VIEWPOINT

Improving the transparency of statistical reporting in Conservation Letters

Fiona Fidler¹,² | Hannah Fraser¹ | Michael A McCarthy¹ | Edward T Game³,⁴

¹School of BioSciences, University of Melbourne, Australia
²School of Historical and Philosophical Studies, University of Melbourne, Australia
³The Nature Conservancy, South Brisbane, Australia
⁴University of Queensland, St. Lucia, Australia

Correspondence
Edward T Game, The Nature Conservancy, South Brisbane, Australia.
Email: egame@tnc.org

Conservation Letters’ new policy on reporting confidence intervals (CIs) with p values is one among many recent calls for change in statistical reporting practices. It sits in line with the recently developed Tools for Transparency in Ecology and Evolution (TTEE; Parker et al., 2016; TTEE_Working_Group, 2016), which are themselves based on the interdisciplinary Transparency and Openness Promotion Guidelines (Nosek et al., 2018). Complete and transparent statistical reporting is essential to building a reliable evidence base for practice, and for accumulating and synthesizing scientific knowledge. Conversely, undisclosed analysis practices such as cherry picking “significant” results and p-hacking (e.g., making decisions about sampling stopping rules, treatment of outliers, transformations, and/or analysis techniques based on whether results meet or fail to meet a statistical significance threshold) have been directly linked to the inability to replicate many important, published experimental effects (Fidler et al., 2017; Forstmeier, Wagenmakers, & Parker, 2017; Simmons, Nelson, & Simonsohn, 2011).

Given Conservation Letters’ focus on publishing science of direct relevance to policy and practice, it is particularly important that the interpretation of statistical analyses and the conclusions supported by this are transparent. From April 2018, Conservation Letters will be requiring: (1) that any article reporting p values must also report 95% CIs in the text and in figures and (2) that all figures that include data used in statistical analyses (whether in the main text or Supporting Information) must show error bars on the figure. Where possible, these error bars should show 95% CIs, but in all cases authors must be explicit about what the error bars show.

The Conservation Letters’ policy encapsulates seven important messages about CIs and p values, which we explicate below.

1 | MESSAGE NO. 1: REPORT AND INTERPRET CIs; INTERVAL LENGTH IS A GUIDE TO PRECISION

While based on the same basic information as p values, CIs make uncertainty in parameter values more explicit than do p values alone. For example, they have been shown experimentally to reduce the misinterpretation of statistical nonsignificance as “no effect” and otherwise improve interpretation (Fidler & Loftus, 2009).

A CI indicates a parameter’s precision, a concept akin to statistical power. A longer CI indicates less precision; a shorter interval indicates relatively high precision. CIs indicate a set of plausible values for the parameter, with longer intervals encompassing a wider range of plausible values. Figure 1 illustrates possible effect sizes with relevant 95% CIs, relative to levels that are considered important and not important for five hypothetical results. CI-A shows a highly imprecise result that while not statistically significant (interval it includes zero) is wide enough to also include values in the ecologically or theoretically important range.
CI-B shows a statistically significant result (interval excludes zero), but is still not precise enough to distinguish between ecologically or theoretically important and unimportant values (we discuss “importance” further in message No. 6). CI-C shows a more precise nonsignificant result; the interval includes zero and is sufficiently narrow to rule out other important values. CI-D is similarly precise, but at the other end of the spectrum; a result that is both ecologically and statistically significant. CI-E demonstrates how a statistically significant result can fail to be ecologically important.

2 | MESSAGE NO. 2: AVOID MERELY DICHOTOMOUS INTERPRETATIONS OF CIs

Simply noting whether zero (or some other null value) is inside or outside a CI ignores other important information CIs have to offer, most notably that the interval width is a guide to the precision of the result. It also fails to recognize that intervals rarely have a uniform distribution. Values closer to the middle of the interval are (usually) more likely to represent the parameter than those toward the edges.

3 | MESSAGE NO. 3: ALWAYS SPECIFY PRECISELY WHAT AN ERROR BAR REPRESENTS

In many instances, authors who report error bars fail to specify precisely what the bars represent. Error bars in figures should clearly identify whether the bars represent standard deviations, standard errors, or CIs, and the source of that variation (e.g., variation among vs. within sites). When reporting CIs, always ensure that the level of the confidence (e.g., 95%) is noted.

If you are also reporting the outcomes of null hypothesis significance tests (i.e., \( p \) values), below are some further important messages.

4 | MESSAGE NO. 4: STATE THE SAMPLING STOPPING RULE ASSOCIATED WITH YOUR HYPOTHESIS TEST

Often, there are practical constraints on sample size, and therefore statistical power. It will not always be possible to increase sample size or to otherwise achieve a statistical power level of 80% or more. This does not necessarily mean the research is not worthwhile. It is essential, however, that even when power is low, it is calculated (through an a priori power analysis) and reported for any hypothesis testing result.

The vast majority of papers in conservation science fail to acknowledge the prospect of type II (false negative) errors. Independent calculations suggest that the average statistical power in ecology and related research is low. For example, Jennions and Møller (2003) estimated the average power of behavioral ecology to be approximately 40% to 47% for medium (typical) effect sizes. Smith, Gammel, and Hardy's (2011) estimate was even lower as 23% to 26% (Smith et al., 2011). This means that the chance of detecting a real effect of medium size in this field is considerably worse than flipping a coin. Parris and McCarthy's (2001) study of the effects of toe-clipping of frogs revealed similarly low power; at best 60% for a large effect of 40% population decline. If power is not reported, it is safest for editors, reviewers, and policy makers to assume it is low.

5 | MESSAGE NO. 5: ENSURE THAT FAILURE TO REJECT A NULL HYPOTHESIS IS NOT INTERPRETED AS EVIDENCE THAT THE NULL HYPOTHESIS IS TRUE; ABSENCE OF EVIDENCE IS NOT EVIDENCE OF ABSENCE

Failure to reject a null hypothesis does not provide evidence that the null hypothesis is true. When power is unknown, statistical nonsignificance is uninterpretable. Although this advice may seem obvious, it is unfortunately common for authors in conservation science to present statistical nonsignificance as evidence that the null is true. For example, in a study by Pavone and Boonstra (1985), the average
lifespan of toe-clipped voles was not significantly different from that of control animals; toe-clipping was interpreted as having no effect on survival. However, the authors should also have noted that the size of the effect was also not significantly different from a 40% reduction in lifespan due to toe-clipping, a potentially large impact. So while they were unable to rule out no effect of toe-clipping, the data were also insufficient to rule out a large effect. Misinterpreting statistical nonsignificance as “no effect” can lead to failures to act to protect biodiversity.

6 | MESSAGE NO. 6: DO NOT EQUATE STATISTICAL SIGNIFICANCE WITH ECOLOGICAL IMPORTANCE

Many studies in conservation science and ecology equate statistical significance with ecological or theoretical importance, as the example above illustrates. Unfortunately, statistical and ecological significance have little to do with one another. Broadly speaking, the effect size measures the magnitude of the change in a parameter that one observes, or expects to observe, from a treatment or exposure to a causal variable. A study result is compelling evidence of an effect only if the effect is large enough to be ecologically or theoretically interesting and unusual enough not to have arisen by chance.

7 | MESSAGE NO. 7: LOOK OUT FOR LESS OBVIOUS INSTANCES OF NULL HYPOTHESIS TESTING; MESSAGE NOs. 4 TO 6 APPLY TO THEM TOO

The messages above are not only relevant to researchers conducting t-tests and ANOVAs as their core analyses. We have on occasion, anecdotally, heard colleagues and peers claim that are involved in modeling, not null hypothesis testing, and as such do not need to consider statistical power or effect size. On closer inspection, many such cases do involve null hypothesis testing as part of a larger procedure, for example, parameters selected for inclusion in models on the grounds that they reached $p < .05$, or goodness-of-fit statistics later subjected to statistical significance analysis. Another often overlooked instance of null hypothesis testing occurs in tests of statistical assumptions (e.g., homogenous variance). Such tests may return nonsignificant results which form the basis of decisions about further analysis, for example, decisions to combine groups of data that show “no difference.” It is important to recognize these instances are null hypothesis testing, and as such, require power calculations and all the same considerations as tests of primary hypotheses.

A number of resources are available to support implementation of Conservation Letters’ policy. Cumming (2012) and Cumming and Calin-Jageman (2014) provide useful information on reporting and interpreting CIs. Cumming also provides explanatory YouTube videos that include visual aids and simulations to improve statistical inference: https://www.youtube.com/user/geoffdcumming. Nakagawa and Cuthill (2007) also provide excellent practical advice for calculating and interpreting effect sizes and CIs for biologists. We thank Conservation Letters’ authors for working to improve the robustness and transparency of statistical reporting, ultimately increasing confidence in the policy importance of the work published here.

ORCID
Edward T Game http://orcid.org/0000-0003-4707-9281

REFERENCES
Cumming, G. (2012). Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis. New York: Routledge.
Cumming, G., & Calin-Jageman, R. (2014). Introduction to the new statistics: Estimation, open science and beyond. New York: Routledge.
Fidler, F., & Loftus, G. R. (2009). Why figures with error bars should replace p values: Some conceptual arguments and empirical demonstrations. Zeitschrift für Psychologie/Journal of Psychology, 217, 27–37.
Fidler, F., Chee, Y., Wintle, B., Burgman, M., McCarthy, M., & Gordon, A. (2017). Meta-research for evaluating reproducibility in ecology and evolution. BioScience. Retrieved from https://doi.org/10.1093/biosci/biw159
Forstmeier, W., Wagenmakers, E. J., & Parker, T. H. (2017). Detecting and avoiding likely false-positive findings—a practical guide. Biological Reviews, 92, 1941–1968. Retrieved from https://doi.org/10.1111/brv.12315
Jennions, M. D., & Möller, A. P. (2003). A survey of the statistical power of research in behavioral ecology and animal behavior. Behavioral Ecology, 14, 438–445.
Nakagawa, S., & Cuthill, I. C. (2007). Effect size, confidence interval and statistical significance: A practical guide for biologists. Biological Reviews, 82, 591–605.
Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., … DeHaven, A. C. (2018). Transparency and openness promotion (TOP) guidelines. Retrieved from osf.io/9f6gx
Parker, T. H., Forstmeier, W., Koricheva, J., Fidler, F., Hadfield, J. D., Chee, Y. E., … Nakagawa, S. (2016). Transparency in ecology and evolution: Real problems, real solutions. Trends in Ecology & Evolution, 31, 711–719. Retrieved from https://doi.org/10.1016/j.tree.2016.07.002
Parris, K. M., & McCarthy, M. A. (2001). Identifying effects of toe-clipping on anuran return rates: The importance of statistical power. Amphibia Reptilia, 22, 275–289.

FIDLER ET AL.
Pavone, L. V., & Boonstra, R. (1985). The effects of toe clipping on the survival of the meadow vole (Microtus pennsylvanicus). *Canadian Journal of Zoology*, 63, 499–501.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, 22, 1359–1366.

Smith, D. R., Gammel, M., & Hardy, I. C. W. (2011). Power rangers: No improvement in the statistical power of analyses published in *Animal Behaviour*. *Animal Behavior*, 81, 347–352.

TTEE Working Group. (2016). *Tools for transparency in ecology and evolution (TTEE)*. Open science framework. Retrieved from https://doi.org/10.17605/OSF.IO/G65CB

**How to cite this article:** Fidler F, Fraser H, McCarthy MA, Game ET. Improving the transparency of statistical reporting in *Conservation Letters*. *Conservation Letters*. 2018;11:e12453. [https://doi.org/10.1111/conl.12453](https://doi.org/10.1111/conl.12453)