Confidence intervals are not a way of moving beyond p-values

Authors: Aaron R. Caldwell (1), Matthew S. Tenan (2)

1 - Natick, MA, USA
2 - Rockefeller Neuroscience Institute, West Virginia University, Morgantown, WV, USA

Correspondence:

Aaron R. Caldwell, arcaldwell49@gmail.com

Word Count: 1000 (of max of 1000)

We read with interest the manuscript by Williams et al (2023) published within the *The Journal of Physiology* on displaying confidence intervals and effect sizes when reporting statistical results. In isolation, this is reasonable advice, but we believe Williams et al (2023) recommendations are unclear and their discussion can benefit from more nuance. Here, we hope to add some of that nuance.

Confidence intervals and p-values are both frequentist statistics and are often derived from the same quantities (i.e., estimate and standard error). However, they address different aspects of uncertainty. While confidence intervals provide valuable information about the range of parameter values compatible with your data given the model assumptions, they are not a direct solution to the problems associated with NHST. Briefly, let us compare and contrast the p-value and the confidence interval:

1. Information Conveyed:
    - P-values provide a probability that can be used to quantify the compatibility of the estimated model with a null model. This requires that a null model/value is of interest to be tested against (i.e., superiority, non-inferiority, or equivalence hypotheses).
    - Confidence intervals provide a range of values for a parameter that are compatible with the data *at a specific threshold* (e.g., 95% frequentist coverage). They can give a sense of the precision and uncertainty of the estimate, much like a standard error.
2. Threshold-based Thinking:

    - P-values are often subject to dichotomous thinking when using NHST, where results are deemed "significant" if $p < 0.05$ and "non-significant" if $p > 0.05$.
    - While confidence intervals provide a range of values, they are still often dichotomized by examining whether the interval contains a specific value (e.g., zero). Dichotomous thinking and misinterpretations can persist even when using confidence intervals (Hoekstra et al, 2014; Fricker et al, 2019).

3. Misinterpretation:
    - P-values are frequently misunderstood and misinterpreted, leading to the misapplication of statistical tests (Gigerenzer, 2004).
    - Confidence intervals can be misunderstood. For example, assuming that the true parameter value lies within the confidence interval and defaulting to an alpha level of 0.05 would replicate many of the problems with NHST (Hoekstra et al, 2014; Fricker et al, 2019).
4. Lack of Effect Size Information:
    - P-values do not provide information about the size or magnitude of an effect; they only indicate divergence from the null model.
    - Confidence intervals do provide information about the effect size and the uncertainty of the effect size estimate. However, they do not convey the practical significance or the importance of an effect on their own. Additionally, the use of benchmark effect size scales (e.g., Cohen's recommendations) or field specific scales to interpret the effect size are likely to mislead rather than inform (Caldwell & Vigotsky, 2020; Panzarella et al, 2021)
5. Subjective Choice of Alpha/Confidence Level:
    - An alpha level (often 0.05) can be utilized to designate results as significant or non-significant. However, when an exact p-value is presented the reader can directly see the degree of incompatibility with the null model and interpret the results accordingly.
    - The width of a confidence interval depends on the chosen confidence level (e.g., 95%, 90%, 99%). Researchers may select a confidence level that suits their goals, which can be seen as analogous to "p-hacking" when testing a specific hypothesis. Unless multiple confidence intervals are presented (Rafi & Greenland, 2020), readers will only see the range of values at one level of confidence/compatibility which may encourage dichotomous interpretations.

In practice, confidence intervals and p-values can be used together to provide a more comprehensive understanding of the data when testing a hypothesis. Confidence intervals offer a range of values for a parameter estimate, while p-values can provide information about the incompatibility between the parameter estimate and the null model. Additionally, neither the p-value nor the confidence interval need to be interpreted in a dichotomous manner (Greenland, 2019; Rafi & Greenland, 2020). Statements from Williams et al (2023) such as, "the effect size and confidence interval tell us everything a P value does about a result and so much more", are not accurate. In most contexts, we believe it is important for researchers to cautiously interpret *both* the p-value and confidence intervals. The only situation where p-value does not make sense to report is one where a researcher has no hypothesis test and therefore no null model from which to compare the data (i.e., truly an estimation based approach). Both values can be evaluated in the context of the research question rather than relying solely on either one to draw conclusions.

Conclusions

We agree with the Williams et al (2023) that in most scenarios researchers should report more than just the significance of a p-value (Caldwell & Vigotsky, 2020; Caldwell & Cheuvront, 2019), and this advice has been echoed in the statistical guidelines released by the American Physiological Society (Curran-Everett & Benos, 2004; Curran-Everett & Benos, 2007). The reporting of an effect size and its confidence interval should be encouraged, but this does not need to happen at the expense of the p-value. Furthermore, more restraint needs to be shown when discussing the effectiveness of confidence intervals for interpreting results. We see no empirical evidence of the claim from Williams et al (2023) that presenting an effect size and confidence will “undoubtedly enhance the interpretation of research findings”. Furthermore, Williams et al (2023) mention that many journals are banning p-values and advocating an estimation approach. Current evidence would suggest that such editorial practices have not had a positive impact (Fricker et al, 2019). Confidence intervals are not a panacea for our statistical woes (Tenan & Caldwell, 2022; Lohse, 2022; Lakens, 2022), and they are subject to many of the same misinterpretations as p-values (Hoekstra et al, 2014). There are many valid ways to analyze, interpret, and present results within a scientific manuscript which may, or may not, include a p-value, confidence interval, or effect size estimate. Therefore, we assert that confidence intervals are not a way of “moving beyond” p-values, but rather another tool that physiologists should keep in their statistical toolbox.

References

1. Caldwell, A. R., & Cheuvront, S. N. (2019). Basic statistical considerations for physiology. *Temperature* (Vol. 6, Issue 3, pp. 181–210). Informa UK Limited. https://doi.org/10.1080/23328940.2019.1624131
2. Caldwell, A., & Vigotsky, A. D. (2020). A case against default effect sizes in sport and exercise science. *PeerJ* (Vol. 8, p. e10314). PeerJ. https://doi.org/10.7717/peerj.10314
3. Curran-Everett, D., & Benos, D. J. (2004). Guidelines for reporting statistics in journals published by the American Physiological Society. *American Journal of Physiology-Regulatory, Integrative and Comparative Physiology* (Vol. 287, Issue 2, pp. R247–R249). American Physiological Society. https://doi.org/10.1152/ajpregu.00346.2004
4. Curran-Everett, D., & Benos, D. J. (2007). Guidelines for reporting statistics in journals published by the American Physiological Society: the sequel. *Advances in Physiology Education* (Vol. 31, Issue 4, pp. 295–298). American Physiological Society. https://doi.org/10.1152/advan.00022.2007
5. Fricker, R. D., Jr., Burke, K., Han, X., & Woodall, W. H. (2019). Assessing the Statistical Analyses Used in Basic and Applied Social Psychology After Their p-Value Ban. *The*

American Statistician (Vol. 73, Issue sup1, pp. 374–384). Informa UK Limited. https://doi.org/10.1080/00031305.2018.1537892

6. Hoekstra, R., Morey, R. D., Rouder, J. N., & Wagenmakers, E.-J. (2014). Robust misinterpretation of confidence intervals. *Psychonomic Bulletin & Review* (Vol. 21, Issue 5, pp. 1157–1164). Springer Science and Business Media LLC. https://doi.org/10.3758/s13423-013-0572-3
7. Gigerenzer, G. (2004). Mindless statistics. *The Journal of Socio-Economics* (Vol. 33, Issue 5, pp. 587–606). Elsevier BV. https://doi.org/10.1016/j.socec.2004.09.033
8. Greenland, S. (2019). Valid P-Values Behave Exactly as They Should: Some Misleading Criticisms of P-Values and Their Resolution With S-Values. *The American Statistician* (Vol. 73, Issue sup1, pp. 106–114). Informa UK Limited. https://doi.org/10.1080/00031305.2018.1529625
9. Lakens, D. (2022). Correspondence: Reward, but do not yet require, interval hypothesis tests. *Journal of Physiotherapy* (Vol. 68, Issue 3, pp. 213–214). Elsevier BV. https://doi.org/10.1016/j.jphys.2022.06.004
10. Lohse, K. (2022). No Estimation without Inference. *Communications in Kinesiology* (Vol. 1, Issue 4). Society for Transparency, Openness, and Replication in Kinesiology. https://doi.org/10.51224/cik.2022.49
11. Panzarella, E., Beribisky, N., & Cribbie, R. A. (2021). Denouncing the use of field-specific effect size distributions to inform magnitude. *PeerJ* (Vol. 9, p. e11383). *PeerJ*. https://doi.org/10.7717/peerj.11383
12. Rafi, Z., & Greenland, S. (2020). Semantic and cognitive tools to aid statistical science: replace confidence and significance by compatibility and surprise. *BMC Medical Research Methodology* (Vol. 20, Issue 1). Springer Science and Business Media LLC. https://doi.org/10.1186/s12874-020-01105-9
13. Tenan, M., & Caldwell, A. (2022). Confidence Intervals and Smallest Worthwhile Change Are Not a Panacea. *Communications in Kinesiology* (Vol. 1, Issue 4). Society for Transparency, Openness, and Replication in Kinesiology. https://doi.org/10.51224/cik.2022.45
14. Williams, S., Carson, R., & Tóth, K. (2023). Moving beyond P values in The Journal of Physiology: A primer on the value of effect sizes and confidence intervals. *The Journal of Physiology*. Wiley. https://doi.org/10.1113/jp285575