Trends in
Ecology & Evolution

CellPress

Letter

Joint reply to 'Rewriting results in the language of compatibility' by V. Amrhein and S. Greenland, and to 'The evidence contained in the *P*-value is context dependent' by F. Hartig and F. Barraquand

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

In our recent opinion article we promoted an approach for interpreting *P*-values as continuous but informal measures of evidence [1]. Our aim was to propagate a pragmatic way forward for those who are willing to leave significance testing behind. Instead of advocating for a total change of approach, which is less likely to be effective, evidence language is easy to use. The potential of this approach is acknowledged by [2] as doing more good than harm.

Compatibility intervals

The problem of how to use *P*-values (and indeed whether they should be used at all) has been discussed for a long time. One reason for these discussions is the weirdness of what *P*-values are meant to do. We are usually interested in a so-called alternative hypothesis H_1 (e.g., that there 'is an effect'), but test it by calculating the probability of obtaining a test statistic at least as extreme as the one observed if the null hypothesis H_0 were true, without explicit reference to H_1 in the calculations. We therefore have to insert H_1 into the inference, that is, it is part of the interpretation of the test. Amrhein and Greenland [2] resolve this issue by suggesting that *P*-values only summarize compatibility with H_0. A significance test is then seen as goodness of fit to H_0, although under the unrealistic premise that our model specification is completely right [3]. They suggested that compatibility language thus solves two problems: it relaxes the assumption of correct model specification and at the same time side-steps the problem of H_1. In practice, this would mean that we report effect sizes, the *P*-value and the 'compatibility interval' (commonly referred to as the confidence interval). This is a fair proposal, as long as we resist the temptation to draw binary decisions based on whether null is in the respective interval. We see compatibility terminology as a complement to evidence language: even if researchers switch from significance and confidence to compatibility, they will also continue to discuss evidence; hence we need to implement additional changes to report and interpret scientific results.

Keeping the status quo?

Hartig and Barraquand [4] take a different approach; they seem to defend (or at least accept) the status quo, both in terms of $P < 0.05$ and the (in)famous old 'star-notation'. They argue that 'statistical significance...is now a term with a well-defined meaning for researchers with appropriate statistical training', and that, thanks to the ample training researchers nowadays receive, the misconception that *P* cannot be translated into the probability of an effect has been effectively eliminated. Sadly, there is strong evidence that the statistical training of the past decades has failed miserably if the aim was to avoid misinterpretations. If that was not the case, we would probably not be having the current discussion in this letter exchange. Even if the statistical training had been as fruitful as claimed [4], we do not see why a more continuous interpretation in terms of evidence would suddenly make people forget to tell students that a *P*-value is not the probability of an effect.

Hartig and Barraquand do, correctly, point out that the probability that a 'significant' *P*-value is a false positive depends on the power of the test. It is then, however, not meaningful to state that 'the significance threshold $P < 0.05$ is...chosen as a compromise between type I and II error rates', since such a compromise would, consequently, have to account for the power of the test, in addition to the costs of making a type I or II error in the respective study.

We acknowledge that the dependency on power remains when reporting results using evidence language. For this reason we have repeatedly stressed two key aspects. First, *P*-values should never be interpreted in a vacuum. As we wrote, '... it is crucial that we always try to understand the relevance and implications of a given result.'. Second, 'little or no evidence' should not be understood as the absence of an effect [5], that is, 'absence of evidence' must not be mixed up with 'evidence of absence' [6]. A large *P*-value is, in fact, often the result of low power.

Is the Bayes factor the solution?

As an alternative to *P*-values, Hartig and Barraquand promote the use of another index: the Bayes factor (BF). Their argument is that this better represents the amount of evidence in a test. However, unlike evidence language, BFs are not a solution that is 'immediately applicable by anyone in the field', and they have never become popular, despite existing for decades. Not a single one of the 137 publications in our literature review [1] reported a BF. The BF also carries with it a greater burden: its calculation is often not trivial, and it is well known to be sensitive to priors (e.g., [7]). The propagated equality of the posterior odds with the BF, for example, only holds under the very strong assumption taken by [4] that the prior odds are 50:50, which is quite unrealistic [8]. Even worse, two priors that result in approximately the same posterior can induce two

very different BFs – but from a Bayesian perspective, it is only really the posterior that matters.

Hartig and Barraquand are, in principle, correct in saying that the evidence (in terms of the BF) of a *P*-value depends on sample size *n*. In practice, however, there are several problems in this reasoning. First, it has been shown that this only matters if *n* is really small [9]. Drawing the reader's attention to a simulation with $n = 13$ to claim that 'most observations of $P > 0.1$ would translate into $Pr(H_0|D) = 1–Pr(H_1|D) > 50\%$' is highly tendentious and thus misleading. Second, given that 'little or no evidence' does not imply the absence of an effect, it is in fact not surprising that a highly underpowered study (e.g., $n = 13$) tends to not reveal evidence for an effect due to the large uncertainty in the estimates. Third, and maybe most importantly, the whole argument is based on assuming the BF to be some sort of 'gold standard', while we should rather accept that BFs and *P*-values are different measures of evidence, each with their own flaws and shortcomings (see e.g., [10,11]).

No foolproof solution exists

We reiterate that evidence is not a mathematical concept, irrespective of whether the context is a *P*-value or a BF. Since it is unlikely that the century-old *P* value debate will come to a harmonic, universal agreement, there are only two choices: the status quo versus a small and pragmatic step forward. Since all pragmatism has its shortcomings, any suggestion for change attracts critique. As we have stressed in our opinion article as well as in our previous reply letter [1,5], the most important thing remains that we do not do 'mindless statistics' [12]. No matter whether we report *P*-values, BFs, effect estimates, confidence, credible or compatibility intervals, the results must be put into context for valid interpretation. The language of evidence with approximate, gradual levels may be misunderstood or overinterpreted, but like others before us have concluded, no foolproof solution exists [11]. To us, the most urgent step forward is to finally leave significance testing behind; thus, if we want to continue using *P*-values, we should at least use a gradual instead of a dichotomous scale. But whatever we do, staying open and modest [2] seems like the only universal recipe to follow.

[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), @BobOHara (R.B. O'Hara), @eb Nilsen (E.B. Nilsen) and @chloe nater (C.R. Nater).

https://doi.org/10.1016/j.tree.2022.03.007

References

1. Muff, S. *et al.* (2022) Rewriting results sections in the language of evidence. *Trends Ecol. Evol.* 37, 203–210
2. Amrhein, V. and Greenland, S. (2022) Rewriting results in the language of compatibility. *Trends Ecol. Evol.* 37, 567–568
3. Box, G.E.P. (1980) Sampling and Bayes' inference in scientific modelling and robustness. *J. R. Stat. Soc. Ser. A* 143, 383–430
4. Hartig, F. and Barraquand, F. (2022) The evidence contained in the *P*-value is context dependent. *Trends Ecol. Evol.* 37, 569–570
5. Muff, S. *et al.* (2022) Response to 'Why p-values are not measures of evidence' by D. Lakens. *Trends Ecol. Evol.* 27, 291–292
6. Altman, D.G. and Bland, J.M. (1995) Absence of evidence is not evidence of absence. *Br. Med. J.* 311, 485
7. Kass, R.E. and Raftery, A.E. (1995) Bayes factors. *J. Am. Stat. Assoc.* 90, 773–795
8. Benjamin, D. *et al.* (2017) Redefine statistical significance. *Nat. Hum. Behav.* 2, 6–10
9. Held, L. and Ott, M. (2016) How the maximal evidence of p-values against point null hypotheses depends on sample size. *Am. Stat.* 70, 335–341
10. Murtaugh, P.A. (2014) In defense of P values. *Ecology* 95, 611–617
11. Greenland, S. *et al.* (2016) Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. *Eur. J. Epidemiol.* 31, 337–350
12. Gigerenzer, G. (2004) Mindless statistics. *J. Socio-Econom.* 33, 587–606