nature ecology & evolution

Analysis

https://doi.org/10.1038/s41559-023-02144-3

Empirical evidence of widespread exaggeration bias and selective reporting in ecology

Received: 29 June 2022

Accepted: 28 June 2023

Published online: 03 August 2023

Check for updates

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

In many scientific disciplines, common research practices have led to unreliable and exaggerated evidence about scientific phenomena. Here we describe some of these practices and quantify their pervasiveness in recent ecology publications in five popular journals. In an analysis of over 350 studies published between 2018 and 2020, we detect empirical evidence of exaggeration bias and selective reporting of statistically significant results. This evidence implies that the published effect sizes in ecology journals exaggerate the importance of the ecological relationships that they aim to quantify. An exaggerated evidence base hinders the ability of empirical ecology to reliably contribute to science, policy, and management. To increase the credibility of ecology research, we describe a set of actions that ecologists should take, including changes to scientific norms about what high-quality ecology looks like and expectations about what high-quality studies can deliver.

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

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

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

We have three aims. First, we seek to provide a primer for new scientists and a refresher for experienced scientists on practices that

[1]Mad Agriculture, Boulder, CO, USA. [2]Department of Earth and Planetary Sciences, Johns Hopkins University, Baltimore, MD, USA. [3]Carey Business School, Johns Hopkins University, Baltimore, MD, USA. [4]Department of Environmental Health and Engineering, a joint department of the Bloomberg School of Public Health and the Whiting School of Engineering, Johns Hopkins University, Baltimore, MD, USA. e-mail: pferraro@jhu.edu

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

To determine the extent to which these practices are prevalent in the ecology literature, we collected data from empirical studies published between January 2018 and May 2020 in five popular journals that publish general-interest ecology studies and include many empirical designs: *Ecology*, *Ecology Letters*, *Journal of Ecology*, *Nature*, and *Science*. We believe that these journals are representative of good quality ecological studies, and thus we assume that the exclusion of other journals does not bias our conclusions. We included only empirical articles that reported statistically estimated parameters and errors in tables in the main or supplemental texts. Simulation, modelling and meta-analysis articles were excluded. Because most statistical tests can be presented in table format and we have no reason to assume that certain tests or types of test are more frequently reported in tables, we assume that including only estimates presented in tables does not bias our results. For every study, we then recorded: (1) every estimate and its associated error, (2) the sample size, (3) whether the study used multiple hypothesis testing, (4) whether there were corrections for multiple hypothesis testing and (5) whether data and code for analyses were available.

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

Practices that lead to low credibility

Under-powered designs

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

Given that most ecologists have training in frequentist statistics and engage in hypothesis testing, we explore the reliability of the estimated effects sizes in the ecology literature through the lens of statistical power. The statistical power of a test is the chance of detecting an effect, if such an effect exists[14]. Statistical power is based on the anticipated effect size, the sample size, the type-1 error rate and the sample variability. A conventional threshold for sufficient statistical power is 0.80, meaning that, if an effect of a given magnitude exists, a study design will detect it 80% of the time. Ecologists often seek to estimate the relationship between two variables and test whether the estimated value is different from a null hypothesis, which is usually that there is no relationship between the two variables. Consider, for example, a study that looks at how plant growth is related to phosphorus addition. A null hypothesis could be that phosphorus addition has no effect on plant growth. If a study is adequately powered, one would be likely to reject this null hypothesis if it were in fact false because the variability of the estimated effect sizes around the true effect size will be low. If, however, the study is under-powered, rejecting the null hypothesis would be unlikely because the variability of the estimated effect sizes around the true effect will be large. Thus, under-powered designs lead to greater prevalence of type-2 errors.

To estimate the statistical power of studies in our data, we followed the methods in ref. 15. First, we calculated an estimate for the magnitude of the true effect sizes that our collection of studies attempts to estimate. We estimated this effect as the weighted average of the partial correlation coefficients (PCCs) for all estimates in our study. A PCC is a measure of the strength and direction of the relationship between two variables when the influence of all other variables is held constant. Like a meta-analysis, this weighted average gives more weight to studies with more precise estimates. Our estimated 'true effect' for our collection of studies was a PCC value of 0.06. Implicitly, we assume that there is no selective reporting or publication bias against small effect size estimates in the literature (that is, we assume ecologists report in the final publication everything that they estimated). Then we calculated the statistical power of the studies to detect this effect size (see 'Power analysis' for details). This approach does not imply that ecological effect sizes are homogenous across sites, studies or variables in our 354 studies. Rather, the approach offers an approximation of the magnitude of the true effect size that a typical ecological study would expect to find.

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

Whether our approach yields an accurate approximation of the statistical power of a typical ecology study also depends on another assumption. We assume that ecologists care about distinguishing small effect sizes from 0 (for example, 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 0 the typically small effect sizes reported in the literature is incorrect, we have underestimated power in our analysis above.

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

Exaggeration bias

The prevalence of under-powered study designs can lead to an exaggeration bias[9,21] in published studies when statistically significant results are preferred over non-significant results by editors, reviewers

https://doi.org/10.1038/s41559-023-02144-3

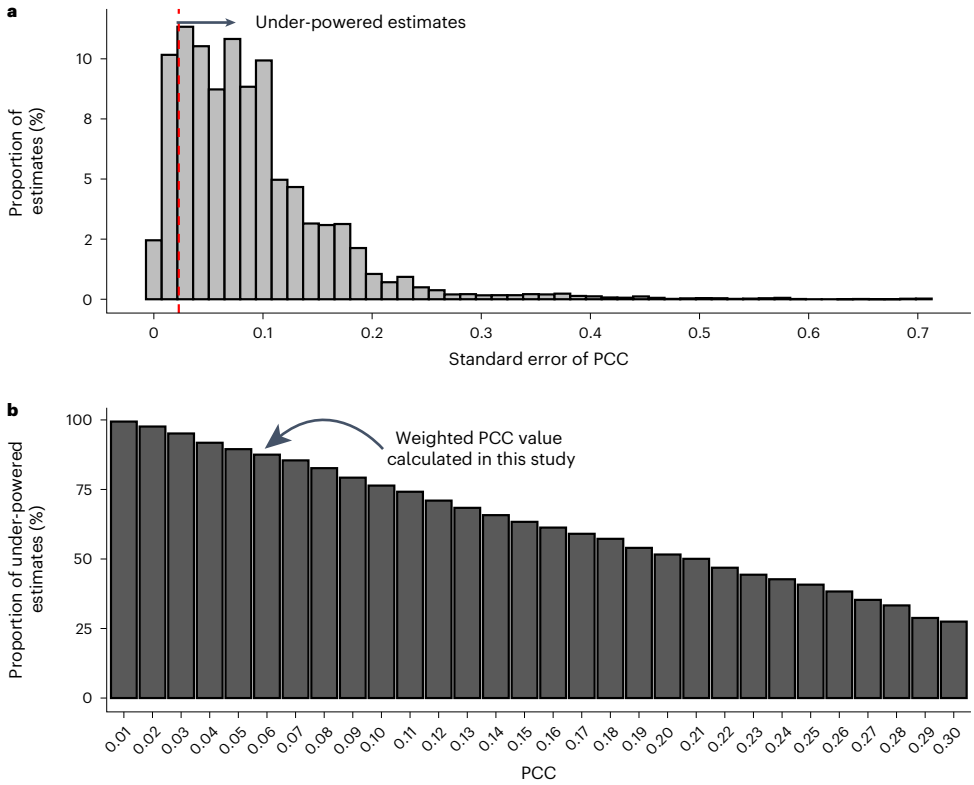

Fig. 1 | Percentage of statistical tests that meet and do not meet the conventional 0.8 threshold for statistical power. a, A histogram of the standard error of the PCCs from ecological studies. All estimates to the right of the red line are under-powered at an 80% power threshold. $n = 18{,}917$ estimates from 354 studies. **b**, The percentage of the 18,917 estimates that would be under-powered for a range of PCC values.

and authors (that is, publication bias[22]). Previous studies have reported evidence of publication biases in ecology[2–4,23], and these biases may be more severe in high-impact journals such as the ones we include in our study[24]. To illustrate how exaggeration bias arises, we consider again the example of a study that seeks to estimate the effect of phosphorus addition on plant growth. Assume that the true treatment effect is a 2% increase in aboveground biomass. In adequately powered studies, most estimated effects would be close to the 2% increase. In under-powered studies, however, the estimated values would vary widely around 2%, such that researchers are likely to report values that are much larger than the true value (type-M error) or even opposite in sign (type-S error)[21]. Yet, in under-powered studies, only the values with exaggerated magnitudes are going to be statistically significant (that is, with confidence intervals that exclude 0).

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

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

Our analysis implies that 63% of the estimates in under-powered studies are exaggerated over the true effect size by a factor of 2 or more (Fig. 2a). Even if we assume a 'true effect' of much greater magnitude, 1 in 4 estimates would still be exaggerated by a factor of 2 or more (Fig. 2b). Our results are similar to a recent study of effect size exaggeration in three types of experimental ecological field study. Using a different methodology, this study found that estimates were exaggerated by anywhere from 0.66 times (drought experiments) to 3.29 times (warming experiments) on average[17].

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

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.

In our study, we use the concept of statistical power simply as a vehicle to illustrate the inconvenient truth about ecological data: the outcome variables are noisy, the target effect sizes are typically smaller than ecologists expect and, given the designs ecologists are using and the incentives they are facing, the estimated parameters in the literature are likely to be unreliable and exaggerated. Our use of statistical

https://doi.org/10.1038/s41559-023-02144-3

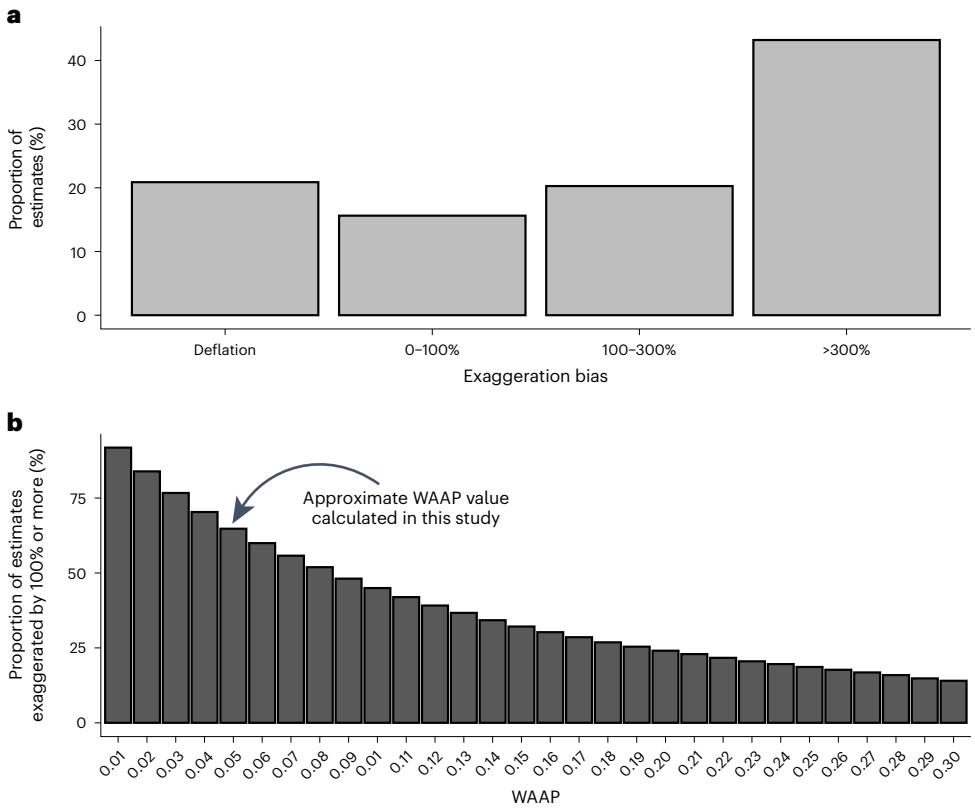

Fig. 2 | The percentage of under-powered estimates from ecological studies that are exaggerated. a, The percentage of estimates from under-powered studies that are exaggerated based on the weighted averages of adequately powered estimates in our sample of studies. Deflation refers to any estimate that is smaller than the hypothesized true effect, while the other categories represent exaggeration. $n = 16{,}407$ estimates from 330 studies with under-powered estimates. **b**, The percentage of the 16,407 estimates that would be exaggerated by 100% or more given a range of WAAP values.

power to explore the reliability of estimated effects in the ecological literature is not an endorsement of null hypothesis statistical testing or the use of binary decision rules based on *P* values to decide when an estimate is ecologically relevant (for example, $P < 0.05$)[27–32].

Selective reporting of results

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

To explore the extent of selective reporting of statistically significant results in ecology, we followed the methods in ref. 33. We plotted the density of reported *t*-statistics and overlaid an Epanechnikov density kernel. We then weighted estimates by the number of estimates per table in each article (see 'Selective reporting' for more details). Without selective reporting, the density kernel should be a smooth function that declines as *t* values increase. In contrast, a dip in the kernel density that creates a bimodal distribution with a second peak before the traditional 1.96 cut-off value for significance (that is, $P = 0.05$) implies the presence of selective reporting practices (not all selective reporting practices lead to a bimodal distribution[37], and thus its absence does not necessarily imply an absence of selective reporting practices).

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

We hypothesize that this pattern of test statistics may arise from three sources. First, a researcher may pose a hypothesis that *X* influences *Y* and then use data on *X* and its covariates to test the hypothesis. The researcher may try multiple model specifications and statistical tests and then choose the combination that yields the most compelling results about the effect of *X* on *Y* to include in the main text, relegating the less compelling results to the supplemental material. Second, the same researcher may be unable to reject the null hypothesis that *X* has no effect on *Y* with any model or test. They then may search for other interesting and statistically significant effects in the data to report and revise the hypothesis they claim to be testing in the main

https://doi.org/10.1038/s41559-023-02144-3

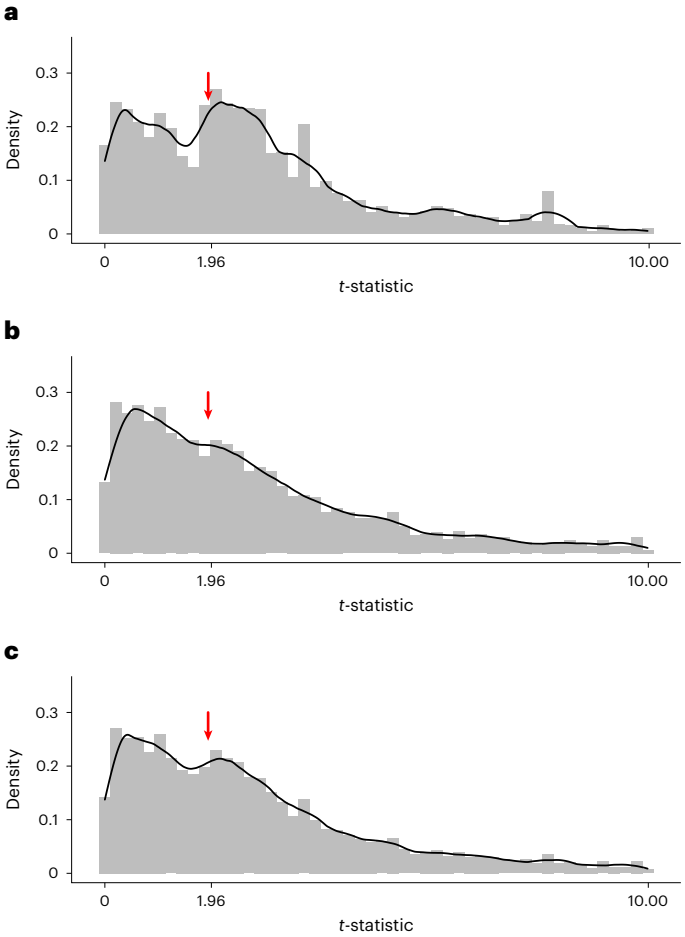

Fig. 3 | Evidence of selective reporting of statistically significant results. **a**–**c**, Evidence of selective reporting of statistically significant results in main text tables (n = 2,286 estimates) (**a**), supplemental text tables (n = 14,680 estimates) (**b**) and all tables (n = 16,966) (**c**)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

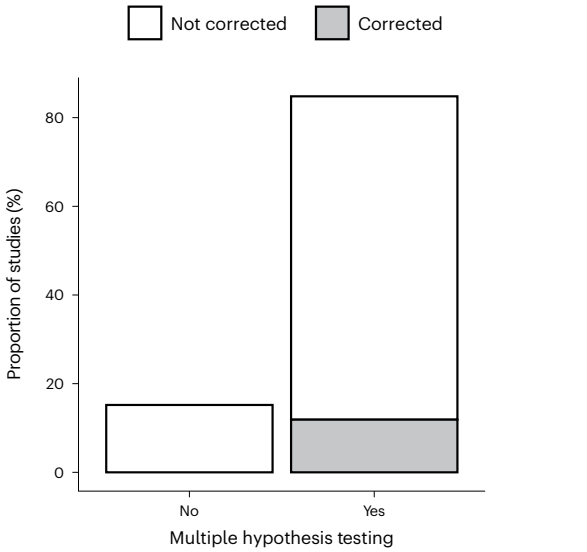

Fig. 4 | The percentage of ecology studies that use multiple hypothesis testing. The grey section represents the percentage of studies that used multiple hypothesis corrections. n = 354 studies.

text (HARKing). The researcher may still present all the tests that they conducted but place the nonsignificant results in the supplement instead of the main text. Third, rather than test a single hypothesis, ecology researchers often posit research questions in the form 'what determines Y?' Such studies yield a range of estimated parameters, at least one estimate for each posited determinant of Y and maybe more if the researcher uses a variety of plausible models. The researcher may then selectively pick the 'most interesting' estimates to report in the main text or, if they report all of the estimates, they may selectively pick the estimates from the 'best' model ('best' could be determined by statistical criteria but may also be determined by criteria that maximize the probability of publication, such as 'how many statistically significant variables are obtained' or 'what understudied variables deliver statistically significant results'). The perceived 'less interesting' estimates or 'inferior' models are relegated to the supplemental materials.

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

Multiple hypothesis testing

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

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

Table 1 | Changes in research practices to help increase the reliability of ecological research

Recommendation	Details	Purpose	References
Checklists	Used at multiple stages of the publication process: for example, they can be used before submitting, during review and by editors	– Ensure researchers include necessary information for evaluating the study – Highlight key features of study design for reviewers – Educate authors and reviewers on best practices	Nosek et al.[1]; Simmons et al.[35]; Parker et al.[14,81]
Data and code archiving	Publicly available except where data privacy is necessary	– Increase the transparency of study workflows and conclusions – Facilitate computational reproducibility and evidence synthesis	Parker et al.[14]; Culina et al.[59]; Munafò et al.[82]; Nosek et al.[83]; Nakagawa & Parker[84]
Pre-registration and pre-analysis plans	Pre-analysis plans: describe the research questions, the design and the methods that will be used in a study; completed before data analysis begins (ideally, before all data have been collected). Pre-registration: process of registering, before the study or data analysis begins, a researcher's intent to undertake a study and the study's pre-analysis plan	– Help authors to be transparent in their research decisions – Reduce, or at least make more transparent, the practices of HARKing, selective reporting of results and presentations of exploratory analyses as if they were confirmatory analyses planned from the outset – Help scholars quantify the 'file drawer' problem: studies that were completed but never published	Parker et al.[52]; Forstmeier et al.[47]; Kaplan & Irvin[85]; Nosek et al.[86]
Registered reports	Two-stage peer review. Before data collection and analysis, authors submit study motivation, design and methods. Reviewers judge submission based on quality of question and design. Second-stage reviews assess how closely study follows original plan	– Reduce selective reporting of results – Reviewers focus on importance of the question and quality of the design, not the sign, magnitude and statistical significance of results	https://www.cos.io/initiatives/registered-reports Allen & Mehler[87]; Scheel et al.[88]; Nosek et al.[89]; Button et al.[90]; Soderberg et al.[91]
Results—blind reviews	Full manuscript submitted for review, but results are not included	– Reviewers focus on importance of the question and quality of the design, not the sign, magnitude and statistical significance of results – No mechanism to reduce selective reporting because no pre-analysis plan is required	Button et al.[90]; Smulders[92]
Incentives	Institutions that matter—namely, employers, funders and publishers—move away from incentivizing 'exciting' results and towards incentivizing best practices	– Align personal values of many researchers to create and disseminate credible science – Value replication studies along with 'ground-breaking' research	Nosek et al.[1]; O'Dea et al.[50]; Anderson et al.[93]. http://sortee.org https://sfdora.org/

See Supplementary Text 'Promising actions ...' for more details on practices.

the credibility of the results reported in ecology publications cannot be fully appreciated.

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

Together with selective reporting (for which we presented evidence in the previous section) and publication bias, multiple hypothesis testing may skew how researchers interpret the evidence base[47]. Researchers may be incentivized to report only 'interesting' and statistically significant results instead of all the tests they performed on the dataset. Thus, we may not even know the extent to which multiple hypothesis testing occurs because some results may be simply excluded from publications.

Fostering a credibility culture in empirical ecology

Strengthening the reliability of ecological evidence will require changes in how ecologists produce and consume research. Ecologists must change their expectations about what high-quality ecological studies should look like and their expectations about what high-quality ecological studies can deliver. While these expectations can be shaped through better statistical knowledge[48,49], knowledge alone will be insufficient.

Changing expectations about what high-quality studies look like and can deliver will require changes in the incentives that ecologists face and in the norms that guide their empirical work. To encourage these changes across scientific fields, scholars have proposed a range of actions, including actions that individual researchers can take and actions that researchers must implement collectively[1]. A few publications describe some of these actions and some of the challenges to scaling these actions in the context of ecology[10,14,50–53]. We believe that most ecologists would readily adopt these actions but are not yet aware of them.

To help foster greater awareness, we highlight in Table 1 some promising actions that we believe will best contribute to improving the credibility and reliability of empirical ecology. Some of these actions, such as pre-registration and registered reports, are not well known in ecology. More widely known is the importance of data and code availability for computational reproducibility[54,55] (a study is computationally reproducible if the same results can be achieved with the data and code used for the original analyses[12,13]). Best practices have been laid out for data and code archiving in ecology[56–60], and several journals (for example, *Journal of Ecology* and Ecological Society of America publications (https://www.esa.org/publications/data-policy/)) and institutions (for example, the National Science Foundation-funded Long-Term Ecological Research network) require public data archiving. Yet, despite these attempts to make data and code more accessible (for example, ref. 61), obtaining data and code can still be challenging[59,62–67]. For example, researchers were only able to obtain data from 19 of 74 articles in wildlife management. Using the data from these 19 publications, the researchers could reproduce the results in only 6 publications, even though code was provided for 9 studies[55]. Therefore, availability does not equate to quality of data or code[60]; most

https://doi.org/10.1038/s41559-023-02144-3

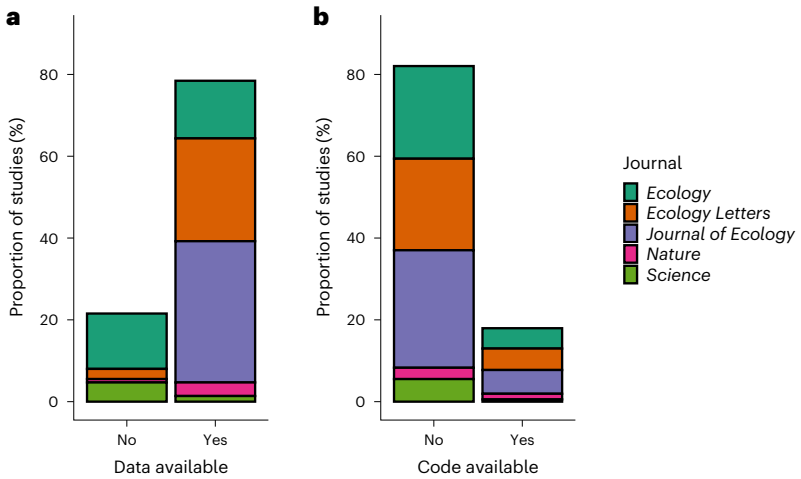

Fig. 5 | Ecology studies that have data available and provide code for their analyses. a,**b**, The percentage of ecology studies that have data available (**a**) and provide code for their analyses (**b**). Bars are coloured by journal. $n = 354$ studies.

ecology and evolution publicly available datasets in a recent analysis were not reusable (a measure of ease with which data can be reused by third parties), and only slightly over half were complete[65]. In our dataset of 354 studies, we found that most studies (78.5%) did make the data available, but only 18% of studies provided code for their analysis (and the code provided did not necessarily show the data cleaning steps; Fig. 5). These percentages are similar to those reported using a sample of 346 articles from ecology journals that had mandatory or encouraged code-sharing policies. In that study, 79% of studies provided data, 27% provided code and 21% had both data and code[59].

Even with broader implementation of actions such as pre-registration and the provision of both data and analysis code, many important decisions will remain in the hands of researchers and thus unobservable to outsiders. Thus, to fully address the issues raised in our article, we need a cultural shift, a shift where we assign more value to important questions and best practices and less value to exciting stories and statistically significant results[50,68]. Given the complexity of ecological systems, we should not expect high-quality empirical studies to provide 'airtight' conclusions or discontinuous jumps in our understanding of ecological processes. Instead, we should expect single studies to incrementally build on previous studies, to have substantial uncertainty arising from many sources (not just sampling variability) and to even present conflicting inferences, implying that we do not fully understand the underlying ecological processes.

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

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

Methods

Data collection

Our methods follow those of ref. 7. We collected data from articles published between January 2018 and May 2020 in five popular journals for ecology publications. We collected data from every empirical article in three ecology journals (*Ecology*, *Ecology Letters* and *Journal of Ecology*) and every empirical ecology article in two general-interest journals (*Nature* and *Science*) ($n = 1{,}568$ papers total). Only empirical articles that statistically estimated parameters from data were included. These articles needed to have reported estimates and errors (standard errors or 95% confidence intervals) in tables either in the main text or in supplemental materials. We focused on results reported in tables so that estimates and associated errors were easy to identify by the research team and to make sure that we were able to collect enough estimates for our analyses. Simulation or modelling articles were excluded. Meta-analyses were also excluded because we sought primary empirical data and did not want to double count any estimates that were found in both an original study and a meta-analysis.

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

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

From the 1,568 papers in the five journals between our target years, we excluded 1,038 that did not report statistical tests in tables. We excluded 136 that were either meta-analyses or not empirical. Fifteen papers were removed that did not report errors and another three were removed that reported 0 for a standard error. One paper was removed because it was duplicated in 2019, and one was removed because the supplemental materials where tables may have been located did not open. Seventeen complete papers were removed because we could not discern sample sizes for any of the tests. When checking our sampled data, one paper was removed because it should not have been classified as an ecology topic from *Science*. During data processing, we removed one publication that had over 6,000 estimates, and one was removed when we discarded the top percentile of *t*-statistics. Thus, our final sample size was 354 publications.

When confidence intervals were reported instead of standard errors, we calculated the upper confidence interval minus the estimate and the lower confidence interval minus the estimate. We then recorded the smaller of the two if the interval was uneven. Thus, we are assuming less error about an estimate and potentially biasing our

results towards a more favourable assessment of the literature than is warranted. These values were divided by 1.96 to obtain an equivalent standard error. Our use of 1.96 may not be correct for small sample sizes, but assuming that 1.96 is the benchmark will attribute less error about the point estimate. Thus, we will be overestimating the power of the tests. In other words, it makes our estimates of power more conservative.

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

To determine if a study used multiple hypothesis testing, we read the methods and looked at results presented in the main text of the manuscript. We categorized a study as using multiple hypothesis testing if the authors investigated multiple outcomes (dependent variables) associated with one cause (independent variable), investigated multiple causes (independent variables) associated with one outcome (dependent variables) or investigated sub-groups within their dataset. We were not concerned with one multiple regression being run (which could fall under multiple causes associated with one outcome) but instead several multiple regressions being run on the same dataset. We tried not to include robustness checks as multiple hypothesis testing. We identified robustness checks by reading how the analysis was referenced and, where possible, by reading figure or table captions. In most cases, robustness checks were easily identified—but the text was not always clear.

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

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

Software used

All data manipulation were done in R version 4.0.0[69], and we utilized the 'here' package (version 1.0.1) for replicability[70]. Throughout our script, we used dplyr (version 1.0.7)[71] and tidyr (version 1.1.4)[72] to manipulate our data. We also relied on ggplot2 (version 3.3.5)[73], ggpubr (version 0.4.0)[74], patchwork (version 1.1.1)[75] and scales (version 1.1.1)[76] for making figures.

Data cleaning

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

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

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

Power analysis

To estimate the statistical power of studies in our dataset and the extent of exaggeration bias, we followed the methods in ref. 15. Power calculations are conditional on some assumption of the size of the effect that the researchers are seeking to estimate. Here we expressed power in the form of the minimum detectable effect (MDE). The MDE of a study design is the smallest effect that, if true, has an X% chance of producing an impact estimate that is statistically significant at the Y% level[77]. X is the level of statistical power (denoted as $(1-\beta)$ and commonly set to 80%), and Y is the type-1 error rate (denoted as α and commonly set to 5%). The MDE can be written in terms of the standard error[78]:

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

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

$$\mathrm{MDE} = (1.96 + 0.84)\,\varepsilon = 2.8\varepsilon \quad (2)$$

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

To calculate the MDE across our sample of studies, we must convert the estimates to a unitless measure with a common scale. This conversion allows us to compare estimates across studies. Here we used the PCC, calculated as[79]:

$$\mathrm{PCC} = \frac{t}{\sqrt{t + \mathrm{d.f.}}} \quad (3)$$

where t is the associated t-statistic of the estimate and d.f. is the degrees of freedom. The standard error of the PCC was then estimated using[79]:

$$\mathrm{SEpcc} = \frac{\mathrm{PCC}}{t} = \frac{1}{\sqrt{t^2 + \mathrm{d.f.}}} \quad (4)$$

Using the absolute values of PCC, we calculated the weighted average PCC for our entire dataset. The PCC values were weighted by the estimates' precision (for example, the standard error about the estimate), so that estimates with higher precision (smaller standard errors) were assigned a larger weight. This weighted average PCC value served as our estimate of the true effect (the MDE in equation (2)) that ecological studies are attempting to estimate. We then divided the weighted average of the PCC values by 2.8 to get the threshold to which we compared the SEpcc values. When the SEpcc of an estimate was less than or equal to the threshold, the estimate had adequate power; otherwise, it was under-powered. We repeated these analyses for 75% and 60% power also where the weighted PCC was divided by 2.63 and 2.21, respectively, to obtain the threshold values. See lines 110–142 in RepCode.R for how these analyses were done.

Most published studies did not provide the information required to calculate the degrees of freedom (d.f.) for each model. To be consistent across studies, we approximate d.f. using the sample size, N. Thus, we are often overestimating the d.f. of a model, even more so when the estimates come from a mixed effects model (42% of the estimates in our dataset are from some sort of mixed effects model). Therefore, most of our calculated PCC values are smaller than they would be if we used d.f.

Because we are using N, we are also likely underestimating the standard error of the PCC values (which are smaller to a greater degree than the PCC values are smaller). This will reduce our standard error of the PCC values which we compare to the MDE threshold. Thus, overall, we are likely overestimating the power of most tests in our sample of studies.

We recognize that each empirical study in ecology seeks to estimate a different effect, whose true value may vary across studies. Given that the true effect size is not known, we also explored how our conclusions changed with changes in the assumed true effect size (Fig. 1b). For a range of 'true effect' values, we computed how many PCC estimates had a standard error greater than the threshold value based on hypothetical true effect sizes divided by 2.8. Our range went up to PCC values of 0.20 (in terms of standard deviations of the outcome variable, this effect size would be analogous to an effect size of roughly 0.5 s.d.). Our estimated weighted PCC value (our MDE in equation (2)) from our entire dataset was 0.06. For only observational studies in our dataset, it was 0.05. For only experimental studies in our dataset, it was 0.19. These values make sense if we assume that experimental studies tend to push the system further than observational studies and, consequently, have larger effects to report. Furthermore, this range spans most of the PCC values recorded from our dataset (Supplementary Fig. 1) and covers the unweighted median PCC value of our sample. Thus, the values we present in Fig. 1b represent a reasonable range of PCC values that we may expect in ecological studies.

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

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

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

Step 3: Let us consider more deeply the two assumptions required to come to the opposite conclusion from the one described in our manuscript. Assumption 1 would require that ecologists think very carefully about the noise in their data and the magnitude of the target effect size before collecting data. Although we acknowledge that statistical power calculations or simulations are not the only way to think about such design attributes, they are likely to be one of the most popular ways of doing so among ecologists. Yet if ecologists conduct power analyses with regularity, they do not report them in their publications: only one study of the 353 publications in our dataset reported conducting a power analysis.

Even if ecologists do carefully think about the noise in their data and the magnitude of the target effect size before collecting data, assumption 2 would require one of two additional conditions. First, when the expected treatment effect sizes are large, the costs of data collection or selecting study units are also large. This pattern of costs could imply that, in comparison to ecologists seeking to estimate small true effect sizes, ecologists seeking to estimate large effects cannot as easily reduce the influence of noise by increasing sample size or by selecting a subset of the target population that has lower outcome variance. If this first condition about differences in relative costs were not satisfied, an alternative condition could support assumption 2. In comparison to ecologists who work on studies seeking to estimate small effect sizes, ecologists who seek to estimate large true effect sizes must be more cognizant that peer reviewers and editors are unlikely to care about the precision of their estimates as long as the confidence interval does not cross the null hypothesis value.

Lastly, we computed the median power for our sample of tests as in ref. 80. The median power is calculated as 1 minus the cumulative normal probability of the difference between 1.96 and the absolute value of the weighted average PCC estimate divided by the median standard error. We calculated this value for six sets of the data: the entire dataset, the set of 'main' estimates, the set of estimates in the main text, the set of estimates in the supplemental text, the set of estimates from observational studies and the set of the estimates from experimental studies (see RepCode.R lines 304–333 for these calculations).

Exaggeration bias

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

Again, because we acknowledge that the WAAP estimate may be different for different types of study, we then explore how our conclusions may change given different WAAP values (Fig. 2b). For a range of WAAP values from 0.01 to 0.2, we calculated how many estimates would be inflated by 100% or more. To do this, we compared the WAAP values in this range to the absolute value of the PCC values for under-powered estimates. Any PCC value divided by the WAAP that was greater than 2 was considered inflated by 100% or more. See RepCode.R lines 335–417 for these calculations and creation of figures.

Selective reporting

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

Multiple hypothesis testing, data and code availability

We calculated the percentage of studies in our dataset that used multiple hypothesis testing and the percentage that used corrections for multiple hypothesis testing (see definitions in 'Data collection' section

above). To quantify the extent to which the data and analysis code from our studies are available for replication, we calculated the percentage of studies that made the data or analysis code, or both, available.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

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

Code availability

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

References

1. Nosek, B. A., Spies, J. R. & Motyl, M. Scientific utopia: II. restructuring incentives and practices to promote truth over publishability. *Perspect. Psychol. Sci.* **7**, 615–631 (2012).
2. Leimu, R. & Koricheva, J. Cumulative meta-analysis: a new tool for detection of temporal trends and publication bias in ecology. *Proc. R. Soc. B* **271**, 1961–1966 (2004).
3. Møller, A. P. & Jennions, M. D. Testing and adjusting for publication bias. *Trends Ecol. Evol.* **16**, 580–586 (2001).
4. Barto, E. K. & Rillig, M. C. Dissemination biases in ecology: effect sizes matter more than quality. *Oikos* **121**, 228–235 (2012).
5. Christensen, G. & Miguel, E. Transparency, reproducibility, and the credibility of economics research. *J. Econ. Lit.* **56**, 920–980 (2018).
6. Collaboration, O. S. Estimating the reproducibility of psychological science. *Science* **349**, aac4716 (2015).
7. Ferraro, P. J. & Shukla, P. Is a replicability crisis on the horizon for environmental and resource economics? *Rev. Environ. Econ. Policy* **14**, 339–351 (2020).
8. Martinson, B. C., Anderson, M. S. & de Vries, R. Scientists behaving badly. *Nature* **435**, 737–738 (2005).
9. Ioannidis, J. P. A. Why most published research findings are false. *PLoS Med.* **2**, 696–701 (2005).
10. Fraser, H., Parker, T., Nakagawa, S., Barnett, A. & Fidler, F. Questionable research practices in ecology and evolution. *PLoS ONE* **13**, e0200303 (2018).
11. Fraser, H., Barnett, A., Parker, T. H. & Fidler, F. The role of replication studies in ecology. *Ecol. Evol.* **10**, 5197–5207 (2020).
12. Fidler, F. et al. Metaresearch for evaluating reproducibility in ecology and evolution. *Bioscience* **67**, 282–289 (2017).
13. Cassey, P. & Blackburn, T. M. Reproducibility and repeatability in ecology. *Bioscience* **56**, 958–959 (2006).
14. Parker, T. H. et al. Transparency in ecology and evolution: real problems, real solutions. *Trends Ecol. Evol.* **31**, 711–719 (2016).
15. Ioannidis, J. P. A., Stanley, T. D. & Doucouliagos, H. The power of bias in economics research. *Econ. J.* **127**, F236–F265 (2017).
16. Jennions, M. D. & Møller, A. P. A survey of the statistical power of research in behavioral ecology and animal behavior. *Behav. Ecol.* **14**, 438–445 (2003).
17. Lemoine, N. P. et al. Underappreciated problems of low replication in ecological field studies. *Ecology* **97**, 2562–2569 (2016).
18. Yang, Y. et al. Publication bias impacts on effect size, statistical power, and magnitude (type M) and sign (type S) errors in ecology and evolutionary biology. *BMC Bio.* **21**, 71 (2023).
19. Button, K. S. et al. Power failure: why small sample size undermines the reliability of neuroscience. *Nat. Rev. Neurosci.* **14**, 365–376 (2013).
20. Fidler, F., Burgman, M. A., Cumming, G., Buttrose, R. & Thomason, N. Impact of criticism of null-hypothesis significance testing on statistical reporting practices in conservation biology. *Conserv. Biol.* **20**, 1539–1544 (2006).
21. Gelman, A. & Carlin, J. Beyond power calculations: assessing type S (sign) and type M (magnitude) errors. *Perspect. Psychol. Sci.* **9**, 641–651 (2014).
22. Nichols, J. D., Oli, M. K., Kendall, W. L. & Scott Boomer, G. A better approach for dealing with reproducibility and replicability in science. *Proc. Natl Acad. Sci. USA* **118**, 1–5 (2021).
23. Koricheva, J. Non-significant results in ecology: a burden or a blessing in disguise? *Oikos* **102**, 397–401 (2003).
24. Ceausu, I. et al. High impact journals in ecology cover proportionally more statistically significant findings. Preprint at *bioRxiv* https://doi.org/10.1093/sw/38.6.771 (2018).
25. Nichols, J. D., Kendall, W. L. & Boomer, G. S. Accumulating evidence in ecology: once is not enough. *Ecol. Evol.* **9**, 13991–14004 (2019).
26. Fanelli, D. Negative results are disappearing from most disciplines and countries. *Scientometrics* **90**, 891–904 (2012).
27. Fanelli, D. Is science really facing a reproducibility crisis, and do we need it to? *Proc. Natl Acad. Sci. USA* **115**, 2628–2631 (2018).
28. Yoccoz, N. G. Use, overuse, and misuse of significance tests in evolutionary biology and ecology. *Bull. Ecol. Soc. Am.* **72**, 106–111 (1991).
29. Fidler, F., Fraser, H., McCarthy, M. A. & Game, E. T. Improving the transparency of statistical reporting in *Conservation Letters*. *Conserv. Lett.* **11**, 1–5 (2018).
30. Murtaugh, P. A. In defense of *P* values. *Ecology* **95**, 611–617 (2014).
31. Anderson, D. R., Burnham, K. P. & Thompson, W. L. Null hypothesis testing: problems, prevalence, and an alternative. *J. Wildl. Manag.* **64**, 912–923 (2000).
32. Callaham, M., Wears, R. L. & Weber, E. Journal prestige, publication bias, and other characteristics associated with citation of published studies in peer-reviewed journals. *J. Am. Med. Assoc.* **287**, 2847–2850 (2002).
33. Brodeur, A., Lé, M., Sangnier, M. & Zylberberg, Y. Star wars: the empirics strike back. *Am. Econ. J. Appl. Econ.* **8**, 1–32 (2016).
34. Gopalakrishna, G. et al. Prevalence of questionable research practices, research misconduct and their potential explanatory factors: a survey among academic researchers in the Netherlands. *PLoS ONE* **17**, 1–16 (2022).
35. Simmons, J. P., Nelson, L. D. & Simonsohn, U. False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychol. Sci.* **22**, 1359–1366 (2011).
36. Head, M. L., Holman, L., Lanfear, R., Kahn, A. T. & Jennions, M. D. The extent and consequences of *P*-hacking in science. *PLoS Biol.* **13**, 1–15 (2015).
37. Hartgerink, C. H. J., Van Aert, R. C. M., Nuijten, M. B., Wicherts, J. M. & Van Assen, M. A. L. M. Distributions of *p*-values smaller than .05 in psychology: what is going on? *PeerJ* **2016**, e1935 (2016).
38. Shaffer, J. P. Multiple hypothesis testing. *Annu. Rev. Psychol.* **46**, 561–584 (1995).
39. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B* **57**, 289–300 (1995).
40. Dunnett, C. W. A multiple comparison procedure for comparing several treatments with a control. *J. Am. Stat. Assoc.* **50**, 1096–1121 (1955).
41. Yekutieli, D. & Benjamini, Y. Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics. *J. Stat. Plan. Inference* **82**, 171–196 (1999).
42. Berry, D. A. & Hochberg, Y. Bayesian perspectives on multiple comparisons. *J. Stat. Plan. Inference* **82**, 215–227 (1999).
43. Gelman, A., Hill, J. & Yajima, M. Why we (usually) don't have to worry about multiple comparisons. *J. Res. Educ. Eff.* **5**, 189–211 (2012).

44. Rubin, M. Do *p* values lose their meaning in exploratory analyses? It depends how you define the familywise error rate. *Rev. Gen. Psychol.* **21**, 269–275 (2017).
45. Rubin, M. When does HARKing hurt? Identifying when different types of undisclosed post hoc hypothesizing harm scientific progress. *Rev. Gen. Psychol.* **21**, 308–320 (2017).
46. Nakagawa, S. A farewell to Bonferroni: the problems of low statistical power and publication bias. *Behav. Ecol.* **15**, 1044–1045 (2004).
47. Forstmeier, W., Wagenmakers, E. J. & Parker, T. H. Detecting and avoiding likely false-positive findings—a practical guide. *Biol. Rev.* **92**, 1941–1968 (2017).
48. Baker, M. & Penny, D. Is there a reproducibility crisis? *Nature* **533**, 452–454 (2016).
49. Gelman, A. & Loken, E. The statistical crisis in science. *Am. Sci.* **102**, 460–465 (2014).
50. O'Dea, R. E. et al. Towards open, reliable, and transparent ecology and evolutionary biology. *BMC Biol.* **19**, 1–5 (2021).
51. Parker, T. H., Nakagawa, S. & Gurevitch, J. Promoting transparency in evolutionary biology and ecology. *Ecol. Lett.* **19**, 726–728 (2016).
52. Parker, T., Fraser, H. & Nakagawa, S. Making conservation science more reliable with preregistration and registered reports. *Conserv. Biol.* **33**, 747–750 (2019).
53. Buxton, R. T. et al. Avoiding wasted research resources in conservation science. *Conserv. Sci. Pract.* **3**, 1–11 (2021).
54. Powers, S. M. & Hampton, S. E. Open science, reproducibility, and transparency in ecology. *Ecol. Appl.* **29**, 1–8 (2019).
55. Archmiller, A. A. et al. Computational reproducibility in the Wildlife Society's flagship journals. *J. Wildl. Manag.* **84**, 1012–1017 (2020).
56. Whitlock, M. C., McPeek, M. A., Rausher, M. D., Rieseberg, L. & Moore, A. J. Data archiving. *Am. Nat.* **175**, 145–146 (2010).
57. Whitlock, M. C. Data archiving in ecology and evolution: best practices. *Trends Ecol. Evol.* **26**, 61–65 (2011).
58. Mislan, K. A. S., Heer, J. M. & White, E. P. Elevating the status of code in ecology. *Trends Ecol. Evol.* **31**, 4–7 (2016).
59. Culina, A., van den Berg, I., Evans, S. & Sánchez-Tójar, A. Low availability of code in ecology: a call for urgent action. *PLoS Biol.* **18**, 1–9 (2020).
60. Wilkinson, M. D. et al. Comment: the FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* **3**, 1–9 (2016).
61. Gopalakrishna, G. et al. Prevalence of responsible research practices among academics in the Netherlands. *F1000Research* **11**, 1–34 (2022).
62. Hardwicke, T. E. et al. Data availability, reusability, and analytic reproducibility: evaluating the impact of a mandatory open data policy at the journal Cognition. *R. Soc. Open Sci.* **5**, 180448 (2018).
63. Stodden, V., Seiler, J. & Ma, Z. An empirical analysis of journal policy effectiveness for computational reproducibility. *Proc. Natl Acad. Sci. USA* **115**, 2584–2589 (2018).
64. Roche, D. G., Kruuk, L. E. B., Lanfear, R. & Binning, S. A. Public data archiving in ecology and evolution: how well are we doing? *PLoS Biol.* **13**, 1–12 (2015).
65. Roche, D. G. et al. Slow improvement to the archiving quality of open datasets shared by researchers in ecology and evolution. *Proc. R. Soc. B* **289**, 20212780 (2022).
66. Lindsey, P. A. et al. The bushmeat trade in African savannas: impacts, drivers, and possible solutions. *Biol. Conserv.* **160**, 80–96 (2013).
67. Roche, D. G. et al. Paths towards greater consensus building in experimental biology. *J. Exp. Biol.* **225**, jeb243559 (2022).
68. Smaldino, P. E. & McElreath, R. The natural selection of bad science. *R. Soc. Open Sci.* **3**, 160384 (2016).
69. R Core Team. R: a language and environment for statistical computing (R Foundation for Statistical Computing, 2019); https://www.R-project.org/
70. Müller, K. here: a simpler way to find your files. R package version 1.0.1 (2017). https://CRAN.R-project.org/package=here
71. Wickham, H., Francois, R., Henry, L. & Muller, K. dplyr: a grammar of data manipulation R package version 1.0.7 (2020). https://CRAN.R-project.org/package=dplyr
72. Wickham, H. & Henry, L. tidyr: tidy messy data R package version 1.1.4 (2020). https://CRAN.R-project.org/package=tidyr
73. Wickham, H. ggplot2: elegant graphics for data analysis (Springer-Verlag, 2016).
74. Kassambara, A. ggpubr: 'ggplot2' based publication ready plots. R package version 0.4.0 (2020). https://CRAN.R-project.org/package=ggpubr
75. Pedersen, T. L. patchwork: the composer of plots. R package version 1.1.1 (2021). https://CRAN.R-project.org/package=patchwork
76. Wickham, H. & Seidel, D. scales: scale functions for visualization. R package version 1.1.1 (2020). https://CRAN.R-project.org/package=scales
77. Bloom, H. S. Minimum detectable effects: a simple way to report the statistical power of experimental designs. *Eval. Rev.* **19**, 547–556 (1995).
78. Djimeu, E. W. & Houndolo, D. G. Power calculation for causal inference in social science: sample size and minimum detectable effect determination. *J. Dev. Eff.* **8**, 508–527 (2016).
79. Havranek, T., Horvath, R. & Zeynalov, A. Natural resources and economic growth: a meta-analysis. *World Dev.* **88**, 134–151 (2016).
80. Stanley, T. D., Carter, E. C. & Doucouliagos, H. What meta-analyses reveal about the replicability of psychological research. *Psychol. Bull.* **144**, 1325–1346 (2018).
81. Parker, T. H. et al. Empowering peer reviewers with a checklist to improve transparency. *Nat. Ecol. Evol.* **2**, 929–935 (2018).
82. Munafò, M. R. et al. A manifesto for reproducible science. *Nat. Hum. Behav.* **1**, 1–9 (2017).
83. Nosek, B. A. et al. Promoting an open research culture. *Science* **348**, 1422–1425 (2015).
84. Nakagawa, S. & Parker, T. H. Replicating research in ecology and evolution: feasibility, incentives, and the cost–benefit conundrum. *BMC Biol.* **13**, 1–6 (2015).
85. Kaplan, R. M. & Irvin, V. L. Likelihood of null effects of large NHLBI clinical trials has increased over time. *PLoS ONE* **10**, 1–12 (2015).
86. Nosek, B. A., Ebersole, C. R., DeHaven, A. C. & Mellor, D. T. The preregistration revolution. *Proc. Natl Acad. Sci. USA* **115**, 2600–2606 (2018).
87. Allen, C. & Mehler, D. M. A. Open science challenges, benefits and tips in early career and beyond. *PLoS Biol.* **17**, 1–14 (2019).
88. Scheel, A. M., Schijen, M. R. M. J. & Lakens, D. An excess of positive results: comparing the standard psychology literature with registered reports. *Adv. Methods Pract. Psychol. Sci.* **4**, 1–12 (2021).
89. Nosek, B. A. et al. Preregistration is hard, and worthwhile. *Trends Cogn. Sci.* **23**, 815–818 (2019).
90. Button, K. S., Bal, L., Clark, A. & Shipley, T. Preventing the ends from justifying the means: withholding results to address publication bias in peer-review. *BMC Psychol.* **4**, 1–7 (2016).
91. Soderberg, C. K. et al. Initial evidence of research quality of registered reports compared with the standard publishing model. *Nat. Hum. Behav.* **5**, 990–997 (2021).
92. Smulders, Y. M. A two-step manuscript submission process can reduce publication bias. *J. Clin. Epidemiol.* **66**, 946–947 (2013).
93. Anderson, M. S., Martinson, B. C. & De Vries, R. Normative dissonance in science: results from a national survey of U.S. scientists. *J. Empir. Res. Hum. Res. Ethics* **3**, 3–14 (2007).

Acknowledgements
We thank the Glenadore and Howard L. Pim Postdoctoral Fellowship in Global Change for funding K.K. We thank T. Parker for his helpful comments on revising the manuscript. We thank M. Buchanan, P. Dye, Z. Ellis, Y. Li, L. Wang and L. Williams for helping in the data collection for this paper. We thank P. Shukla for providing sample code for the analyses.

Author contributions
P.J.F. and K.K. designed the study. K.K. analysed the data. M.L.A., P.J.F. and K.K. wrote the paper.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41559-023-02144-3.

Correspondence and requests for materials should be addressed to Paul J. Ferraro.

Peer review information *Nature Ecology & Evolution* thanks Timothy Parker, Antica Culina, Dominique Roche and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

nature portfolio

Corresponding author(s): Paul Ferraro

Last updated by author(s): Jun 13, 2023

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	All data was collected from manuscripts published between 2018 and 2020 in Science, Nature, Journal of Ecology, Ecology, and Ecology Letters.
Data analysis	All data manipulation were done in R version 4.0.0 70, and we utilized the 'here' package (version 1.0.1) for replicability. Throughout our script, we used dplyr (version 1.0.7) and tidyr (version 1.1.4) to manipulate our data. We also relied on ggplot2 (version 3.3.5), ggpubr (version 0.4.0), patchwork (version 1.1.1), and scales (version 1.1.1) for making figures. Our code is available at https://osf.io/9yd2b/?view_only=d3e18f3437bf49289cc5448d9e5a2e36.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

Our dataset is available at https://osf.io/9yd2b/?view_only=d3e18f3437bf49289cc5448d9e5a2e36.

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	NA
Reporting on race, ethnicity, or other socially relevant groupings	NA
Population characteristics	NA
Recruitment	NA
Ethics oversight	NA

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☐ Behavioural & social sciences ☒ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	We conducted a meta-science study whereby we estimated statistical power, exaggeration bias, and selective reporting from estimates reported in published studies.
Research sample	Our sample comprised peer-reviewed empirical ecology manuscripts published between 2018 and 2020 in Science, Nature, Ecology, Journal of Ecology, and Ecology Letters.
Sampling strategy	We collected estimates from empirical papers where data were reported in tables. We aimed to get a sample size >15,000 estimates so that our analysis would be robust based on a similar paper published in environmental economics.
Data collection	Data was recorded from empirical studies that statistically estimated parameters. Means and some measure of error needed to be reported. Data were collected by six undergraduate research assistants.
Timing and spatial scale	We collected data from papers published between 2018-2020.
Data exclusions	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.
Reproducibility	This is not applicable to our study.
Randomization	This is not applicable to our study.

Blinding | This is not applicable to our study.

Did the study involve field work? ☐ Yes ☒ No

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging