Trends in
Ecology & Evolution

CellPress
OPEN ACCESS

Opinion

Rewriting results sections in the language of evidence

Stefanie Muff [1,2,*,@], Erlend B. Nilsen,[2,3,4,@] Robert B. O'Hara,[1,2,@] and Chloé R. Nater[2,3,@]

Despite much criticism, black-or-white null-hypothesis significance testing with an arbitrary *P*-value cutoff still is the standard way to report scientific findings. One obstacle to progress is likely a lack of knowledge about suitable alternatives. Here, we suggest language of evidence that allows for a more nuanced approach to communicate scientific findings as a simple and intuitive alternative to statistical significance testing. We provide examples for rewriting results sections in research papers accordingly. Language of evidence has previously been suggested in medical statistics, and it is consistent with reporting approaches of international research networks, like the Intergovernmental Panel on Climate Change, for example. Instead of re-inventing the wheel, ecology and evolution might benefit from adopting some of the 'good practices' that exist in other fields.

Highlights

It has been known for decades that there are severe problems associated with null-hypothesis significance testing (NHST) based on arbitrary *P*-value thresholds (e.g., $P = 0.05$).

A small literature review indicates that much of the current research in ecology and evolution is still disregarding the warnings and frequently relies on binary decisions based on *P*-values to report statistical significance.

While the *P*-value itself is a sound mathematical concept that does not have to be banned when used correctly, we should stop using the term 'statistical significance' and replace it with a gradual notion of evidence.

Language matters and 'evidence' is an intuitive concept that honestly reflects what the data really tell us.

To facilitate rewriting scientific results, we offer generic examples of how to translate (binary) significance language into a gradual language of evidence.

The century-long debate around the *P*-value

The *P*-value is probably the most commonly used and yet the most hotly debated statistical measure employed for the interpretation of quantitative research outcomes (e.g., [1–5]). The *P*-value is, in essence, the main ingredient in null-hypothesis significance testing (NHST), where the existence of an effect of interest is evaluated following a recipe-like procedure. In the almost 100-year-long history of *P*-values, the respective practice of NHST has continually been criticized in literally hundreds of articles (e.g., [6–12]). Eventually, statisticians shocked their audience with articles entitled 'Why most published research findings are false' [10] and 'The statistical crisis in science' [13], essentially condemning the reliance on *P*-values and NHST to assess the statistical significance of effects. One key problem is the mistaking of statistical significance for scientific importance, even though the myth that lower *P*-values automatically imply higher relevance was debunked a long time ago (e.g., [8,14]). In addition, the *P*-value is often misinterpreted (see for instance [14] for a list of 12 *P*-value misconceptions), illustrating that understanding what the *P*-value actually means is not as simple as it seems. Formally, the *P*-value is the probability of observing an outcome that is at least as extreme as an observed data summary, under the assumption that a certain hypothesis, the so-called null hypothesis (H_0), is true (Box 1). The null hypothesis thereby implies that a specific mathematical model is correct, for example that the data are normally distributed with a prespecified mean. In NHST we can only do two things: we can either reject H_0 or we can not reject it. If we cannot reject H_0, that is, when *P* lies above a predefined threshold (usually $P > 0.05$), it is incorrect to conclude that '...there was no effect...' or that 'the null hypothesis is true'. In fact, H_0 cannot be proven and 'absence of evidence is not evidence of absence' [15], reflecting that NHST is an intrinsically asymmetric procedure. Similarly, the *P*-value is often interpreted as the probability that H_0 is true, a misconception that is persistent despite it having been pointed out repeatedly (e.g., [8,11,14]).

More recently, several high-profile papers have brought the same old controversy to the attention of the broader scientific community [3,4,12,16]. The discussion was boosted by a statement on

[1]Department of Mathematical Sciences, Norwegian University of Science and Technology NTNU, 7491 Trondheim, Norway
[2]Centre for Biodiversity Dynamics, Norwegian University of Science and Technology NTNU, 7491 Trondheim, Norway
[3]The Norwegian Institute for Nature Research (NINA), 7485 Trondheim, Norway
[4]Nord University, Faculty of Bioscience and Aquaculture, 7713 Steinkjer, Norway

*Correspondence: stefanie.muff@ntnu.no (S. Muff).
[@]Twitter: @stefaniemuff (S. Muff), @eb_nilsen (E.B. Nilsen), @BobOHara (R.B. O'Hara), and @chloe_nater (C.R. Nater).

Box 1. The *P*-value

Definition: the *P*-value is the probability of observing a specific data summary (e.g., an average) that is at least as extreme as the one observed, given that the null hypothesis (H_0) is correct.

Example: a prominent example is a case where H_0 assumes that a certain data summary (denoted as test statistic) has a standard normal distribution. Given that the observed value z of the test statistic is derived from the data, the *P*-value is thus the probability that we would see such an extreme, or an even more extreme, value given that H_0 was in fact true. The *P*-value thus reflects how likely it is that we see a specific outcome if H_0 holds.

The graphical example in Figure I shows the meaning of the *P*-value, once for an observed value of $z = 1.96$ for a relatively clear positive effect (left) and once for an observed value $z = -0.84$ for a negative, but less clear, effect. The shaded areas under the curve represent the *P*-values, that is, the probabilities that the observed values of z or more extreme values occur under H_0. The *P*-value for $z = 1.96$ is thus $P = 0.025 + 0.025 = 0.05$, and the *P*-value for $z = -0.84$ is $P = 0.2 + 0.2 = 0.4$.

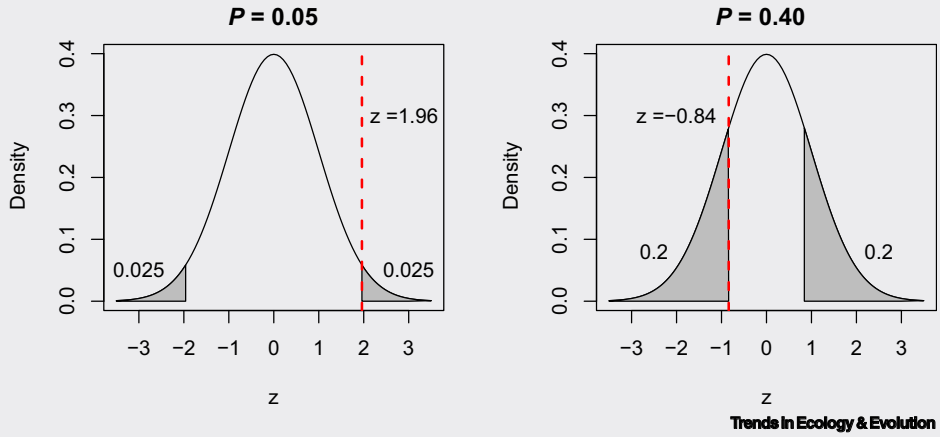

Figure I. Graphical illustration of the *P*-value for two different values of the test statistics z in the two-sided case.

The asymmetric nature of the *P*-value: one major misunderstanding about the *P*-value is that it is often believed to say something about the probability that H_0 is true. This is not the case. Instead, the probability that H_0 is true, given a certain data summary (i.e., a test statistic) from the data, is given as:

$$P(H_0 \mid \text{data summary}) = \frac{P(\text{data summary} \mid H_0) \cdot P(H_0)}{P(\text{data summary})} \quad [\text{I}]$$

This implies that we would have to specify a prior guess for P(H_0). Even if this was possible, we would need to calculate the prior density for the observed data summary P(data summary), which is not trivial in most circumstances.

the misuse and misinterpretation of the *P*-value and statistical significance that the American Statistical Association (ASA) published in March 2016 [11]. This was the first time in the history of the Association that such a policy statement had been released, underlining the importance the ASA Board assigned to the topic.

The confusion around the *P*-values' use is exacerbated by the fact that, ironically, it is not actually the definition of the *P*-value that is the problem. Rather, the issues arise from the way the *P*-value is used in NHST to make binary decisions (significant versus nonsignificant, there is an effect versus there is no effect) based on a sharp, arbitrary cutoff, typically $P = 0.05$ (though recent arguments speak for lower limits, see [5]). When it was originally developed, the *P*-value was indeed not meant to be used the way it is used today. Fisher, who suggested the *P*-value [17], used the term 'significance' only to indicate that an observed outcome was worth closer investigation,

and emphasized that H_0 would be rejected only if follow-up experiments also 'rarely failed to achieve significance', while he opposed using the *P*-value for 'automatic inference' (see, e.g., [4]).

No agreement on best practices in sight

Long-term misuse of *P*-values has fostered questionable research practices (e.g., [18]), like *P*-hacking, model selection based on *P*-values, and hypothesizing after the results are known (HARKING). Combining these with the virtually unlimited degrees of freedom researchers have in building models and stating assumptions, and the tendency to publish 'significant' results more often than 'nonsignificant' ones (i.e., publication bias), has led to a flood of false positive findings that has contributed to a severe scientific reproducibility crisis (see, e.g., [10,19,20]). However, guidelines and solutions for appropriate use of the *P*-value are still hotly debated and no agreement on a way forward is in sight ([5,11,12,21] to mention just a few ongoing discussions).

In the wake of the debate, we observe in our everyday collaborations that many applied scientists have become uncertain about how to report their findings and some hardly dare to report *P*-values anymore. The confusion is also reflected, for example, by the choice of the editors of the journal *Basic and Applied Social Psychology* to ban *P*-values [22]. However, abolishing the *P*-value would be a case of throwing the baby out with the bath water. Despite repeated misuse, many statisticians still believe that the *P*-value is a very informative statistical index when interpreted correctly [23].

Alternatives to using the *P*-value have, of course, existed for a long time. The most prominent examples are information criteria like the Akaike or Bayesian information criterion (AIC, BIC), Bayes factors, or confidence intervals (CIs) (see, e.g., [24] for an overview). However, when these alternatives are used to make binary decisions, for example regarding the inclusion of variables in model selection, when checking whether the null effect lies in a CI, or when employing a certain threshold of a Bayes factor, we are not doing anything different from an NHST. It can, for example, be shown that model selection based on the AIC criterion can be converted into *P*-value limits (e.g., [2]), and even Bayes factors have an approximate equivalent in terms of *P*-values [4,25,26]. Finally, checking whether a certain value (often 0) lies outside the CI is equivalent to checking the *P*-value limit, like $P < 0.05$ if the 95% CI is used, for example.

Did reporting behavior change?

Has the debate had an impact on how we report and interpret our findings in the ecology and evolution research community? In order to get a better feeling for this question, we carried out a small literature review. We used the January 2021 issues (December 2020 if January 2021 was a special issue) of eight major journals in ecology and evolution and checked all research papers containing at least one statistical analysis ($n = 137$, see the supplemental information online). Of those, 113 (82.5%) reported results based on the NHST philosophy: 104/113 (92%) of the dichotomous decisions were based on the *P*-value, while seven used the 95% CIs, and two used an information criterion. A total of 110/113 (97.3%) reported their findings using the 'significance' terminology. It appears as if the decades with waving warning flags had relatively little impact on the routines in our field when it comes to writing the results sections of scientific papers.

The gradual evidence language: a simple proposal

So, how can we do better? And what should we teach as good practice to the next generation of ecologists and evolutionary biologists? There is ample agreement, maybe the lowest common denominator of the whole discussion, that we should retire statistical significance by eliminating binary decision making from most scientific papers [12]. Such a transition will take time, but the show must go on today and we urgently need simple and safe ways to bypass the current

state of disorientation. If not, researchers might remain stuck in old habits, as our literature review suggests.

A central aspect of an alternative reporting standard is that it must be immediately applicable by anyone in the field. We propose that one of the easiest and most practical measures would be to replace the wording around binary decision making by a more gradual notion of evidence, which better reflects the actual information provided by the data. We can learn from those that have thought about these issues for over 30 years, such as medical statisticians. Instead of introducing a new terminology (see, e.g., [21]), we could take advantage of tried and tested practices from fields that have already progressed further in the debate. Very useful guidelines were, for instance, given in an introductory medical statistics book that first appeared in 1986 [27,28], where it was suggested to regard *P*-values as what they are, namely, continuous measures of statistical evidence (Figure 1). Instead of reporting a binary yes/no test outcome, the results sections of scientific papers should rather report the exact *P*-values and interpret that 'there was no/weak/moderate/strong/very strong evidence' for a certain finding or effect, depending on approximate ranges into which the actual *P*-value falls (Figure 1). In Tables 1 and 2 we present some generic and real examples for how statements in results sections may be adapted from statistical significance terminology to the language of evidence.

There are several reasons why we think that the notion of 'evidence' is more appropriate than 'significance'. Most prominently, the notion of (accumulated) evidence is the main concept behind meta-analyses. Meta-analyses became popular in the 1970s in medical research [29], but are nowadays used for aggregating the essence of previous research in all scientific fields that learn from data (e.g., [30,31]). International research networks like the enhancing the quality and transparency of health research (EQUATOR) network in the context of medicine, epidemiology, and health (https://www.equator-network.org/) and the Intergovernmental Panel on Climate Change (https://www.ipcc.ch/) have clear guidelines on how to carry out meta-analyses, which result in so-called synthesis reports. In the meta-analysis philosophy, each single study is contributing one piece of evidence to the global knowledge in a cumulative manner. It is then irrelevant

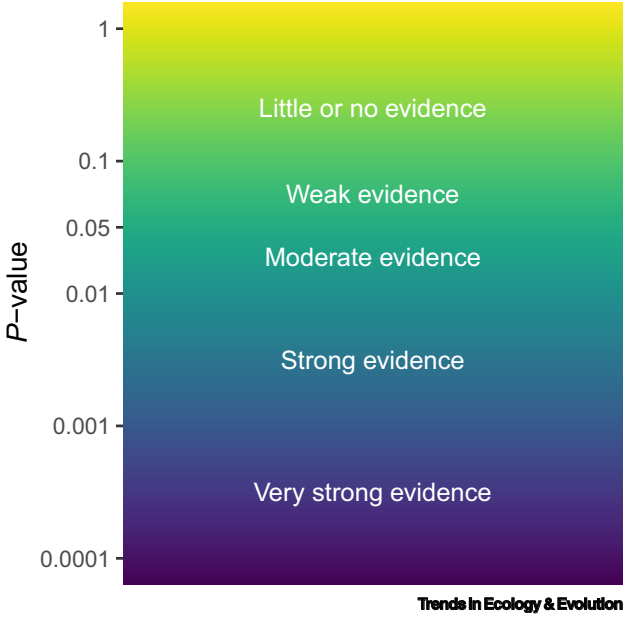

Figure 1. Suggested ranges to approximately translate the *P*-value into the language of evidence. The ranges are based on Bland (1986) [27], but the boundaries should not be understood as hard thresholds.

Table 1. Generic examples of how to rewrite results from the statistical significance to evidence-based language[a]

Statistical significance language	Evidence language (suggestions)
The effect of *x* on *y* was not significant ($P = 0.53$).	There was no evidence that *x* has an effect on *y* [(give effect estimate), $P = 0.53$].
	The data did not have any evidence about the direction of any association of *x* with *y* [(give effect estimate), $P = 0.53$].
The effect of *x* on *y* was not significant ($P = 0.08$).	There was (only) weak evidence that *x* (positively/negatively) affects *y* [(give effect estimate), $P = 0.08$].
	There was (only) weak evidence that *x* is (positively/negatively) associated with *y* [(give effect estimate), $P = 0.08$].
The effect of *x* on *y* was significant ($P = 0.03$).	There was (only) moderate evidence that *x* has a (positive/negative) effect on *y* [(give effect estimate), $P = 0.03$].
	The data revealed moderate evidence that *x* is (positively/negatively) associated with *y* [(give effect estimate), $P = 0.03$].
The effect of *x* on *y* was significant ($P = 0.0003$).	There was very strong evidence for a (positive/negative) effect of *x* on *y* [(give effect estimate), $P < 0.001$].
	The data revealed very strong evidence that *x* is (positively/negatively) associated with *y* [(give effect estimate), $P < 0.001$].

[a]Note that we recommend reporting $P < 0.001$ if that is the case. A general recommendation is that *P*-values should be accompanied by effect size estimates whenever possible. The remark whether an effect was positive or negative should be added when this is possible (e.g., for continuous variables *x*).

whether an individual study was 'significant' or not. Rather, all existing effect sizes and standard errors are integrated into a new estimator with a respective uncertainty. When we see our own study (with potentially quite limited sample size) as a contribution to the global scientific effort to explain, learn, and understand through the accumulation of evidence and knowledge, we might understand even better that evidence-based language is more appropriate than significance testing. Binary decision making can then be reserved for contexts where we do need practical reasons to act [32], like in drug development, public health, or policy making in ecosystem management and

Table 2. Text examples from papers published in the December 2020 issue of the journal *Evolution* and in some of the publications included in the literature review

Published text example	Rewriting suggestion using evidence language
Glider and arborealist disparities are not significantly different ($P = 0.44$).	There is no evidence that glider and arborealist disparities differ [(give effect estimate), $P = 0.44$].
We found no significant differences between hypercarnivorous and generalist species for the shape of the cranium ($F = 1.07$, $P = 0.34$).	There was no evidence that the shape of the cranium is different between hypercarnivorous and generalist species ($F = 1.07$, $P = 0.34$).
By contrast, we found significant shape differences, mainly related to bone robustness, for the humerus ($F = 3.13$, $P = 0.022$) and the femur ($F = 2.7$, $P = 0.017$).	By contrast, there was moderate evidence for shape differences, mainly related to bone robustness, for the humerus ($F = 3.13$, $P = 0.022$) and the femur ($F = 2.7$, $P = 0.017$).
Our results revealed significant disparity differences between generalist and hypercarnivorous species for the cranium ($P = 0.002$) and the mandible ($P = 0.006$).	There was strong evidence for disparity differences between generalist and hypercarnivorous species for the cranium [(give effect estimate), $P = 0.002$] and the mandible [(give effect estimate), $P = 0.006$].
(...) we show here that body size decreased significantly in the treatments ($F_{3,\ 7710} = 76.30$, $P < 2.20 \cdot 10^{-16}$).	(...) there was very strong evidence that body size decreased in the treatments ($F_{3,\ 7710} = 76.30$, $P < 0.001$).
I_A was affected by conditions in males ($P = 8.87 \cdot 10^{-5}$) but not in females ($P = 0.07$).	There was very strong evidence that I_A was (positively/negatively) affected by conditions in males [(give effect estimate), $P < 0.001$], but only weak evidence that this was the case in females [(give effect estimate), $P = 0.07$].
Foliar 10% did not significantly increase production of extrafloral nectar (estimate = –0.13, $P = 0.061$).	There was (only) weak evidence that foliar 10% increased production of extrafloral nectar (estimate = –0.13, $P = 0.061$).
The relationship between mean light transmittance and basal area was not significant ($R^2_{adj} = 0.146$, $P = 0.168$, $n = 9$), but light transmittance decreased slightly with diameter at breast height (DBH) of transplant trees across sites ($R^2_{adj} = 0.022$, $P < 0.014$, $n = 225$).	There was no evidence for a relationship between mean light transmittance and basal area ($R^2_{adj} = 0.146$, $P = 0.168$, $n = 9$), but moderate evidence that light transmittance decreased slightly with DBH of transplant trees across sites [$R^2_{adj} = 0.022$, $P =$ (give exact *P*-value), $n = 225$].
The sex ratio for immigrants was female biased (58.9% females, $n = 569$, binomial test $P < 0.001$) in wandering albatrosses (but not for residents: 49.7%, $n = 2844$, binomial test $P = 0.750$).	There was very strong evidence that the sex ratio for immigrants was female biased (58.9% females, $n = 569$, binomial test $P < 0.001$) in wandering albatrosses, but there was no evidence for such a bias for residents (49.7%, $n = 2844$, binomial test $P = 0.75$).
There was no difference detected among contemporary Great Lakes and East Coast anadromous alewives (ANOVA: $F_{2,224} = 2.74$, $P = 0.067$).	There was (only) weak evidence that contemporary Great Lakes and East Coast anadromous alewives differ (ANOVA: $F_{2,224} = 2.74$, $P = 0.067$).

conservation (see Outstanding questions; also several of the comments to the ASA statement by [11]).

Of course, given that 'no single index should substitute for scientific reasoning' [11], it is crucial that we always try to understand the relevance and implications of a given result. It is rarely enough to know that there is some level of evidence against a hypothesis. Instead, we should also assess whether effects are important, but maybe just too poorly estimated, by looking at their uncertainty. In this way, we can, for example, assess weak evidence against the null hypothesis to see if more data are needed for a clear assessment. A minimal requirement thus is that we report effect sizes, CIs, and (if applicable) Bayes factors. In addition, we strongly recommend that researchers truly attempt to interpret the biological meaning and implications of their quantitative findings, for example by giving numerical examples and/or graphical descriptions of how a variable affects an outcome and how uncertain those findings are.

Concluding remarks

There is hopefully not much doubt that it is time to move away from the cult around binary decision making and statistical significance. Intuitively, we have all known that it cannot really matter whether $P = 0.049$ or $P = 0.051$. Here we are promoting a relatively simple guide to replacing cutoff-based decision-making by a gradual language of evidence. The change in terminology suggested here might help us to see our results as what they truly are, namely, as pieces of information in the context of cumulative science. Nonetheless, we are aware that any type of guidelines bear the danger of being applied as a recipe to follow blindly, which may entail a reduction of critical reflection. The most important thing remains that we do not, by any means, do mindless statistics [33]. We can therefore not stress enough how important it is that any result is interpreted, not only based one a single index (like the *P*-value, AIC, etc.), but under consideration of the context in which the research is actually intended to make an impact. We can, for example, illustrate how an increase in mean temperature by 0.5°C or a change in the inbreeding coefficient by 0.1 units affects expected outcomes, such as the abundance of a species or the predicted viability of a population. By plugging in concrete values (and the associated uncertainties), we can communicate our findings in a critically reflected and meaningful way. In addition, a graphical representation usually helps make the findings intuitively more accessible and can be worth a thousand statistics [34].

Will a seemingly trivial change from a language built around binary statistical significance to a more continuous language of evidence make a real change? We think so, because by reporting results in this way, we automatically move away from drawing unfounded binary conclusions. At the same time, we can free ourselves from hunting arbitrary cutoffs that magically determine whether our research was a success or a failure.

Another suggestion to rewrite results has recently been made [21]. However, they argued for the replacement of statistical significance by 'statistical clarity'. The underlying idea is that the terms statistically 'clear' or 'unclear' slightly better reflect the actual information contained in a *P*-value. By saying that a result is 'statistically unclear' when $P > 0.05$, for example, we do not imply that the respective effect does not exist, in contrast to the misconception that a nonsignificant effect is absent. The respective suggestion is definitely valid, and complements our suggestion, but we believe that the evidence-based language has three key advantages. First, the term 'evidence' is better suited to reflect the information content in our data and it allows for a nuanced interpretation via the gradual shifts from very strong to no evidence. Second, evidence-based language has been around for a long time and we believe that statistical terminology should be as consistent as possible across fields. And third, evidence is a very intuitive concept that may help our increasingly

Outstanding questions

The future of NHST: (when) is it still useful? NHST and similar methods of asymmetric binary decision making might be justified when decisions are needed based on only one or a few studies, but the underlying studies then need to fulfill very high-quality standards. In drug development, for example, we request carefully planned, prospective randomized controlled trials that are based on a long list of requirements, including sample size calculations, preregistration, statistical analysis plans, etc.

How do we make decisions based on evidence? When each paper only contributes a piece of evidence in the cumulative process of creating knowledge, scientists should run meta-analyses that pull the information from the literature together. The responsibility for making decisions may then be returned to the 'practical decision makers', but the respective translation requires training and, ideally, an ongoing dialogue with the researchers.

How will we measure scientific success in the future? When the potential to publish a 'significant' result falls away and science is rather seen as a joint effort to accumulate knowledge, we might also have to rethink the measures of scientific success.

How do we break our habits? Statistical significance is what we teach because we use it, and it is what we use because we teach it. To break the cycle, our habits need a change at both levels simultaneously; also, journals have an obligation to act.

more open research to be more accessible to broader audiences, such as non-scientists, including the public, stakeholders, and media.

Of course, simply replacing NHST and statistical significance with a language of evidence will not automatically overcome all the fundamental issues of how we generate, report, and think about scientific results. The replacement of NHST and statistical significance by a language of evidence is one stone in the mosaic of the reforms that science urgently needs. Other issues such as preregistration, model selection, differences in handling exploratory and confirmatory analyses, and the distinction between experimental and observational studies (e.g., [35,36]) are also important topics where a deeper knowledge of the methodological issues will improve the reproducibility, replicability, and understanding of scientific results. In addition, the discussion inevitably raises the question about how scientific achievements of individuals should be measured when the possibility to statistically 'prove' new findings falls away (see Outstanding questions). These topics deserve further assessment and we hope we could at least stimulate the discussion and help authors overcome a potential *P*-value paralysis. Most importantly, reporting scientific results using evidence language instead of significance testing is a straightforward step anyone can take immediately to move ecology and evolution forward and help overcome the reproducibility crisis.

Acknowledgments

We thank Jonathan Dushoff and a second, anonymous reviewer for their comments that helped improve our opinion piece.

Declaration of interests

No interests are declared.

Supplemental information

Supplemental information associated with this article can be found online https://doi.org/10.1016/j.tree.2021.10.009.

References

1. Goodman, S.N. (1999) Toward evidence-based medical statistics. 1: the P value fallacy. *Ann. Intern. Med.* 130, 995–1004
2. Murtaugh, P.A. (2014) In defense of P values. *Ecology* 95, 611–617
3. Nuzzo, R. (2014) Statistical errors. *Nature* 506, 150–152
4. Goodman, S.N. (2016) Aligning statistical and scientific reasoning. *Science* 352, 1180–1182
5. Benjamin, D. *et al.* (2017) Redefine statistical significance. *Nat. Hum. Behav.* 2, 6–10
6. Berkson, J. (1942) Tests of significance considered as evidence. *J. Am. Stat. Assoc.* 219, 325–335
7. Rozenboom, W.W. (1960) The fallacy of the null-hypothesis significance test. *Psychol. Bull.* 57, 416–428
8. Cox, D.R. (1982) Statistical significance tests. *Br. J. Clin. Pharmacol.* 14, 325–331
9. Cohen, J. (1994) The earth is round ($p < .05$). *Am. Psychol.* 49, 997–1003
10. Ioannidis, J.P.A. (2005) Why most published research findings are false. *PLoS Med.* 2, e124
11. Wasserstein, R.L. and Lazar, N.A. (2016) The ASA's statement on p-values: context, process, and purpose. *Am. Stat.* 70, 129–133
12. Amrhein, V. *et al.* (2019) Retire statistical significance. *Nature* 567, 305–307
13. Gelman, A. and Loken, E. (2014) The statistical crisis in science. *Am. Sci.* 102, 460–465
14. Goodman, S.N. (2008) A dirty dozen: twelve P-value misconceptions. *Semin. Hematol.* 45, 135–140
15. Altman, D.G. and Bland, J.M. (1995) Absence of evidence is not evidence of absence. *Br. Med. J.* 311, 485
16. Claridge-Change, A. and Assam, P.N. (2016) Estimation statistics should replace significance testing. *Nat. Methods* 13, 108–109
17. Fisher, R.A. (1926) The arrangement of field experiments. *J. Minist. Agric.* 33, 503–515
18. Fraser, H. *et al.* (2018) Questionable research practices in ecology and evolution. *PLoS One* 13, e0200303
19. Peng, R.D. (2015) The reproducibility crisis in science. *Signif. (Oxf)* 12, 30–32
20. Baker, M. (2016) 1500 scientists lift the lid on reproducibility. *Nature* 533, 452–454
21. Dushoff, J. *et al.* (2019) I can see clearly now: reinterpreting statistical significance. *Methods Ecol. Evol.* 10, 756–759
22. Trafimow, D. and Marks, M. (2015) Editorial. *Basic Appl. Soc. Psych.* 37, 1–2
23. Spiegelhalter, D. (2017) Too familiar to ditch. *Signif. (Oxf)* 14, 41
24. Halsey, L.G. (2019) The reign of the p-value is over: what alternative analyses could we employ to fill the power vacuum? *Biol. Lett.* 15, 20190174
25. Goodman, S.N. (2001) Of p-values and Bayes: a modest proposal. *Epidemiology* 12, 295–297
26. Held, L. and Ott, M. (2018) On *p*-values and Bayes factors. *Annu. Rev. Stat. Appl.* 5, 393–419
27. Bland, J.M. (1986) *An Introduction To Medical Statistics*, Oxford Medical Publications, New York
28. Bland, M. (2015) *An Introduction to Medical Statistics*, Oxford University Press
29. Haidich, A. (2010) Meta-analysis in medical research. *Hippokratia* 14, 29–37
30. Cooper, H. *et al.* (2009) *The Handbook of Research Synthesis and Meta-Analysis*, Russell Sage Foundation
31. Koricheva, J. *et al.* (2013) *Handbook of Meta-Analysis in Ecology and Evolution*, Princeton University Press
32. Poole, C. (1987) Beyond the confidence interval. *Am. J. Public Health* 77, 195–199

33. Gigerenzer, G. (2004) Mindless statistics. *J. Socio. Econ.* 33, 587–606
34. Fagerlin, A. *et al.* (2005) Reducing the influence of anecdotal reasoning on people's health care decisions: is a picture worth a thousand statistics? *Med. Decis. Mak.* 25, 398–405
35. Nilsen, E.B. *et al.* (2020) Exploratory and confirmatory research in the open science era. *J. Appl. Ecol.* 57, 842–847
36. Tredennick, A.T. *et al.* (2021) A practical guide to selecting models for exploration, inference, and prediction in ecology. *Ecology* 102, e03336