References

1. Elkins MR, et al. *Journal of Physiotherapy*. 2021. https://doi.org/10.1016/j.jphys.2021.12.001
2. Boos DD, et al. *The American Statistician*. 2011;65:213–221. https://doi.org/10.1198/tas.2011.10129
3. Miller J, et al. *Psychological Methods*. 2011;16:337–360. https://doi.org/10.1037/a0023347
4. Lakens, D. 2022. Sample Size Justification. PsyArXiv. https://doi.org/10.31234/osf.io/9d3yf
5. Murphy KR, et al. *Journal of Applied Psychology*. 1999;84:234–248. https://doi.org/10.1037/0021-9010.84.2.234
6. Lakens D. *Social Psychological and Personality Science*. 2017;8:355–362. https://doi.org/10.1177/1948550617697177
7. Lakens D. *Perspectives on Psychological Science*. 2021;16:639–648. https://doi.org/10.1177/1745691620958012
8. Scheel AM, et al. *Perspectives on Psychological Science*. 2021;16:744–755. https://doi.org/10.1177/1745691620966795 32

Correspondence: Response to Lakens

We thank Associate Professor Lakens for his interest in our editorial. We will address each of his five comments.

'*… only God knows the probability that the null hypothesis is true given the data observed, and no statistical method can provide it. Estimation will not tell you anything about the probability of hypotheses.*' Bayesian analyses can estimate the probability of a hypothesis given the data. In any case, as the editorial explains, there is little point in knowing the probability that the null hypothesis is true.

'A p-value does not constitute evidence. Neither do estimates, so their proposed alternative suffers the same criticism.' It is true that estimates, like *p* values, are not evidence. However, proponents of null hypothesis testing imply that *p* values are useful or meaningful because they provide evidence that can be used to reject the null hypothesis (Fisherian significance testing) or that can be used to choose between the null and alternative hypotheses (Neyman-Pearson hypothesis testing). In contrast estimates, unlike *p* values, are intrinsically meaningful.

'*It is not possible to determine the probability a study will replicate based on a single value (Miller & Schwarz, 2011). Furthermore, well-designed replication studies do not use the same sample size as an earlier study, but are designed to have high power for an effect size of interest (Lakens, 2022).*' We don't disagree with either assertion. Neither changes the substantive point: experimenters who obtain a significant test finding cannot expect that, if an exact replication of their study were possible, it too would obtain a significant finding. As Amhrein and Greenland[1] state: 'random variation alone can easily lead to large disparities in *P* values, far beyond falling just to either side of the 0.05 threshold. For example, even if researchers could conduct two perfect replication studies of some genuine effect, each with 80% power (chance) of achieving $P < 0.05$, it would not be very surprising for one to obtain $P < 0.01$ and the other $P > 0.30$' (p306).

'*Fourth, the editors argue, without any empirical evidence, that in most clinical trials the null-hypothesis must be false.*' The assertion that the null hypothesis is false in most clinical trials does not require empirical evidence, because it is self-evidently true. The null hypothesis is that there is *exactly* no effect – it is not, as A/Prof Lakens implies, that the null is true within the bounds we can detect with available resources. While the latter may often be true, the former never is. The null hypothesis may often be approximately true, but it is rarely if ever exactly true. Moreover, empirical estimates of effects are always at least a little bit biased. So exactly null hypotheses must always be false. The only reason they are not always found to be false is that almost all studies lack the precision to detect tiny effects. For that reason, empirical evidence is unable to demonstrate that the null hypothesis is not always true. And that is why van der Laan and Rose[2] state that 'We know that for large enough sample sizes, every study, including one in which the null hypothesis of no effect is true, will declare a statistically significant effect' (p xvi).

'*Finally, the fifth point that "Researchers need to know more than just whether an effect does or does not exist" is correct, but the "more than" is crucial. It remains important to prevent authors from claiming there is an effect, when they are actually looking at random noise, and therefore, effect sizes complement, but do not replace, hypothesis tests.*' We disagree for reasons explained in the previous paragraph. There is no reason to worry about authors claiming there is an effect when there truly is exactly no effect, because there truly always is at least some effect (although the effect may be microscopically small). Instead of being concerned with whether there is or is not an effect we need to know if the effect is big enough to be of any substantive interest. *p* values convey no useful information on this issue, and they convey no information that cannot be gleaned from a confidence interval. In contrast, confidence intervals contain much useful information that cannot be gleaned from a *p* value. Confidence intervals can replace *p* values without any loss of useful information.

A/Prof Lakens argues that the suggestion of how to interpret a confidence interval is not estimation but is 'minimum effect testing'. In our opinion, the key feature of estimation is that it seeks to estimate the value of a population parameter. That should be the key objective of most inferential statistical analyses, and is the approach advocated in the editorial. Interpretation of the data from clinical trials inevitably involves consideration of the importance or clinical significance of the estimated average effect of the intervention. A/Prof Lakens points out that, if that is done formally using the tools of significance or hypothesis testing then it becomes minimum effect testing. And, as he points out, that requires formal enumeration of the smallest important effect. However, like Amrhein and Greenland,[1] we do not see the need to use the machinery of significance testing or hypothesis testing to rationally interpret estimates of effect. In the absence of a well-established threshold for interpretation, authors can still interpret a confidence interval by describing the practical implications of all values inside the confidence interval.[1] And there is another reason not to conduct minimum effect tests: researchers who supply confidence intervals, rather than conducting minimum effect tests, devolve the responsibility of distinguishing between important and unimportant effects to their readers. Arguably that is where that responsibility should lie.

Mark R Elkins, Rafael Zambelli Pinto, Arianne Verhagen, Monika Grygorowicz, Anne Söderlund, Matthieu Guemann, Antonia Gómez-Conesa, Sarah Blanton, Jean-Michel Brismée, Shabnam Agarwal, Alan Jette, Michele Harms, Geert Verheyden and Umer Sheikh

https://doi.org/10.1016/j.jphys.2022.06.003

References

1. Amrhein V, et al. *Nature*. 2019;567:305–307.
2. van der Laan MJ, et al. *Targeted Learning. Causal Inference for Observational and Experimental Data*. Springer; 2011.