estimates? Isn't this assumption against what is known (and even demonstrated in this MS).

7. Why was the survey only conducted regarding the power, and not to estimate other practices that this MS deals with?

8. I have checked the data and codes but have not run the analyses. Data and codes seem complete. They lack Readme or similar files to aid the understanding of what files represent, and what variable names mean.

- Conclusions: Do you find that the conclusions and data interpretation are robust, valid and reliable?

They generally are, but please see my other comments.

- References: Does this manuscript reference previous literature appropriately? If not, what references should be included or excluded?

I find some important references lacking, and think this work would benefit from considering these (and their findings)
1. Smalidno and McElreath (2016): Natural selection of bad science. https://royalsocietypublishing.org/doi/10.1098/rsos.160384
2. Fostemier et al. (2017): Detecting and avoiding likely false-positive findings – a practical guide. https://onlinelibrary.wiley.com/doi/10.1111/brv.12315
3. Ulrich, R. & Miller, J. (2020). Meta-Research: Questionable research practices may have little effect on replicability. eLife 9: e58237
4. Lakens (2021): Sample size justification. https://psyarxiv.com/9d3yf/
5. on the prevalence of QRP in the Netherlands https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0263023#sec008
6. On the prevalence of responsible research practices (to inform solutions) https://f1000research.com/articles/11-471
7. reference for line 74-76 is missing.

- Clarity and context: Is the abstract clear, accessible? Are abstract, introduction and conclusions appropriate?

Generally, the paper is well written and accessible. I would suggest that both the title and the abstract are more clear about what is studied in this work: some of the suboptimal research practices (rather than all of them). The abstract could also provide the values for the estimates.

Some suggestions:
1) line 27 predictive models are part of understanding how the world works.
2) line 29 please provide a reference (if any) for this statement. Also think this sentence needs rewording e.g. 'if the estimates of effect sizes are inaccurate, biased, and presented without uncertainty).
3) line 48/49 – this is unclear to me.
4) line 55 – please highlight that this is only about power, rather than about the other aspects of the empirical design
5) line 63 – think academic/research institutions should also be mentioned here
6) main MS would benefit from more information on how the data were collected [lines 68-69]
7) the first paragraph of the Underpowered design section could do better work in stating why power

is such a problem in ecology (please see some refs that I have provided previously). Currently, it just states what components will determine the power of the test, rather than what we know about these components (e.g. that most of the effects in ecology are small).
8) lines 81-83 are unclear to me.
9) lines 93-94 do not provide an explanation on why rejecting a null hypothesis will be unlikely in this case.
10) lines 147-149 sentence is a bit biased itself – most of the QRPs when coupled with publication bias can lead to exaggeration bias.
11) lines 184-188 are unclear to me.
12) if space needs to be spared I would suggest removing paragraphs 228-247. This is already mentioned in the next paragraph and can be just a bit elaborated on.
13) line 313 should be 'reproducibility' instead of replication.
14) not mentioned how many papers provide both code and data (lines 319/320)

- Please indicate any particular part of the manuscript, data, or analyses that you feel is outside the scope of your expertise, or that you were unable to assess fully.

I am not an expert in power analysis so could not judge that analysis nor what is represented in the paragraph lines 97-110. Especially I am unclear about the validity of the assumption that there is no selective reporting or publication bias against small effect sizes. Isn't this assumption against what is known (and even demonstrated in this MS).
I have also never conducted a survey thus not competent to judge the general approach.

Reviewer #3 (Remarks to the Author):

I was pleased to review NATECOLEVOL-220516528 "Reliability of empirical evidence in ecology and a proposal for action". The study is on an important and timely topic, and is well suited for Nature Ecology & Evolution. The authors carried out a considerable amount of work and the analyses appear sensible and rigorous. However, I was unable to carefully check the data and analysis code as there was no metadata provided alongside the data (i.e., it is challenging to make sense of what each data file is and how it fits into the broader analysis without running the script). The file names for the csv files are not particularly informative. Otherwise, the manuscript is clear, neatly structured, and well written. The supplementary material is informative. I have only minor comments and commend the authors on their excellent work.

Main text

Please include metadata (i.e., a readme file) describing the content of each data file, what the column headings are, abbreviations, and units. Column headings should not contain units, values, etc – this

information would be best placed in the readme file. For useful resources on this topic, see refs in Table 3 under "Open data" here: https://doi.org/10.1242/jeb.243559 - and (Towse et al. 2021; Contaxis et al. 2022). These resources are also relevant in the context of the manuscript's Table 1.

L35. It might be worthwhile clarifying (here or elsewhere) that the authors consider p-hacking as a component of selective reporting. I was surprised that there was no mention of p-hacking in the manuscript.

L94-95. This sentence does not appear to logically flow from the previous two IMO. Please consider rephrasing?

L99-100. Consider providing a more detailed explanation of the PCC as not all readers will be familiar with this statistical concept.

L137. It is unclear from the SM that the respondents were asked specifically about the authors' dataset. Perhaps clarify this in the SM?

L138-140. I was slightly confused by this sentence given that the authors only asked experimentalists how often they carry out power analyses (see SM L234-236). Consider rephrasing?

L200. Consider replacing insignificant with nonsignificant.

L290-291. It is unclear from the text if the authors evaluated whether studies reported why corrections were or were not used.

L299 – Data and Code Availability. It might be worth mentioning in this section that most publicly available datasets in E&E are unusable – i.e. open data are not necessarily FAIR data (Wilkinson et al. 2016; Roche et al. 2022) – see also https://www.go-fair.org/fair-principles/ and Box 2 in https://doi.org/10.1242/jeb.243559.

L305. Ecological Society of America (not American)

L317-320. There is no need to cite this paper but the authors might be interested in recent data for comparative physiology and behavioural ecology journals, which colleagues and I have reported here (see Table 1 and Fig 2): https://doi.org/10.1242/jeb.243559

L332. Again, no need to cite but perhaps of interest: Buxton et al (2021) Avoiding wasted research resources in conservation science. Conservation Science and Practice 3: e329 https://doi.org/10.1111/csp2.329. See section 5 " Openly and comprehensively report research outputs" and Table S1.

L340. Consider using "positive results" instead of "statistically significant results".

L346. I was pleased to see SORTEE mentioned here. If there is space, DORA might also be worth highlighting (https://sfdora.org/) – although it is not specific to E&E.

L355. Missing 'to'.

Table 1. Parker et al 2019 mention results-blind review but only in the context of registered reports, hence it might not be an appropriate reference here. Again, no need to cite but perhaps of interest – we list the potential costs and benefits of sharing open data in (Roche et al. 2014).

Supplementary Material

L19. Were 1,568 papers examined in total? Slightly unclear…

L21. Perhaps consider explaining why only in tables and not in the text?

L47. "We tried not to include robustness checks" – how was this done and why was it challenging?

L48. Were these exact words searched or did you employ stemming to broaden the search (https://libguides.mit.edu/c.php?g=175963&p=1158679)? For example, I would have expected "correct*" to be used as a keyword.

L68. 2 x 'the'

L164-165. It is unclear from this sentence why the costs of data collection or selecting study units are large when the expected treatment effect sizes are large.

L266. Reference to SM in the SM.

L294-296. The American Naturalist and the Journal of Evolutionary Biology now have data editors: see https://jevbio.net/data-editing-at-jeb/ and http://comments.amnat.org/2021/01/note-since-fall-2020-robert-montgomerie.html

L329-330: Consider explaining why science suffers when exploratory analyses are repackaged in publications as confirmatory analyses. This might not be obvious to many readers.

L408. Could you provide this list of 9 journals in a table?

I hope these comments are helpful in revising the manuscript.
Regards,
Dom Roche

References:

Contaxis N, et al. 2022. Ten simple rules for improving research data discovery. PLOS Computational Biology 18:e1009768.

Roche DG, Berberi I, Dhane F, Lauzon F, Soeharjono S, Dakin R, Binning SA. 2022. Slow improvement to the archiving quality of open datasets shared by researchers in ecology and evolution. Proceedings of the Royal Society B: Biological Sciences 289:20212780.

Roche DG, Lanfear R, Binning SA, Haff TM, Schwanz LE, Cain KE, Kokko H, Jennions MD, Kruuk LE. 2014. Troubleshooting public data archiving: suggestions to increase participation. PLoS Biology 12:e1001779.

Towse AS, Ellis DA, Towse JN. 2021. Making data meaningful: guidelines for good quality open data. The Journal of Social Psychology 161:395-402.

Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten J-W, da Silva Santos LB, Bourne PE. 2016. The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data 3:160018.

Reviewer #4 (Remarks to the Author):

Manuscript NATECOLEVOL-220516528A reports on an investigation into the manner in which ecologists conduct their research that could be potentially problematic. The authors present a range of pieces of evidence that they find cast doubt on the appropriateness of various research practices. They conclude that these practices present a challenge to the credibility of ecology as a discipline because they can lead to unreliable and exaggerated results being presented.

While I consider myself fairly proficient in statistical methods, I am not an expert in statistics. Therefore, I will refrain from commenting on the manuscript components that are focused on these aspects. Instead, I focus here on the authors' use of a survey of ecologists and the reporting of the methods and results from this survey. While I am a trained field ecologist, today I am using surveys frequently to conduct environmental social sciences research. The authors used the results from this survey as one piece of evidence among various others in support of their conclusions.

1. The survey was sent to the listserv of the Ecological Society of America (ESA) and Twitter. With >9,000 members, ESA is the largest national-level organization of its kind globally. While there may be many reasons why ESA members may not visit the listserv or might not participate in the survey, the survey garnered only 238 responses, which is less than 3% of the ESA membership. Given the low response rate of the survey (even lower when considering the reach of Twitter), there is potential for substantial non-response bias in the survey responses. Best practices would call for an attempt to estimate the size of the non-response bias and its effect on the survey results.
2. The survey was open for a rather short timeframe, namely 2 weeks. There is no indication that the survey was sent out repeatedly or whether any reminders were sent to the ESA membership (or Twitter re-tweets). Best practices in survey research suggest that repeat contacts with potential survey participants are important for recruitment and for increasing response rates. Short response periods increase the potential for excluding potential participants with characteristics that might make a swift response difficult.

3. While not specified, it appears that the survey may have been conducted anonymously. Despite being an anonymous survey, it would have been informative to collect demographic data on the survey participants. From the results presented, it does not appear as if any other demographic data were collected than their "position". While it was not the goal of this investigation to understand how research practices vary by demographic factors, this information could have been potentially illuminating in understanding any patterns in research practices. This would also have been an avenue for assessing a non-response bias.
4. The investigation is drawing on data published in five major scientific journals with global authorship and readership. However, the listserv survey was accessible only to members of the ESA. While membership in ESA is open to foreign nationals, it may be reasonable to assume that most ESA members are US citizens. Implicit in here is the assumption that the research practices of US researchers are the same as in other countries across the globe. This assumption might be true or not. The survey was also shared via Twitter, which potentially increases its reach across the globe. However, the authors did not report how many responses were gathered via the ESA listserv and via Twitter.
5. The use of Twitter does not allow any control to ascertain that survey participants indeed were ecologists. Next to ESA and instead of the use of Twitter, it might have been better to reach out to other ecological associations across the globe such as the British Ecological Society, the Société Française d'Écologie et d'Évolution, the Ecological Society of Germany, Austria and Switzerland, the European Ecological Federation, the Ecological Society of Australia, and the various other national and international organizations that unite ecologists.
6. The authors collected data on participants' "position" and report how many participants belonged to various groups such as faculty or graduate student. It is not quite clear why this data was collected and reported as it was not used for any other purpose. However, it might have been interesting to investigate whether survey responses varied by "position". It might be expected, or not, that statistical sophistication is increasing with increasing seniority.
7. The authors collected data on participants' "position" including "graduate student". However, the group "graduate student" captures both Masters and PhD students. It might be expected that PhD students show advanced statistical sophistication. But this cannot be confirmed given the presented data.
8. It is unclear how the survey was administered through the listserv and through Twitter. Given this lack of information, it is unclear whether, and if so how, comparability of the two survey modes was achieved.

The above observations put into question whether the authors followed best practices in survey research. Not following these practices has the potential to introduce errors and biases of unknown direction and magnitude; it also makes it very difficult to replicate this part of the investigation (which might be considered ironic since this entire investigation is based on the premise that there is an issue with bias and reliability in ecological research). In my opinion, while the survey results are interesting, it may be best to present them with a much higher degree of caution, or remove them from the manuscript.

**********END**********

Author Rebuttal, first revision:

NATECOLEVOL-220516528A

Below, **the reviewer comments are in bold** and our responses are in regular font.

Summary to All Reviewers

We appreciate the time and attention that the four reviewers invested in reviewing our manuscript and their favorable responses. We are pleased that you thought that our study was interesting, important, and timely for a broad audience of ecologists. You all made constructive comments, which we have taken on board. Here, we summarize the most important revisions:

1. In response to Reviewer 1, we have toned down the assertions regarding the analysis of effect sizes and power, and the implications of this analysis for the credibility of ecological designs. In particular, we completely revised the text in lines 83-144 to acknowledge the reviewer's central point: if ecologists were not interested in detecting and reporting on small effect sizes, despite their ubiquity, then the lack of statistical power to detect the likely small expected effect size in an ecological study would not be a problem. We also revised the abstract to reflect the downweighing of claims about power, and we extended Figure 2 and Figure 3 to include even larger effect sizes. To address the reviewer's concerns about the ambiguity surrounding the implications of multiple hypothesis testing, we re-arranged the text in the multiple hypothesis testing section to make the debate about when corrections should be used more salient, as well as to better describe the implications of multiple hypothesis testing when there is selective reporting.
2. In response to Reviewer 2 who requested more justification for the journals we selected and the analyses we perform (i.e., why not other potentially problematic practices?), we added text to the manuscript. Regarding the five journals from which we take our data, we make our underlying assumption explicit so that a reader make judge its validity (lines 69-71): the empirical studies in these journals are representative of good quality ecological studies in the broader ecological literature. Regarding why we chose to study the practices we chose to study, we now make it clearer in the text (lines 53-55) that we study them because they are important factors that affect the credibility of empirical studies and because they can be empirically detected, which we believe is a missing component in debates about scientific credibility in ecology.
3. In response to Reviewer 3, we have created a metadata file and uploaded it to our Open Science Framework project page, where, upon acceptance of our manuscript, the public will also have access to our data and analysis code: https://osf.io/9yd2b/?view_only=d3e18f3437bf49289cc5448d9e5a2e36 .
4. In response to Reviewer 4's critique of our survey of 238 ecologists about their expectations and practices regarding statistical power, we have removed the survey from the manuscript, as the editor has instructed us to do. We do not include R4's comments below because they are all addressed by removing the survey from the manuscript.

Throughout the revised manuscript, we adopted many of your suggestions for improvements in the exposition (see responses below).

Reviewer #1

This manuscript presents empirical evidence (based on a large sample of the recent ecology literature and a survey of ecologists) with the aim of assessing the degree of bias in the ecology literature. The evidence in this paper is thorough and diverse, and taken together, will be interesting to many ecologists as well as to meta-scientists because it represents an empirical advancement in our understanding of the issue of bias in the ecology literature.

The core evidence presented in this manuscript is a set of many thousands of statistical effects extracted from ecology articles published over 2018, 2019, and part of 2020 in five high profile journals. Based on these effect sizes, the authors calculate an average effect size and then calculate the average power of the studies in this sample to reach nominal statistical significance if the average effect size across studies is assumed to be the true effect size. The authors are explicit in their acknowledgement that they do not view the average effect size as the true effect size. Instead, they present it as a useful benchmark for comparison. I think it is valuable to present this analysis in this manuscript, but as I discuss in more detail in several places below, I think the authors have overstated the utility of this approach.

We are pleased that you believe the analysis is valuable and we are willing to revise the manuscript to make your critiques of our analysis salient in the text.

This average effect size, though credible, is not as useful as the authors argue (because it is probably actually an underestimate of the sizes of the effects that researchers want power to distinguish from zero). So, many of my comments below are related to my request that the authors moderate their statements about the conclusions that can be drawn from these data and acknowledge additional shortcomings of these data. All that said, the manuscript does present evidence that allows us to evaluate the hypothesis that, under a range of plausible conditions, many effects published in ecology are inflated, and this is important and should be of broad interest.

Please see our specific revisions, which are described below.

The authors also present valuable evidence to evaluate the rate of selective reporting in published papers and their supplements, useful evidence regarding rates of data and code archiving, and useful evidence regarding multiple hypothesis testing (and corrections thereof) within individual papers, as well as results of an interesting survey of ecologists regarding statistical power.

I make specific suggestions regarding most of these sections below, organized by line number.

7-8: "We show that most ecology studies are underpowered" – this statement rests on an important assumption that I don't think is sufficiently supported and therefore I think this statement should be moderated somewhat. I explain my concerns about this assumption below.

We respond in more detail below, but we believe that we have addressed this comment through a revision that acknowledges the ambiguity about what inferences readers can draw about the statistical power of ecology studies based on our analysis. More specifically, we re-arranged the analysis in this section in three paragraphs. The first para presents our main analysis. The second para acknowledges the key assumption in our analysis – i.e., that ecologists care about small effect sizes – and makes clear that a violation of that assumption means that we have under-estimated power. We introduce this paragraph with the transition sentence, "Whether our approach yields an accurate approximation of the statistical power of a typical ecology study depends on several assumptions." The third para presents the results from the extended analysis that considers power for effect sizes five times larger than the one that we used in the main analysis. We introduce this final paragraph with the transition sentence, "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 'true effect sizes.'"

8: "that interacts with publication biases" – I think this statement should also be moderated to something like "can interact with" or even "may be likely to interact with"

To moderate the statement, we have adopted your suggestion and revised the text from "interacts with" to "can interact with."

25-26: ideas, hypotheses, models etc. are also important. Maybe you could say something like "an essential component of advancement"

We agree that these other components are important for knowledge generation and did not mean to imply advances only take place through the generation of credible empirical evidence. We think we can accommodate the spirit of your comment without changing the subject of the sentence to "an essential component of advancement," which is a bit challenging to read and no longer agrees with the subject of the introductory clause (disciplines). The revised the first sentence reads: "Like all scientific disciplines, ecology advances, in part, through the generation of credible empirical evidence." That revision also allows us to retain the parallel structure of the subjects in the first two sentences of the paragraph (ecology and ecologists).

26: maybe "and often to inform" (not all ecologists are doing these things)

To improve the flow in this paragraph, we have revised the second sentence. The first sentence ends with the topic of evidence and thus the second sentence should pick up where that sentence left off. It now reads, "Ecologists rely on this empirical evidence in their efforts to understand how the natural world works and to inform policy and management decisions." The revised sentence does not imply that all ecologists do these things and thus we do not think we need to add "often" before "to inform," which would make reading the revised sentence awkward. Note that, Reviewer 2 requested that we delete "predictive models" from the list in this sentence in the original manuscript, and we have done so.

33: references 2-4 are a surprising choice to support your statement about debate about incentives. These references all address the issue of publication bias, but they do not really explore or give contrasting perspectives regarding the incentives that might drive these biases.

We think the problem here may arise from our use of the word "debate" in this sentence. Our intent was to assume that publication bias arises from misaligned incentives between what is good for scientists and what is good for science, and thus any publication that addresses whether publication bias is a concern or not is a relevant citation for this sentence. We have revised the sentence to remove the word "debate" and instead emphasize that these articles raise concerns about the credibility of the empirical evidence base (the subject of our first paragraph).

"Concerns about whether scientists have the correct incentives to generate credible 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."

42: you might consider replacing "imply" with "demonstrate"

We have replaced "imply" with "demonstrate."

52: it seems to me that this paragraph would be best if it addressed just your second aim. The first aim is simply the background info that would normally go into an introduction to justify the study and the third aim is simply expected content of a discussion section on such a topic.

We respectfully disagree. We want not only experts on open science to read this article, but all ecologists. Many ecologists are not aware of the intuition for why the issues we raise are a problem and how they map to the credibility of the evidence base. By emphasizing in the Introduction that our article provides a primer on these ideas, we make it more likely that novices will continue reading rather than assuming the meat of the article will be very technical and thus to be skipped. Regarding our third aim, we believe that whether the topic of "solutions" is

"simply expected content of a discussion section" is debatable. In fact, some readers of our manuscript have encouraged us to delete that section and simply focus on the quantification exercises. We want to provide a signpost to readers that they can expect us to not to simply highlight problems but also to recommend solutions.

76-80: this is an important point that merits more explanation. A naïve reader would benefit from some more detailed explanations of how and why these statements are true

For the reviewer's benefit, we reproduce the two sentences here: "When ecological data are highly variable and sample sizes are small relative to the true effect sizes, the estimated effect sizes are unreliable. This unreliability, when paired with publication biases against replications and against estimates of a particular magnitude, sign, or precision, can lead to exaggerated effect sizes in a literature." These two sentences are part of a short introductory paragraph that aims to summarize the essential ideas of the subsection for readers. More detail on the first sentence could be found in the next paragraph of this section. In our revision, we help give the reader a better idea of the meaning of the sentence in the first paragraph and prepare the reader for more details on this claim in the next paragraph. We now write, "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)." The second sentence in the quote marks above was designed to serve as a signpost for the next subsection, but in re-reading the sentence, we believe that it does not belong in this paragraph because we don't address the issue in this subsection. In this subsection, we address unreliability (variability) of effect sizes. Then, in the next subsection, we describe how that unreliability interacts with publication biases to create exaggeration biases. Thus, we cut the second sentence from this paragraph.

94: your "Thus" statement here is not sufficiently clearly linked to the prior statement. Again, think of the naïve reader

We have revised the text. Here, we reproduce the three sentences that precede the revision, as well as the revision itself that forms the last two sentences: "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."

106: I have long had a concern about this method of estimating the average effect size for the sake of estimating power. Given that some hypotheses are wrong, in other words, the hypothesized relationships are unimportant/trivial/tiny, we don't want power to detect those. Instead, we want high power to detect (distinguish from zero) relationships that are large enough to be meaningful (this threshold between meaningful and not meaningful is obviously both subjective and contextual, but it will often exist). So, if an estimate of the average effect size we want to detect is based in part on many effects that are trivial and which we therefore do not want to detect (as yours surely is), we will dramatically underestimate our power. (see below for two further discussions of the relevance of hypothesis testing vs estimation)

That said, I appreciate that you also presented analyses based on other (more plausible) potential target effect sizes (though I would like to see presentation of wider range [extending larger than 0.2]). I would further appreciate an explicit de-emphasis of your calculated mean effect size (explained to the reader in the context of the reasoning I just presented above), and a corresponding additional emphasis on larger potential target effect sizes.

(I want to acknowledge that in some cases, ecologists will want a precise estimate of a very small effect [but see more on this below]. However, my sense is that most ecologists are happy to consider very small effect sizes as approximately zero, and thus do not require precise estimates of very small effects. I also want to point out that, unfortunately, many researchers in ecology remain more focused on dichotomous hypothesis testing than on estimating sizes of effects)

The target effect sizes – i.e., the magnitudes of effect sizes that are ecologically relevant to researchers – are unknown. Yet, you raise an important point that needs to be acknowledged in the subsection (see next comment and response).

107: for reasons stated above, I think you probably actually underestimate power (despite your confident statement of overestimation).

As noted above, we have reorganized and revised this subsection. We now explicitly acknowledge that ecologists may not be interested in distinguishing small effect sizes from zero and, if that were true, we have under-estimated power in our analysis. That text is on the next page in quotes. We have also revised the following sentence (now lines 116-117): "Rather, the approach offers an approximation of the magnitude of the true effect size that a typical ecological study seeks to estimate" to read "Rather, the approach offers an approximation of the magnitude of the true effect size that a typical ecological study would expect to find." We have also extended the range of effect sizes to a PCC of 0.30, which is considered large by modern

analysts (e.g., Doucouliagos 2011, How Large is Large…") [Note also that we are presenting the results in terms of PCCs rather than fractions of SDs, which is the typical units for applying labels like "small," "medium", or "large"]

125, also Fig 1: see comments above

As noted above, we have extended the range of Fig 1 to PCC=0.30.

129: effects of 0.2 are often described as moderate or between small and moderate (usually not large). Of course this is all relative to expectation, but effects can certainly be much larger than 0.2 in ecology studies.

We hope that the revisions to our text and figure that we described above address this issue. As an aside, we note that the adjectives being applied to effect sizes are changing since the classic Cohen benchmarks from many decades ago (which were in SD units not PCC units). As pre-registered studies become more common in some disciplines, expectations about effect sizes in those disciplines also change, and thus so will the categories to which adjectives like "small" and "moderate" apply.

130-135: these arguments rest on the assumption that you have used accurate estimates of effect sizes that researchers want to distinguish from zero. I want to be clear that I am not saying I think that you are wrong, only that I don't think we understand as much about average effect size as you assume.

We now try to make your point explicit in our text. Recall that the first paragraph in this subsection focuses on the big picture about information that is available in an empirical design about a target parameter's value. The second paragraph defines power and provides some more intuition about how it connects to the topic in the first paragraph. The third paragraph describes the methods we use. Following that paragraph, we reordered the text and added new text. The revised fourth paragraph presents our first result using the PCC from the weighted average of PCCs in the studies in our data set. Then the fifth paragraph transitions to your critique and the last paragraph extends the analysis to larger effect sizes. Thus, we end the subsection with what we believe is your main critique: that our analysis may under-estimate power and that if ecologists were to only be interested in much larger effect sizes, most designs would be considered adequately powered. The reader can then make their own conclusion based on our presentation of a range of effect sizes.

"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 effects." 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 Fig S1 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)."

148-9: and authors!

We have added this additional text.

163: instead of "only the large effects", maybe you should say something like "mostly the large effects" or "disproportionately the large effects" (or something to that effect)

We have revised the text as you suggested: "However, where there is publication bias against results that do not pass conventional thresholds of statistical significance or have unexpected signs[5–7], mostly the large effect sizes with expected signs end up being published."

171-174, Fig 2: These estimates are again based on the assumption that our goal should be power to detect or precisely estimate even trivially small effects. The range of WAAP values you selected to display in Fig 2B is skewed towards small effects, and the calculated value of WAAP is probably an underestimate given that it presumable incorporates many trivial effects. Similarly, the values shown in Fig 2a are also based on a calculation that incorporates trivial effects (into the estimate of the 'true' effect size).

As we did for Figure 1, we have extended Figure 2 to include values as high as 0.30. Readers can now draw their own conclusions.

178: some ecology is applied, but it is not universally an applied field. Maybe instead say something like "In a field where results often have application" or "In a field where results often have real-world application" or "In a field where results often have management implications"

We have revised the sentence as you suggested: "In a field where results often have real-world applications…"

178-182: I very much agree with this statement that magnitude matters and that a major problem with publication bias is the bias in magnitude of published effects. However, much of your empirical work in this paper misrepresents the extent of the problem to some unknown amount because you focus on power, and 'power' is not primarily about magnitude, it is about crossing a threshold of 'statistical significance'. As true effect sizes get smaller and smaller, we need greater and greater power to distinguish these effect sizes from zero. However, as they get smaller and smaller, distinguishing these effect sizes from zero may often become less and less important to the researcher. In other words, if we focus on SE relative to the size of the estimate, we will need to make the SE smaller and smaller as the effect size gets smaller to make sure we cross the significance threshold. However, it may instead be the case that we want a certain amount of confidence in a point estimate (a certain absolute SE) regardless of the distance of the estimate from zero. In other words, we may sometimes want the same degree of absolute precision in our estimates regardless of the magnitude of the effect. If that is the case, then focusing on power to distinguish tiny/trivial effects from zero (as your study currently does inadvertently), is misguided.

We hope that our revisions to the text, describe above, have mitigated your concerns.

I want to emphasize here that I like this study and I don't want to see very many substantial changes to your methods. Instead, I want to see more acknowledgement and discussion of the limits of your inference.

Thank you.

197: Amrhein (citation 30) is also an ecologist, but maybe you mean that the paper was published in a general science journal, in which case, this is fine

We were using the Amrhein citation as an example of a paper published for more of a general audience rather than one for ecologists explicitly.

220: I think "necessarily" and "not" are reversed

We have fixed this typo.

262: This section of the study is the one I am least excited about. There is legitimate debate about the circumstances that require correcting for multiple hypothesis tests and the methods that should be used for these corrections.

We agree with your concerns about multiple hypothesis testing. See our detailed replies below.

267-268: You state that "Even in the absence of selective reporting, multiple hypothesis testing without statistical corrections can inflate the prevalence of spurious results reported in the ecology literature". I'm not sure what you mean here. If you mean that the absolute number of spurious results in the literature will increase as the number of tests within individual studies increases, then I agree. However, I do not see the problem with that. As the number of different studies conducted on a topic increases, the number of spurious results will increase also, but this is not a bad thing. That is because, as the number of different studies increases, the number of effects contributing to our meta-analytic mean also increases, and so the precision of our mean increases. More effects will mean better precision of average effects whether those different effects come from the same study or from a different study. Thus, more effects calculated = better. Instead, on the scale of the literature at large, multiple hypothesis testing is primarily a problem if it is combined with selective reporting or other deceptive practices.

We agree with the reviewer, but also note that implicit in this argument is a professional context that rewards and encourages replications.

On the scale of an individual study, multiple hypothesis testing is a problem for interpretation of the paper in question if the authors do not acknowledge the increase in the likelihood of false positives or strong relationships due to chance. This is clearly important when considering studies in isolation, but the much greater concern to me is multiple testing combined with some form of publication or reporting bias since that biases that literature at large.

We agree with your concerns here about multiple hypothesis testing. We have removed the beginning clause of the sentence you highlight. Further, we have now added more text to be explicit about when multiple hypothesis testing can be misleading - when there is selective reporting of 'interesting' results and when authors do not report on all the statistical tests done in the current study such that reviewers and subsequently readers cannot make informed decisions about the interpretation of the paper. See lines 262-292.

273-274: this is an example of a debatable statement on this topic (and it is not backed up by citation). Some authors argue (and I apologize for not going back and finding any

citations for you here) that corrections for multiple comparisons are only necessary when testing different predictions of the same hypothesis.

We believe that you are saying that there are debates about when to make corrections for multiple comparisons and thus while our example of multiple hypothesis testing is an example of multiple comparisons, it may not an example where corrections for multiple tests are necessary, given the ongoing debates about when to apply these corrections. If we have correctly characterized your concern, we agree and have cut the sentence you highlighted. The revised first paragraph focuses on the interaction of multiple comparisons with publication biases. The second paragraph highlights that there are ways for ecologists to present all their hypotheses and adjust their inferences when multiple hypotheses are tested, but the paragraph also emphasizes that there is an ongoing debate about when exactly one needs to do these corrections, as well as other challenges. We also bookend our results with the statement, "While correcting for multiple tests may not always be necessary (e.g., [40,42,46]), reporting why corrections were or were not used is necessary for readers to make judgements about the credibility of the analyses." That's our punchline – with few studies using corrections (adjustments) and no studies with multiple tests justifying why no correction (adjustment) was applied, judging the reliability of the results is made more challenging.

279: OK, good. Thanks for including this. However, this debate is not reflected in your earlier statements.

As noted above, we believe our revision addresses this concern.

308: you may wish to add this recent citation on data archiving:
https://doi.org/10.1098/rspb.2021.2780

Thank you for this suggestion. We have added it to the citations about archiving.

Supplement

56-59: please provide version numbers for packages

We added the version numbers.

170: I am sympathetic to the argument you are making here, and I suspect that you are correct that many ecologists are not adjusting their sampling effort in accordance with anticipated effect size. However, I'm confident that such adjustments happen (I've seen them happen in some ecological disciplines), and so the real question should be, how often

do they happen. I do not think we know the answer to this question, and so I think you should express a bit less certainty in your argument here. In other words, your statement on line 174-175 is overly confident.

We appreciate your concern. We thus have removed all sentences in this section of the SM that contain opinions or speculations about the plausibility of the assumptions described in this section. We instead simply state the assumptions and let the reader assess their plausibility. We believe that, with the revisions to the main text where we detail other assumptions, we have provided readers with the relevant information to make their own judgements and have removed our own opinions.

230: could you please provide more information about the survey? For instance, what introductory text preceded the questions you present here? What sample of ecological tests were they told to hypothesize about?

Another reviewer and the editor asked that the survey be removed from our study. Because the survey was not a critical element of our analysis, we have removed the text about the survey from the main paper and supplement.

337: please include a citation in this paragraph

The text in this paragraph is our opinion. Thus, we do not include a citation. We reproduce it here: "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."

444: where does this "may not be a top criterion" come from? Is there evidence of this in the Nosek paper? This is certainly a top criterion at the hiring stage for many (most?) academic research institutions

We cut this sentence in our revision. We were trying to address a critique from another reviewer of an earlier version of the manuscript that not all researchers work in academia. In re-reading the SM, we do not believe that this sentence is critical to making our point and thus we deleted it.

Reviewer #2 (Remarks to the Author):

The manuscript deals with an important topic in ecology (and other sciences) - that of credibility and reliability, which has been on the radar for the past years. Generally, methods seem sound (though see my comments below) and findings are in line with previous work, calling for urgent action. I congratulate the authors on all the work conducted. I provide my comments below, structured following as Reviewers' guidelines.

We are glad that you enjoyed the paper. We provide detailed responses to your comments below.

- Key results: Please summarise what you consider to be the outstanding features of the work.

This work estimates the prevalence of several suboptimal practices in ecology based on publications from five popular journals. Results correspond to some previous estimates for ecology and definitely support calls for action on these issues. The dataset encompasses a large and comprehensive set of articles published in 5 journals.

- Validity: Does the manuscript have flaws which should prohibit its publication? If so, please provide details.

I find that MS is valid, with no major flaws, but I do have some issues with how the topic was presented, as I note below (Data & methods)

- Originality and significance: If the conclusions are not original, please provide relevant references. On a more subjective note, do you feel that the results presented are of immediate interest to many people in your own discipline, and/or to people from several disciplines?

I am unclear as to what the overlap between data collected here and those collected in previous studies on these topics is. Good references can be found in the supplementary materials of Purgar et al. 2022, https://doi.org/10.5281/ zenodo.6566100) - They list several studies that estimate power in ecology and publication bias (these studies are based on a larger corpus of primary studies) and also a study by Culina et al on reproducibility.
I think that the results are of wide interest, however with some adjustments (as I explain in my comments).

- Data & methodology: Please comment on the validity of the approach, quality of the data and quality of presentation. Please note that we expect our reviewers to review all data,

including any extended data and supplementary information. Is the reporting of data and methodology sufficiently detailed and transparent to enable reproducing the results?

The methodology seems generally good, however, I have several comments.

My main comment is about the conceptual approach: It is unclear why were these particular practices (power, selective reporting, and multiple hypothesis testing) chosen, among all others that are known to happen (and could also be assessed). There are many other practices that contribute to the low reliability of scientific results. E.g. use of flowed data collection designs? While there is a very limited explanation on why (line 47 'widespread') research shows that other suboptimal practices are equally widespread. Even in the specification of research aims, it is stated that 'degree to which ecologists use empirical designs ….' [line 55] but later the analysis concentrates only on power, which is only one of the aspects of the empirical design.

We chose these practices because they can be detected empirically directly from data in empirical papers. We now include a sentence to make this clear [bolded]: "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." The practices we quantify are also inter-related, which allows us to build the narrative of the manuscript. Moreover, we have not read a publication that includes the quantification of these suboptimal practices for a broad set of ecology publications.

Second, I would argue that while the first three practices are suboptimal (questionable) and can lead to unreliable results, the fourth practice (availability of codes and data) is of a different kind as it does not lead to unreliable but to unverifiable evidence, and also to the lower impact of work. My suggestion would thus be to make a distinction between low power, selective reporting, and multiple hypothesis testing vs availability of data and code. Indeed, the latter allow for identifying of the QRPs, and their correction (as recognized by authors in Table 1).

We have now moved our discussion of data and code availability to the "Creating a culture that fosters greater credibility in empirical ecology" section in our manuscript because, like you point out, these actions are more about identifying QRPs than a QRP in themselves. Therefore, we also remove these actions in our introduction text from our list of QRPs that we are quantifying.

Other comments:
1. Why were these specific 5 journals selected? Can the results be representative for other journals publishing ecological research? While I do not think that other journals have to be

considered, some discussion on the implication of this particular selection on the results should be there.

We now make explicit both our reasoning for selecting these journals and the underlying assumption in our analysis. Specifically, we state (lines 69-71): "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."

2. Given that only articles that represent data in tables were considered, can this introduce any bias in the results (e.g. I know that results of particular types of analysis are more likely to be presented in a table format).

We do not believe that choosing data in tables would create bias towards higher or lower powered designs. Few studies do not include key results in table format in either the main text or supplement. We now make our assumption explicit in the text (lines 73-75): "Because most empirical studies report estimates in tables in the main or supplemental text, we assume that only including studies with estimates presented in tables does bias our results."

3. It is unclear how were articles scored as appropriate for the research, i.e. being empirical (reading abstract? Or full texts? How many people independently scored these to avoid misclassification?). How many articles passed this stage?

Two people looked at each article (between Dr. Kimmel and six trained research assistants). Each person looked at the full text to make sure that it was empirical. We now add text explicitly stating this in the methods section of the supplemental material: "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."

In our revised supplemental text, we also report the exclusion criteria so that readers know how many articles were excluded for each criterion. "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 work. 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."

4. Were data on statistical tests extracted manually? How many articles did have results reported in tables?

Paper_trackerFeb21.csv provides all the data on how many papers had results reported in tables that fit our requirement. All data were extracted manually and then double checked for accuracy by at least one other individual.

5. The final sample size differs from the main MS (353) and the supplement (366-29 dropped as of no sample size) articles. Maybe Im missing something.

The sample size reported in the main manuscript was our final sample size. The sample size reported in the SM was before papers were dropped when cleaning data. To ensure that readers understand this data processing step, we have now included a detailed exclusion criterion section in the SM.

6. How does the assumption that there is no selective reporting or publication bias against small effect sizes (109-110) affect the reliability of the estimates? Isn't this assumption against what is known (and even demonstrated in this MS).

In the case that small effect sizes were selectively not reported, the benchmark effect size that we calculate would be inflated and thus our results would be a lower bound on the proportion of underpowered tests (i.e., the number of tests that are underpowered would be underestimated).

For example, let us say our minimum detectable effect (MDE) was a PCC of 0.01 from the published results. If we assume that there is a publication bias against small effect sizes, the 'real' MDE may be closer to 0.005. Therefore, when we divide the inflated MDE by 2.8 we get ~ 0.00357 and the 'real' MDE by 2.8 we get ~ 0.00179. Thus, the number we are comparing to is much larger and would thus classify more estimates as adequately powered than if the 'real' MDE was used.

Unfortunately, we do not know the extent of the bias against small effect sizes. Therefore, we use the calculated MDE for our analyses and present a range of other plausible effect sizes, both smaller and larger than our weighted estimate.

7. Why was the survey only conducted regarding the power, and not to estimate other practices that this MS deals with?

Another reviewer and the editor requested that the survey be removed from our study. Because the survey was not a critical element of our analysis, we have removed the text on the survey from the main text and supplement.

8. I have checked the data and codes but have not run the analyses. Data and codes seem complete. They lack Readme or similar files to aid the understanding of what files represent, and what variable names mean.

We apologize for that important omission. We have uploaded a Readme file to both the NEE manuscript system and our OSF page that also includes the code and data: https://osf.io/9yd2b/?view_only=d3e18f3437bf49289cc5448d9e5a2e36 .

- Conclusions: Do you find that the conclusions and data interpretation are robust, valid and reliable?

They generally are, but please see my other comments.

- References: Does this manuscript reference previous literature appropriately? If not, what references should be included or excluded?

I find some important references lacking, and think this work would benefit from considering these (and their findings)
1. Smalidno and McElreath (2016): Natural selection of bad science.
https://royalsocietypublishing.org/doi/10.1098/rsos.160384
Now added as reference 70.
2. Fostemier et al. (2017): Detecting and avoiding likely false-positive findings – a practical guide. https://onlinelibrary.wiley.com/doi/10.1111/brv.12315
Now added as reference 49.
3. Ulrich, R. & Miller, J. (2020). Meta-Research: Questionable research practices may have little effect on replicability. eLife 9: e58237
4. Lakens (2021): Sample size justification. https://psyarxiv.com/9d3yf/
5. on the prevalence of QRP in the Netherlands
https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0263023#sec008
Now added as reference 35.
6. On the prevalence of responsible research practices (to inform solutions)
https://f1000research.com/articles/11-471
Now added as reference 63.

Thank you for these suggestions. Given the journal limits, we do not have room in the manuscript to include them all along with accompanying text to describe their implications in the context of our study. Reading through them, we believe that the Smalidno & McElreath,

Forstmeier et al, and both Gopalakrishna et al references are the most critical to cite, and we now do so. We appreciate you bringing these publications to our attention.

7. reference for line 74-76 is missing.

We do not believe we need a reference for this line.

- Clarity and context: Is the abstract clear, accessible? Are abstract, introduction and conclusions appropriate?

Generally, the paper is well written and accessible. I would suggest that both the title and the abstract are more clear about what is studied in this work: some of the suboptimal research practices (rather than all of them). The abstract could also provide the values for the estimates.

We revised the abstract to be clearer about what we study. We believe that the title and revised abstract avoid implying that we are looking at every suboptimal research practice. We do not include the values for the estimates in the abstract because of word limit constraints, and because our text presents a range of estimates, allowing readers to come to their own conclusions about the magnitudes of the problems we identify. Reviewer 1, in particular, wanted more explicit acknowledgement of uncertainty in our estimates, rather than emphasizing a single estimate.

Some suggestions:

1) line 27 predictive models are part of understanding how the world works.

We removed the phrase "and predictive models." To improve the flow in this paragraph, we also revised the sentence. The first sentence ends with the topic of evidence and thus the second sentence should pick up where that sentence left off. It now reads, "Ecologists rely on this empirical evidence in their efforts to understand how the natural world works and to inform policy and management decisions."

2) line 29 please provide a reference (if any) for this statement. Also think this sentence needs rewording e.g. 'if the estimates of effect sizes are inaccurate, biased, and presented without uncertainty).

There is no reference for this statement because this is a hypothetical example of the implications of not having credible empirical evidence. The suggested revision does not match what we're trying to communicate in this sentence. If the wording issue identified by the

reviewer lies in what we mean by accurate estimation of "the uncertainty about these estimates," we could drop that phrase without substantially altering our point.

3) line 48/49 – this is unclear to me.

The full sentence is: "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[8–10]." We are unsure what you think is unclear about the sentence. The sentence was included because prior readers of the manuscript thought that, given the motivation in the previous paragraphs, a reader may expect us to define what we mean by "credibility" or "replicability" – a debate we do not wish to enter. Thus, we added a "signpost" to ensure that the reader's expectations are aligned with what we deliver in the rest of the paper.

4) line 55 – please highlight that this is only about power, rather than about the other aspects of the empirical design

We have now added a clause to acknowledge that we are focusing on power. The sentence now reads (with the new clause in bold here): "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,…"

5) line 63 – think academic/research institutions should also be mentioned here

We agree and have added "research institutions".

6) main MS would benefit from more information on how the data were collected [lines 68-69]

We now include details on data collection that states (lines 71-78): "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 empirical studies report estimates in tables in the main or supplemental text, we assume that only including studies with estimates presented in tables does 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."

7) the first paragraph of the Underpowered design section could do better work in stating why power is such a problem in ecology (please see some refs that I have provided

previously). Currently, it just states what components will determine the power of the test, rather than what we know about these components (e.g. that most of the effects in ecology are small).

We believe one has to start this subsection with a definition of power rather than "what we know about these components."

8) lines 81-83 are unclear to me.

The complete sentence here is: "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." In the Bayesian framework, one would explore the reliability of estimated effect sizes through a different lens. This sentence is simply to justify, to our Bayesian colleagues, why we focus on statistical power rather than other forms of assessing the "information content" of an empirical design.

9) lines 93-94 do not provide an explanation on why rejecting a null hypothesis will be unlikely in this case.

We have updated the text to read (lines 101-103): "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."

10) lines 147-149 sentence is a bit biased itself – most of the QRPs when coupled with publication bias can lead to exaggeration bias.

We did not mean to suggest that underpowered designs are the only practice that, when coupled with publication bias, can lead to exaggeration bias. We make clear in the abstract and introduction that the other practices can also lead to (exacerbate) exaggeration bias.

11) lines 184-188 are unclear to me.

We now clarify that we are asserting that there exists an exaggeration bias in published results by including a sentence between the two we originally wrote (in bold): "Based on our empirical results, we are not asserting that most of the ecological relationships reported in the literature are likely to be spurious – in fact, we doubt ecologists are studying relationships for which the sharp null hypothesis of zero effect is widely true. **Instead, we are asserting that the magnitude of these relationships is inflated.** In other words, we are asserting that we have indirect empirical evidence - "fingerprints", if you will - that the published effect sizes in ecology journals exaggerate the importance of many ecological relationships."

12) if space needs to be spared I would suggest removing paragraphs 228-247. This is already mentioned in the next paragraph and can be just a bit elaborated on.

We believe this text is relevant to understanding the text that follows in the subsequent paragraph.

13) line 313 should be 'reproducibility' instead of replication.

We have changed from 'replication' to 'reproducibility.'

14) not mentioned how many papers provide both code and data (lines 319/320)

We have now updated the sentence to state the percentage of papers provide both data and code (lines 325-326): "In that study, 79% of studies provided data, 27% provided code, and 21% had both data and code[11]" Note that the subsection on data and code availability has been integrated into the Discussion section entitled "Creating a culture that fosters greater credibility in empirical ecology."

- Please indicate any particular part of the manuscript, data, or analyses that you feel is outside the scope of your expertise, or that you were unable to assess fully.

I am not an expert in power analysis so could not judge that analysis nor what is represented in the paragraph lines 97-110. Especially I am unclear about the validity of the assumption that there is no selective reporting or publication bias against small effect sizes. Isn't this assumption against what is known (and even demonstrated in this MS).
I have also never conducted a survey thus not competent to judge the general approach.

In the case that small effect sizes were selectively not reported, the benchmark effect size that we calculate would be inflated and thus our results would be a lower bound on the proportion of underpowered tests (i.e., the number of tests that are underpowered would be underestimated).

Reviewer #3 (Remarks to the Author):

I was pleased to review NATECOLEVOL-220516528 "Reliability of empirical evidence in ecology and a proposal for action". The study is on an important and timely topic, and is well suited for Nature Ecology & Evolution. The authors carried out a considerable amount of work and the analyses appear sensible and rigorous. However, I was unable to carefully check the data and analysis code as there was no metadata provided alongside the data (i.e., it is challenging to make sense of what each data file is and how it fits into the broader analysis without running the script). The file names for the csv files are not particularly informative. Otherwise, the manuscript is clear, neatly structured, and well written. The supplementary material is informative. I have only minor comments and commend the authors on their excellent work.

We thank you for acknowledging the time and rigor of our study. We have uploaded a Readme file to both the NEE manuscript system and our OSF page that also includes the code and data: https://osf.io/9yd2b/?view_only=d3e18f3437bf49289cc5448d9e5a2e36

Your other minor comments are addressed below

Main text

Please include metadata (i.e., a readme file) describing the content of each data file, what the column headings are, abbreviations, and units. Column headings should not contain units, values, etc – this information would be best placed in the readme file. For useful resources on this topic, see refs in Table 3 under "Open data" here: https://doi.org/10.1242/jeb.243559 - and (Towse et al. 2021; Contaxis et al. 2022). These resources are also relevant in the context of the manuscript's Table 1.

We have uploaded a Readme file to both the NEE manuscript system and our OSF page that also includes the code and data:
https://osf.io/9yd2b/?view_only=d3e18f3437bf49289cc5448d9e5a2e36

L35. It might be worthwhile clarifying (here or elsewhere) that the authors consider p-hacking as a component of selective reporting. I was surprised that there was no mention of p-hacking in the manuscript.

We originally did not use the term "p-hacking" because it has a pejorative connotation: it connotes the practice of deliberately altering statistical tests and thus may make some readers less receptive to the key points we are trying to make. However, we acknowledge that p-hacking is a common term used in metascience research. Thus, in our revision of the selective reporting section of the manuscript, we now clarify that we include p-hacking as part of selective

reporting: "Because of publication biases in favor of statistically significant results[4,7,12], researchers may seek to find and publish such results over those that are statistically insignificant[13]. 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 significant[14,15]. Such choices are more likely when they can transform "marginally nonsignificant" results into significant results (e.g., "*p*-hacking")."

L94-95. This sentence does not appear to logically flow from the previous two IMO. Please consider rephrasing?

We have updated the text to read: "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."

L99-100. Consider providing a more detailed explanation of the PCC as not all readers will be familiar with this statistical concept.

We have added a definition of the partial correlation coefficient: "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."

L137. It is unclear from the SM that the respondents were asked specifically about the authors' dataset. Perhaps clarify this in the SM?

Another reviewer and the editor requested that the survey be removed from our study. Because the survey was not a critical element of our analysis, we have removed the text on the survey from the main text and supplement.

L138-140. I was slightly confused by this sentence given that the authors only asked experimentalists how often they carry out power analyses (see SM L234-236). Consider rephrasing?

As noted above, we removed the survey from our manuscript based on another reviewer and the editor's request.

L200. Consider replacing insignificant with nonsignificant.

We have made the change to "nonsignificant."

L290-291. It is unclear from the text if the authors evaluated whether studies reported why corrections were or were not used.

From our reading of the papers that did not use a correction, none mentioned multiple hypothesis corrections, let alone stated a reason for the absence of corrections.

L299– Data and Code Availability. It might be worth mentioning in this section that most publicly available datasets in E&E are unusable – i.e. open data are not necessarily FAIR data (Wilkinson et al. 2016; Roche et al. 2022) – see also https://www.go-fair.org/fair-principles/ and Box 2 in https://doi.org/10.1242/jeb.243559.

We have now added a sentence on the quality of publicly available data and included the suggested citations (lines 319-321): "Therefore, availability does not equate to quality of data or code[53]; most ecology and evolution publicly available datasets in a recent analysis where not reusable and only slightly over half were even complete[16]."

L305. Ecological Society of America (not American)

We have fixed this typo.

L317-320. There is no need to cite this paper but the authors might be interested in recent data for comparative physiology and behavioural ecology journals, which colleagues and I have reported here (see Table 1 and Fig 2): https://doi.org/10.1242/jeb.243559

We thank you for sharing this publication. We have added it in support of this statement (lines 314-316): "Yet, despite these attempts to make data and code more accessible (e.g., [63]), obtaining data and code can still be challenging[61,64–69]."

L332. Again, no need to cite but perhaps of interest: Buxton et al (2021) Avoiding wasted research resources in conservation science. Conservation Science and Practice 3: e329 https://doi.org/10.1111/csp2.329. See section 5 " Openly and comprehensively report research outputs" and Table S1.

We have now included this citation to support this statement (lines 303-305): "A few publications describe some of these actions and some of the challenges to scaling these actions in the context of ecology[15,22–26]."

L340. Consider using "positive results" instead of "statistically significant results".

"Positive results" could be interpreted by a reader as the direction of the effect instead of statistical significance and thus we have decided not to incorporate your suggested revision.

L346. I was pleased to see SORTEE mentioned here. If there is space, DORA might also be worth highlighting (https://sfdora.org/) – although it is not specific to E&E.

We are unable to specifically mention DORA in the main text of the manuscript because of word limits. However, we now highlight DORA in Table 1.

L355. Missing 'to'.

We have fixed this typo.

Table 1. Parker et al 2019 mention results-blind review but only in the context of registered reports, hence it might not be an appropriate reference here. Again, no need to cite but perhaps of interest – we list the potential costs and benefits of sharing open data in (Roche et al. 2014).

We have removed the Parker et al 2019 reference from the results-blind review section.

Supplementary Material

L19. Were 1,568 papers examined in total? Slightly unclear…

1,568 papers were examined in total. We excluded papers that did not meet our criteria. We now include the exclusion criterion in the supplemental text so that readers know how many articles were excluded for each reason.

"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 work. 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."

L21. Perhaps consider explaining why only in tables and not in the text?

We standardized our protocol before looking at papers. We do not think this protocol created any bias and wanted to make sure we had a robust sample size where effects and errors were easily identified by the research assistants. We now include text to explain our rationale: "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." In the main text, we now also make explicit an assumption that accompanies this decision: "Because most empirical studies report estimates in tables in the main or supplemental text, we assume that only including studies with estimates presented in tables does bias our results."

L47. "We tried not to include robustness checks" – how was this done and why was it challenging?

We erred on the side of caution when we could not determine if a test was part of the main analysis or a robustness check and did not include it as a multiple hypothesis test. We now

include our rationale for determination in the text: "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."

L48. Were these exact words searched or did you employ stemming to broaden the search (https://libguides.mit.edu/c.php?g=175963&p=1158679)? For example, I would have expected "correct*" to be used as a keyword.

These were the exact words we used, as we followed the methods in Ferraro and Shukla 2020. We assumed that authors would not mention multiple hypothesis corrections without stating the type of correction done.

L68. 2 x 'the'

We have corrected this typo.

L164-165. It is unclear from this sentence why the costs of data collection or selecting study units are large when the expected treatment effect sizes are large.

We also do not believe that this statement is true – we were using it as an example of what one would have to think in order to interpret our results in a different way. In other words, we are elaborating on what one would have to believe in order to interpret our results in a different way.

L266. Reference to SM in the SM.

We have removed the reference to the SM and included a reference to the specific section of the SM in its place.

L294-296. The American Naturalist and the Journal of Evolutionary Biology now have data editors: see https://jevbio.net/data-editing-at-jeb/ and http://comments.amnat.org/2021/01/note-since-fall-2020-robert-montgomerie.html

We have now included these as examples as well.

L329-330: Consider explaining why science suffers when exploratory analyses are repackaged in publications as confirmatory analyses. This might not be obvious to many readers.

We have added text to explain this concept further. "Indeed, these repackages exploratory analyses never have the chance to be falsified and may need complex hypothesis to accommodate the results[32]."

L408. Could you provide this list of 9 journals in a table?

Yes - we have now included a table (Table S1) with the journals offering Registered Report format.

nature portfolio

Decision Letter, second revision:

24th April 2023

Dear Dr. Ferraro,

Thank you for submitting your revised manuscript "Reliability of empirical evidence in ecology and a proposal for action" (NATECOLEVOL-220516528B). It has now been seen again by the original reviewers and their comments are below. The reviewers find that the paper has improved in revision, and therefore we'll be happy in principle to publish it in Nature Ecology & Evolution, pending minor revisions to satisfy the reviewers' final requests and to comply with our editorial and formatting guidelines.

If the current version of your manuscript is in a PDF format, please email us a copy of the file in an editable format (Microsoft Word or LaTex)-- we can not proceed with PDFs at this stage.

We are now performing detailed checks on your paper and will send you a checklist detailing our editorial and formatting requirements in about a week. Please do not upload the final materials and make any revisions until you receive this additional information from us.

Thank you again for your interest in Nature Ecology & Evolution. Please do not hesitate to contact me if you have any questions.

[REDACTED]

Reviewer #1 (Remarks to the Author):

I am satisfied with the revisions made by the authors. I have no further comments.

Reviewer #2 (Remarks to the Author):

The authors have thoroughly addressed the reviewers comments and resulting MS is much clearer and better reasoned. I have a few remaining comments, but I would leave the authors to decide whether they want to incorporate them, as opinions on some matters might just differ.

1) Line 55 'We focus on practices that can be empirically detected via analyses of published articles.' I guess other practices can be empirically detected too, but maybe with less precise estimates as of the low completeness of reporting of e.g. methods. So maybe add ' relatively reliably/easily' or similar

2) Lines 70/71 ' 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'

I am unsure about this. High impact journal do tend to attract different type of studies than lower impact journals, as well as might invite for more of some suboptimal practices as it is more difficult to get a study published if results are 'not interesting' . E.g. see doi: https://doi.org/10.1101/311068

3) 'lines 73-75: "Because most empirical studies report estimates in tables in the main or supplemental text, we assume that only including studies with estimates presented in tables does bias our results."

Maybe I misunderstand the first sentences, but about 1568 papers originally selected, 1038 do not report statistical test in tables. So, the opposite seems to be true – most articles do not report results in tables.

Also, I think you wanted to write 'does NOT bias our results'. While we do not have any evidence of differences between studies that do and don't report their results in tables, I would still not be sure that the difference does not exist.

4) Lines 84/85 'The amount of information that ecologists can extract from their data depends on the variability …' 'Information' is quite a vague term, and the sentence is more about the reliability of information (as implied in the next sentence of the paragraph).

5) Seems that quite many estimates had to be excluded as sample size was not possible to determine. Maye this observation is something to briefly mention in the Discussion, saying that this 'side ' result amplifies the need for better reporting.

Reviewer #3 (Remarks to the Author):

The authors have carefully addressed the reviewer comments. I have no further recommendations.

Our ref: NATECOLEVOL-220516528B

18th May 2023

Dear Dr. Ferraro,

Thank you for your patience as we've prepared the guidelines for final submission of your Nature Ecology & Evolution manuscript, "Reliability of empirical evidence in ecology and a proposal for action" (NATECOLEVOL-220516528B). Please carefully follow the step-by-step instructions provided in the attached file, and add a response in each row of the table to indicate the changes that you have made. Please also check and comment on any additional marked-up edits we have proposed within the text. Ensuring that each point is addressed will help to ensure that your revised manuscript can be swiftly handed over to our production team.

We would like to start working on your revised paper, with all of the requested files and forms, as soon as possible (preferably within two weeks). Please get in contact with us immediately if you anticipate it taking more than two weeks to submit these revised files.

When you upload your final materials, please include a point-by-point response to any remaining reviewer comments.

If you have not done so already, please alert us to any related manuscripts from your group that are under consideration or in press at other journals, or are being written up for submission to other journals (see: https://www.nature.com/nature-research/editorial-policies/plagiarism#policy-on-duplicate-publication for details).

In recognition of the time and expertise our reviewers provide to Nature Ecology & Evolution's editorial process, we would like to formally acknowledge their contribution to the external peer review of your manuscript entitled "Reliability of empirical evidence in ecology and a proposal for action". For those reviewers who give their assent, we will be publishing their names alongside the published article.

Nature Ecology & Evolution offers a Transparent Peer Review option for new original research manuscripts submitted after December 1st, 2019. As part of this initiative, we encourage our authors to support increased transparency into the peer review process by agreeing to have the reviewer comments, author rebuttal letters, and editorial decision letters published as a Supplementary item. When you submit your final files please clearly state in your cover letter whether or not you would like to participate in this initiative. Please note that failure to state your preference will result in delays in accepting your manuscript for publication.

Cover suggestions

As you prepare your final files we encourage you to consider whether you have any images or illustrations that may be appropriate for use on the cover of Nature Ecology & Evolution.

Covers should be both aesthetically appealing and scientifically relevant, and should be supplied at the best quality available. Due to the prominence of these images, we do not generally select images featuring faces, children, text, graphs, schematic drawings, or collages on our covers.

We accept TIFF, JPEG, PNG or PSD file formats (a layered PSD file would be ideal), and the image should be at least 300ppi resolution (preferably 600-1200 ppi), in CMYK colour mode.

If your image is selected, we may also use it on the journal website as a banner image, and may need to make artistic alterations to fit our journal style.

Please submit your suggestions, clearly labeled, along with your final files. We'll be in touch if more information is needed.

Nature Ecology & Evolution has now transitioned to a unified Rights Collection system which will allow our Author Services team to quickly and easily collect the rights and permissions required to publish your work. Approximately 10 days after your paper is formally accepted, you will receive an email in providing you with a link to complete the grant of rights. If your paper is eligible for Open Access, our Author Services team will also be in touch regarding any additional information that may be required to arrange payment for your article.

Please note that Nature Ecology & Evolution is a Transformative Journal (TJ). Authors may publish their research with us through the traditional subscription access route or make their paper immediately open access through payment of an article-processing charge (APC). Authors will not be required to make a final decision about access to their article until it has been accepted. Find out more about Transformative Journals

Authors may need to take specific actions to achieve compliance with funder and institutional open access mandates. If your research is supported by a funder that requires immediate open access (e.g. according to Plan S principles) then you should select the gold OA route, and we will direct you to the compliant route where possible. For authors selecting the subscription publication route, the journal's standard licensing terms will need to be accepted, including <a href="https://www.nature.com/nature-portfolio/editorial-policies/self-archiving-and-license-to-publish. Those licensing terms will supersede any other terms that the author or any third party may assert apply to any version of the manuscript.

Please note that you will not receive your proofs until the publishing agreement has been received through our system.

For information regarding our different publishing models please see our Transformative Journals page. If you have any questions about costs, Open Access requirements, or our legal forms, please contact ASJournals@springernature.com.

Please use the following link for uploading these materials:
[REDACTED]

If you have any further questions, please feel free to contact me.

[REDACTED]

Reviewer #1:
Remarks to the Author:
I am satisfied with the revisions made by the authors. I have no further comments.

Reviewer #2:
Remarks to the Author:
The authors have thoroughly addressed the reviewers comments and resulting MS is much clearer and better reasoned. I have a few remaining comments, but I would leave the authors to decide whether they want to incorporate them, as opinions on some matters might just differ.

1) Line 55 'We focus on practices that can be empirically detected via analyses of published articles.' I guess other practices can be empirically detected too, but maybe with less precise estimates as of the low completeness of reporting of e.g. methods. So maybe add ' relatively reliably/easily' or similar

2) Lines 70/71 ' 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'

I am unsure about this. High impact journal do tend to attract different type of studies than lower impact journals, as well as might invite for more of some suboptimal practices as it is more difficult to get a study published if results are 'not interesting' . E.g. see doi: https://doi.org/10.1101/311068

3) 'lines 73-75: "Because most empirical studies report estimates in tables in the main or supplemental text, we assume that only including studies with estimates presented in tables does bias our results."

Maybe I misunderstand the first sentences, but about 1568 papers originally selected, 1038 do not report statistical test in tables. So, the opposite seems to be true – most articles do not report results in tables.

Also, I think you wanted to write 'does NOT bias our results'. While we do not have any evidence of differences between studies that do and don't report their results in tables, I would still not be sure that the difference does not exist.

4) Lines 84/85 ‘The amount of information that ecologists can extract from their data depends on the variability …’ ‘Information’ is quite a vague term, and the sentence is more about the reliability of information (as implied in the next sentence of the paragraph).

5) Seems that quite many estimates had to be excluded as sample size was not possible to determine. Maye this observation is something to briefly mention in the Discussion, saying that this ‘side ‘ result amplifies the need for better reporting.

Reviewer #3:
Remarks to the Author:
The authors have carefully addressed the reviewer comments. I have no further recommendations.

Final Decision Letter:

28th June 2023

Dear Dr Ferraro,

We are pleased to inform you that your Analysis entitled "Empirical evidence of widespread exaggeration bias and selective reporting in ecology", has now been accepted for publication in Nature Ecology & Evolution.

Over the next few weeks, your paper will be copyedited to ensure that it conforms to Nature Ecology and Evolution style. Once your paper is typeset, you will receive an email with a link to choose the appropriate publishing options for your paper and our Author Services team will be in touch regarding any additional information that may be required

After the grant of rights is completed, you will receive a link to your electronic proof via email with a request to make any corrections within 48 hours. If, when you receive your proof, you cannot meet this deadline, please inform us at rjsproduction@springernature.com immediately.

You will not receive your proofs until the publishing agreement has been received through our system

Due to the importance of these deadlines, we ask you please us know now whether you will be difficult to contact over the next month. If this is the case, we ask you provide us with the contact information (email, phone and fax) of someone who will be able to check the proofs on your behalf, and who will be available to address any last-minute problems . Once your paper has been scheduled for online publication, the Nature press office will be in touch to confirm the details.

Acceptance of your manuscript is conditional on all authors' agreement with our publication policies (see www.nature.com/authors/policies/index.html). In particular your manuscript must not be published elsewhere and there must be no announcement of the work to any media outlet until the

publication date (the day on which it is uploaded onto our web site).

Please note that Nature Ecology & Evolution is a Transformative Journal (TJ). Authors may publish their research with us through the traditional subscription access route or make their paper immediately open access through payment of an article-processing charge (APC). Authors will not be required to make a final decision about access to their article until it has been accepted. Find out more about Transformative Journals

Authors may need to take specific actions to achieve compliance with funder and institutional open access mandates. If your research is supported by a funder that requires immediate open access (e.g. according to Plan S principles) then you should select the gold OA route, and we will direct you to the compliant route where possible. For authors selecting the subscription publication route, the journal's standard licensing terms will need to be accepted, including <a href="https://www.nature.com/nature-portfolio/editorial-policies/self-archiving-and-license-to-publish. Those licensing terms will supersede any other terms that the author or any third party may assert apply to any version of the manuscript.
gi
In approximately 10 business days you will receive an email with a link to choose the appropriate publishing options for your paper and our Author Services team will be in touch regarding any additional information that may be required.

You will not receive your proofs until the publishing agreement has been received through our system.

If you have any questions about our publishing options, costs, Open Access requirements, or our legal forms, please contact ASJournals@springernature.com

An online order form for reprints of your paper is available at https://www.nature.com/reprints/author-reprints.html. All co-authors, authors' institutions and authors' funding agencies can order reprints using the form appropriate to their geographical region.

We welcome the submission of potential cover material (including a short caption of around 40 words) related to your manuscript; suggestions should be sent to Nature Ecology & Evolution as electronic files (the image should be 300 dpi at 210 x 297 mm in either TIFF or JPEG format). Please note that such pictures should be selected more for their aesthetic appeal than for their scientific content, and that colour images work better than black and white or grayscale images. Please do not try to design a cover with the Nature Ecology & Evolution logo etc., and please do not submit composites of images related to your work. I am sure you will understand that we cannot make any promise as to whether any of your suggestions might be selected for the cover of the journal.

You can now use a single sign-on for all your accounts, view the status of all your manuscript

submissions and reviews, access usage statistics for your published articles and download a record of your refereeing activity for the Nature journals.

To assist our authors in disseminating their research to the broader community, our SharedIt initiative provides you with a unique shareable link that will allow anyone (with or without a subscription) to read the published article. Recipients of the link with a subscription will also be able to download and print the PDF.

You can generate the link yourself when you receive your article DOI by entering it here: <a href="http://authors.springernature.com/share">http://authors.springernature.com/share<a>.

Yours sincerely,

Simon Harold PhD
Senior Editor
Nature Ecology and Evolution

P.S. Click on the following link if you would like to recommend Nature Ecology & Evolution to your librarian http://www.nature.com/subscriptions/recommend.html#forms

** Visit the Springer Nature Editorial and Publishing website at www.springernature.com/editorial-and-publishing-jobs for more information about our career opportunities. If you have any questions please click here.**c