No Estimation without Inference:

A Response to the International Society of Physiotherapy Journal Editors

Keith Lohse, PhD[1]

[1] Physical Therapy and Neurology, Washington University School of Medicine, Saint Louis, MO

NOTE THIS IS AN AUTHOR'S PRE-PRINT AND HAS NOT BEEN PEER-REVIEWED.
Please cite this pre-print as: Lohse, K.R. (2022). No Estimation without Inference: A Response to the International Society of Physiotherapy Journal Editors. *SportRxiv*.

Date Submitted
2022-07-17

Keywords:
"physical therapy"; "statistical significance"; "inference"; "estimation"

Corresponding Author:
Keith Lohse, PhD, PStat; lohse@wustl.edu

Recently, Elkins et al.[1] (hereafter referred to as "the Editorial") published an editorial on behalf of the International Society of Physiotherapy Journal Editors (ISPJE), recommending that researchers stop using null-hypothesis significance tests and adopt "estimation methods". Further, the editorial warns that this is not merely an idea to consider, but a coming policy of journals: "the [ISPJE] will be expecting manuscripts to use estimation methods instead of null hypothesis statistical tests".

I commend the Editorial for encouraging researchers to think deeply about the statistical tools available to them, to consider "practical significance" as well as "statistical significance", and for bringing important methodological discussions to the forefront of physical therapy research. However, the Editorial is also deeply flawed in its statistical reasoning. If these practices were adopted, they could damage the statistical literacy and scientific integrity of the field. We detail each of these critiques below, but in short the Editorial: (1) fails to adequately grapple with the inherent connection between "statistical inference" and "estimation" methods, (2) presents several misleading arguments about the flaws of significance tests, and (3) presents an alternative that is, in itself, a form of significance test – the minimal effects test[2] (the proposed alternative also muddles two-sided and one-sided hypothesis testing). Finally, I end with a short list of more urgent problems that the ISPJE could work to address.

Inference and Estimation are Inescapably Intertwined

The editorial presents statistical inference and estimation as two distinct methodological approaches. However, a very rudimentary example shows that these ideas are two sides of the same coin. Consider the mean differences and 95% confidence intervals shown in Figure 1. The 95% confidence interval shows values in the population that would be *compatible* with what we observed in the sample.[3] That is, if you move outside of the confidence interval, any of those parameter values ($\mu's$) would be statistically different from the mean observed in the sample ($\bar{x}$'s) at the p<0.05 level. Inside of the confidence interval, any of those parameter values would not be sufficiently different (p>0.05) from the observed mean difference. Similarly, the null-hypothesis significance test (NHST) assumes that the true value in the population is 0 (i.e., $H_0: \mu_1 - \mu_2 = 0$). The further the sample mean difference is away from

0, the lower the probability of observing that sample mean [if the null-hypothesis were true; i.e., $p(\geq \bar{x}|H_0)$]. The exact value of *p* will depend on the degrees of freedom and the standard error (under a *t*-distribution), but in general if a sample is ≥2 standard errors away from the hypothesized null value, then *p* will be less than 0.05. Thus, inference and estimation cannot be fully disentangled: estimation (frequentist or Bayesian) asks about plausible *values of the parameter* in the population, inference asks about the plausibility of *a specific parameter value* (and in the frequentist paradigm, uncertainty is accounted for with long-run error control, e.g., setting the Type 1 Error rate, $\alpha = 0.05$). This can be seen in the behavior of the confidence intervals relative to the p-values in Figure 1A-E: any confidence interval that does not contain zero also has p<0.05, for the null hypotheses significance test (NHST).

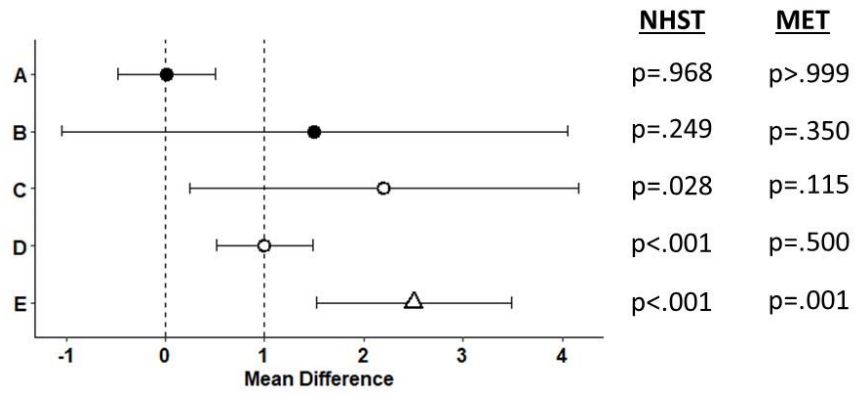

Figure 1. Five hypothetical scenarios (A-E) in which the means of two independent groups are compared using a t-test. The magnitude of the mean difference is shown on the x-axis. Solid points are not statistically significant (p>0.05) for a two-sided null hypothesis significance test assuming H_0: $\mu_1 - \mu_2 = 0$. Circles are not statistically significant (p>0.05) for a one-sided minimal-effects test assuming H_0: $\mu_1 - \mu_2 \leq 1$. Only case E is statistically significant in both cases, and note the 95% confidence interval does not contain either of the tested values (0 or 1).

Importantly, the Editorial does not address the fact that we can set H_0 to be any value. For instance, rather than setting $H_0 = 0$ (sometimes referred to as the "nil-hypothesis")[4], we can set H_0 equal to some clinically meaningful value of interest. This is referred to as a minimal effects test (or minimum effect test, MET[2,5]). For the sake of argument, let's say this value is 1. Comparing the

confidence intervals to our new null value, you can see that any confidence intervals that only contain values larger than 1 also have a $p<0.05$ for the minimal effects test (i.e., Figure 1E).[i] Thus, we have both an inference (i.e., do we reject the specific value of the population parameter in our null hypothesis?) and an estimate (i.e., the lower and upper limits of our 95% confidence intervals) in both the NHST and the MET, but the inference and the estimate are complementary and connected.

Misleading Arguments about flaws with Significance Tests

The Editorial bases many arguments on a previous list of perceived problems from Herbert (2019).[6] First, the Herbert paper is in itself an editorial that presents informed arguments, but is not an objective demonstration of any mathematical facts. So, reinforcing the Editorial's list through a citation does not provide an evidentiary foundation: it is layering opinion on top of opinion. Second, each of the five "problems" outlined by the Editorial is either not really a problem inherent to p-values or is a true but misleading statement. I address each stated problem from the Editorial (in quotes) below:

1. ***"A p-value is not the probability that a hypothesis is (or is not) true."*** – This is correct, but it does not follow that this makes p-values (or even statistical significance tests) unhelpful or uninformative. Knowing that the observed data are incompatible with the null hypothesis is a crucial step for many research questions.

2. ***"A p-value does not constitute evidence"*** – This is an oversimplification and misleading. The Editorial is correct that a single p-value cannot tell us about the probability of the null hypothesis being true, but p-values can be used as evidence against the null-hypothesis. First, as the p-value is calculated assuming that the null is true, so the Editorial is correct that we cannot simply flip the question around, assume the

[i] METs are typically directional, using one-sided hypothesis tests (e.g., $H_0 \leq 1$) whereas NSHTs are often non-directional, using two-sided hypothesis tests (e.g., $H_0 = 0$). Thus, although the confidence interval for Figure 1A does not contain the null value of 1, the whole of the confidence interval is below 1, thus yielding a non-significant minimal effects test.

data, and get the probability of the null, $p(\bar{x}|H_0) \neq p(H_0|\bar{x})$. To truly determine the probability of the null hypothesis, we would need Bayesian statistics in which we formalize some *prior* probability about the null hypothesis.[7] If we have a strong enough prior probability that the null is true, then the current data in the sample may not lead us to change our beliefs in the *posterior* distribution (i.e., 0 may still have a high likelihood of being true even though we observed extreme data). However, for any given prior distribution, observing larger effects (and hence smaller p-values from an NHST) will also lead to lower likelihoods in the posterior distribution.[ii] Second, as shown in Figure 2A, p-values have a uniform distribution under the null hypothesis, so if the null is true then any p-value is equally likely.[8] However, if the null is not true, then we will see a shift in the distribution of p-values, with small p-values becoming more common, Figure 2B. Thus, p-values are *relatively* less likely to be observed when the null hypothesis is true compared to when an alternative hypothesis is true, but to make quantifiable judgments about evidence for/against a specific hypothesis, we need more than a single p-value.[7–9]

3. ***"Statistically significant findings are not very replicable."*** – This is not true or at least very misleading. First, it is difficult to precisely define replication,[10,11] but across many different definitions we would expect more statistically significant findings to "replicate" provided that hypothesis tests have adequate statistical power, researchers have not engaged in p-hacking, there is not select reporting of results, etc. (more on this below in our section on threats to statistical integrity). Thus, not all statistically significant findings will replicate,[12] but replicable findings should mostly be statistically significant in well-designed studies.[13–15] Second, any threats to replicability are also going to affect confidence intervals (the Editorial's proposed solution) as much they affect p-values, because again, the p-value is inextricably linked to the upper and lower limits of the confidence interval.

4. ***"In most clinical trials, the null hypothesis must be false."*** – This is true but very misleading. It is true that "real" treatment effects are unlikely to be precisely 0 (e.g., they might be +0.001), but it begs the

[ii] For a humorous demonstration see: https://xkcd.com/1132/ ; for a more quantitative visualization of the relationship between priors, p-values, and posteriors see: https://rpsychologist.com/d3/bayes/.

question of do we really care if it is 0 or 0.001? And we will ever have the statistical precision to discern that difference? Scientists are often working on the frontiers of human knowledge; this is costly work where we need to explore a lot of different ideas and many them do not pan out (perhaps most[13]). However, the Editorial is specifically critiquing the "nil" hypothesis (i.e., $H_0 = 0$), when we could hypothesize any value.[2,5] First, simply because a point estimate of $H_0 = 0$ is unlikely to be true does not mean that it is unhelpful to ask (i.e., it should be a very low bar to show that your clinical treatment has a non-zero effect!). And second, we can set that null value to be anything we want (i.e., $H_0 \leq 0.4$ m/s for an improvement in gait speed or $H_0 \leq 30\%$ change on a standardized pain scale).

5. **"*Researchers need information about the size of effects*."** – This is a true statement, but it is not a problem with p-values nor null hypothesis significance tests. To my knowledge, no statistician has ever recommended that applied researchers ignore measures of effect size (either raw or standardized). Measures of effect size are a great addition to any results section. I would even take this one step further and encourage authors to share their data whenever possible[16], enabling other researchers to calculate their own effect sizes (as there can be limitations with and confusion about standardized effects sizes, and there is no one-size-fits-all solution to effect sizes[17–19]).

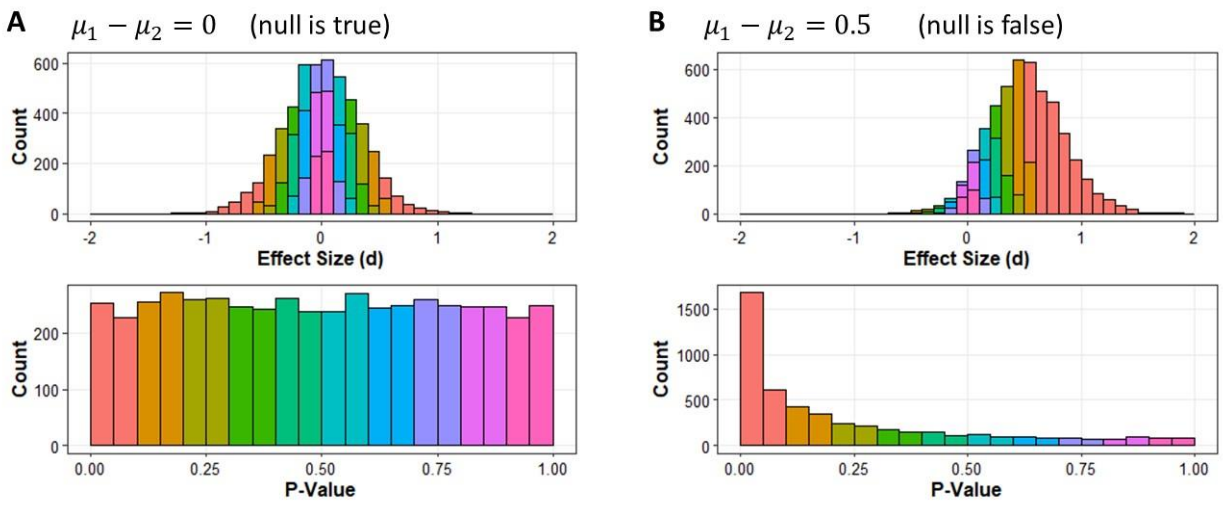

Figure 2. Simulated experiments (k=5,000), in which the means of two-independent groups are compared using a t-test assuming equal variances (n/group=20 with both samples from normal populations with $\sigma = 1$). Bars are color coded with respect to their p-values (10 levels from 0 to 1 by 0.1). Note that when the null is true in Panel A, the effect sizes (calculated as Cohen's d) cluster around 0 and have a normal distribution. However, the p-values have a uniform distribution that spans from 0 to 1 (all p-values are equally likely under the null). In Panel B, when the null is false, the effect sizes cluster around the new population parameter, 0.5, retaining their normal distribution. The p-values, however, now have a heavily right-skewed distribution with most values falling below p<0.05.

The Editorial's "Alternative" is essentially an NHST – The Minimal Effects Test

After detailing the potential problems with the NHST, the Editorial proposes an alternative solution in which they encourage authors to compare their 95% confidence to some minimum clinically meaningful value (which we will write as δ).[iii] This is absolutely a good practice and we would encourage researchers to report 95% confidence intervals and interpret their upper and lower limits in context, when appropriate. However, what the Editorial is suggesting is an MET where $H_0: \mu \leq \delta$. That is, if the test is to see if the 95% confidence interval does not contain δ, then that is mathematically equivalent to an MET assuming $H_0: \mu \leq \delta$ and finding $p < 0.025$. (Note p<0.025, not p<0.05, because most METs are one-sided hypothesis tests whereas confidence intervals are two sided; see Figure 1 and Footnote 1.) After

[iii] We caution that it is difficult to find a single measure of δ; it changes as a function of the study population, the study context, and has its own uncertainty due to sampling error.[20,21]

heavily critiquing hypothesis testing as a method of inference, the Editorial ends up proposing a hypothesis test. This is clearly an illogical proposition.

I want to emphasis that it is absolutely valid for the Editorial to recommend that authors consider their 95% confidence interval relative to some clinically meaningful value. However, this is not an "alternative" to conducting a null hypothesis significance test, it is in fact mathematically identical to conducting a null hypothesis test with a carefully chosen null hypothesis. Both are valid.

Finally, it is important to stress that history provides us with several examples of how authors will view their data through rose-tinted glasses when quantitative statistical safeguards are removed. For instance, when *Basic and Applied Social Psychology* banned p-values, authors were found to overstate their conclusions well beyond what would have been considered if "statistical significance" had been a benchmark.[22] In sport and exercise science, "magnitude-based inference" was leveraged as a niche method that allowed authors to interpret differences as meaningful when they had very little statistical support (e.g., p's >0.25).[23–25] Statistical significance in an NHST does not necessarily need to be the benchmark nor 0.05 the default value[26–29], but it is always important to have a statistically sound framework for dealing with uncertainty.

Virtues of Hypothesis Testing and Greater Threats to Statistical Integrity

One of the great virtues of null-hypothesis significance testing is Type I error control while making minimal assumptions about the nature of the data or the world at large; if we set $\alpha = 0.05$, then we can be confident we will only get data ≥ to what we observed 5% of the time when the null is true. Importantly, this works not only for *t*-, and *z*-statistics but also *F*- and χ^2-statistics that have multiple degrees of freedom, asking questions about multiple effects simultaneously (a situation not covered by the Editorial and for which the Editorial's confidence interval alternative cannot apply). However, the benefits of error control are extinguished by questionable research practices such as p-hacking, sub-group analyses, flexible stopping rules, selective exclusion of outliers, selective reporting, or hypothesizing after

results are known[30–34]. Given the mathematical connection between p-values and confidence intervals, these questionable research practices pose an equal threat to confidence intervals. Switching to a fully Bayesian method of analysis is similarly not an antidote for poor study design, small samples, and questionable research practices. As others have argued[35,36], p-values get a disproportionate amount of attention in popular discussions of research methodology. I encourage the ISPJE to instead focus their attention on methods for improving data/code sharing, transparency, and replicability through tools like preregistration, results blind review and registered reports, or even "data papers" whose primary function is to report a design and archive data, without trying to draw inferences from limited sample sizes.

It is entirely valid to say that p-values are often mis-used and mis-interpreted, and "statistical significance" may not ultimately be the best term for applied researchers to use.[37] However, it is incorrect to present these human errors as inherent flaws in hypothesis testing. For instance, if someone mis-interprets $p>0.05$ as evidence of "no difference", the correct action is to teach them about equivalence tests and non-inferiority designs, not the banning of p-values. Similarly, there are times when Bayesian inference is what authors are really interested in and in those cases Bayesian inference can and should be used (e.g., what is the probability that the null is true, given the evidence?). However, Bayesian analysis is not a panacea (it comes with its own limitations and assumptions) and needs to be used thoughtfully like any statistical tool. So, although a simple heuristic of $p<0.05$ may well be overused as "the" test in physical therapy research, frequentist hypothesis tests are still valid and useful tools for physical therapy researchers. Moreover, the scientific integrity of the field has much larger concerns, and both p-values and confidence intervals will be corrupted by p-hacking, under-powered subgroup analyses, surrogate outcomes, and other questionable research practices.

In conclusion, I agree with the Editorial on the importance of reporting effect sizes and interpreting them in context. However, the Editorial makes numerous statistical faux pas that could harm the statistical literacy in our field (if readers take them at face value) and harm the scientific integrity of our field (if put into editorial practice).

References

1. Elkins, M. R. *et al.* Statistical inference through estimation: recommendations from the International Society of Physiotherapy Journal Editors. *Phys. Ther.* **102**, pzac066 (2022).
2. Murphy, K. R. & Myors, B. Testing the hypothesis that treatments have negligible effects: Minimum-effect tests in the general linear model. *J. Appl. Psychol.* **84**, 234–248 (1999).
3. Rafi, Z. & Greenland, S. Semantic and cognitive tools to aid statistical science: replace confidence and significance by compatibility and surprise. *BMC Med. Res. Methodol.* **20**, 244 (2020).
4. Cohen, J. The earth is round ($p < .05$). *Am. Psychol.* **49**, 997–1003 (1994).
5. Lakens, D. The Practical Alternative to the p Value Is the Correctly Used p Value. *Perspect. Psychol. Sci.* **16**, 639–648 (2021).
6. Herbert, R. Research Note: Significance testing and hypothesis testing: meaningless, misleading and mostly unnecessary. *J. Physiother.* **65**, 178–181 (2019).
7. Goodman, S. N. & Royall, R. Evidence and scientific research. *Am. J. Public Health* **78**, 1568–1574 (1988).
8. Lakens, D. Why p-values are not measures of evidence. (2021). PsyRxiv.
9. Goodman, S. N. Toward Evidence-Based Medical Statistics. 1: The P Value Fallacy. *Ann. Intern. Med.* **130**, 995–1004 (1999).
10. Collaboration, O. S. Estimating the reproducibility of psychological science. *Science* **349**, aac4716 (2015).
11. Patil, P., Peng, R. D. & Leek, J. T. What Should Researchers Expect When They Replicate Studies? A Statistical View of Replicability in Psychological Science. *Perspect. Psychol. Sci.* **11**, 539–544 (2016).
12. Scheel, A. M., Schijen, M. R. M. J. & Lakens, D. An Excess of Positive Results: Comparing the Standard Psychology Literature With Registered Reports. *Adv. Methods Pract. Psychol. Sci.* **4**, 25152459211007468 (2021).

13. Ioannidis, J. P. Why most published research findings are false. *PLoS Med.* **2**, e124 (2005).
14. Anderson, S. F. & Maxwell, S. E. Addressing the "Replication Crisis": Using Original Studies to Design Replication Studies with Appropriate Statistical Power. *Multivar. Behav. Res.* **52**, 305–324 (2017).
15. Nosek, B. A. *et al.* Replicability, robustness, and reproducibility in psychological science. *Annu. Rev. Psychol.* **73**, 719–748 (2022).
16. Borg, D. N. *et al.* Sharing data and code: a comment on the call for the adoption of more transparent research practices in sport and exercise science. (2020).
17. Caldwell, A. & Vigotsky, A. D. A case against default effect sizes in sport and exercise science. *PeerJ* **8**, e10314 (2020).
18. McGrath, R. E. & Meyer, G. J. When effect sizes disagree: the case of r and d. *Psychol. Methods* **11**, 386 (2006).
19. Levine, T. R. & Hullett, C. R. Eta Squared, Partial Eta Squared, and Misreporting of Effect Size in Communication Research. *Hum. Commun. Res.* **28**, 612–625 (2002).
20. Tenan, M. & Caldwell, A. A Critical Review of Phyiotherapy Editor's Comments on Statistical Practice. SportRxiv.
21. Dabija, D. I. & Jain, N. B. Minimal Clinically Important Difference of Shoulder Outcome Measures and Diagnoses: A Systematic Review. *Am. J. Phys. Med. Rehabil.* **98**, 671–676 (2019).
22. Fricker Jr, R. D., Burke, K., Han, X. & Woodall, W. H. Assessing the statistical analyses used in basic and applied social psychology after their p-value ban. *Am. Stat.* **73**, 374–384 (2019).
23. Sainani, K. L. The Problem with" Magnitude-based Inference". *Med. Sci. Sports Exerc.* **50**, 2166–2176 (2018).
24. Sainani, K. L., Lohse, K. R., Jones, P. R. & Vickers, A. Magnitude-based inference is not Bayesian and is not a valid method of inference. *Scand. J. Med. Sci. Sports* **29**, 1428 (2019).
25. Lohse, K. R. *et al.* Systematic review of the use of "magnitude-based inference" in sports science and medicine. *PloS One* **15**, e0235318 (2020).

26. Benjamin, D. J. *et al.* Redefine statistical significance. *Nat. Hum. Behav.* **2**, 6–10 (2018).
27. Lakens, D. *et al.* Justify your alpha. *Nat. Hum. Behav.* **2**, 168–171 (2018).
28. Amrhein, V. & Greenland, S. Remove, rather than redefine, statistical significance. *Nat. Hum. Behav.* **2**, 4–4 (2018).
29. McShane, B. B., Gal, D., Gelman, A., Robert, C. & Tackett, J. L. Abandon statistical significance. *Am. Stat.* **73**, 235–245 (2019).
30. Simmons, J. P., Nelson, L. D. & Simonsohn, U. Life after p-hacking. in *Meeting of the society for personality and social psychology, New Orleans, LA* 17–19 (2013).
31. Simmons, J. P., Nelson, L. D. & Simonsohn, U. False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. (2016).
32. Sun, X. *et al.* Credibility of claims of subgroup effects in randomised controlled trials: systematic review. *Bmj* **344**, (2012).
33. Kerr, N. L. HARKing: Hypothesizing after the results are known. *Personal. Soc. Psychol. Rev.* **2**, 196–217 (1998).
34. Rosenthal, R. The file drawer problem and tolerance for null results. *Psychol. Bull.* **86**, (1979).
35. Borg, D. N., Lohse, K. R. & Sainani, K. L. Ten common statistical errors from all phases of research, and their fixes. *PM&R* **12**, 610–614 (2020).
36. Leek, J. T. & Peng, R. D. Statistics: P values are just the tip of the iceberg. *Nature* **520**, 612–612 (2015).
37. Wasserstein, R. L., Schirm, A. L. & Lazar, N. A. Moving to a world beyond "p< 0.05". *The American Statistician* vol. 73 1–19 (2019).