A Critical Review of Phyiotherapy Editor’s Comments on Statistical Practice

Matthew Tenan1 & Aaron R. Caldwell2

1 A University, A City, A State, A Country

2 Society for Transparency, Openness, and Replication in Kinesiology

PREPRINT: NOT PEER REVIEWED

The authors made the following contributions. Matthew Tenan: Conceptualization, Writing - Original Draft Preparation, Writing - Review & Editing; Aaron R. Caldwell: Writing - Original Draft Preparat, Writing - Review & Editing.

**Corresponding Author:** Matthew Tenan, Postal address. E-mail: [my@email.com](mailto:my@email.com)

**Keywords:** Statistics, Physiotherapy, Estimation, Significance, Confidence Interval

Abstract

Something

We read with interest the recent Editorial written by [Elkins et al.](#ref-elkins2022) ([2022](#ref-elkins2022)), who are the Editor-in-Chief members of the International Society of Physiotherapy Journal Editors (heretofore referred to as “The Editorial”). We applaud the author group for encouraging clinical researchers to look beyond null-hypothesis testing (NHST) and into the realm of effect estimation. In the Frequentist framework, upon which The Editorial solely describes, NHST and effect estimation are two sides of the same coin with fundamental mathematical relationships, so it makes perfect sense to fully describe the analyses and results of clinical research to the fullest extent possible. As methodological tutorials have described previously ([Rafi & Greenland, 2020](#ref-rafi2020)), using estimation, or “unconditional,” approach to reporting statistics is entirely valid. However, The Editorial also contains a multitude of incorrect or misleading statements and the central thesis that Frequentist Confidence Intervals (CIs) should be contrasted against a point estimate of Smallest Worthwhile Effect (SWE) is fundamentally flawed. In this short response, we will briefly detail a non-exhaustive list of misleading statements in The Editorial and expand on the statistical and practical issues with suggesting that CI overlap with SWE metrics be used instead of NHST.

# 1 Misleading Statements in The Editorial

At a foundational level, the goal of NHST is to make inferences with an eye towards error control and the goal of Estimation, whether Frequentist or Bayesian, is to quantify the magnitude of an effect and the certainty of that effect (which is also directly related to error control in a frequentist paradigm). As The Editorial seems to understand (page 2, paragraph 6), there is a mathematical relationship between the p-values calculated through NHST and confidence intervals around model estimates ([Altman & Bland, 2011](#ref-altman2011)). For this reason, it is surprising the number of misstatements made within The Editorial regarding NHST and CIs. For example, Table 1 in The Editorial states “Statistically significant findings are not very replicable”; however, if exactly reproducing a study repeatedly in the same population with different samples, one would have the exact same replication characteristics for both p-values and CIs . This would seem to be a misunderstanding of the replication crisis which, while tangentially related to p-values, is largely due to systematic publication practices and the behavior of researchers. The authors also seem to forget that a move to CIs would suffer from these exact same issues and would not magically solve the problem of replicability ([Hoekstra, Morey, Rouder, & Wagenmakers, 2014](#ref-hoekstra2014); [Morey, Hoekstra, Rouder, Lee, & Wagenmakers, 2015](#ref-morey2015)).

Table 1 also makes some very peculiar assumptions about interventions in clinical trials by stating without evidence that “Almost all interventions would be expected to have some effect, even if that effect was trivially small.” It is possible this is inelegant wording, and the intention was to state that, within a given trial, it is highly unlikely for a measured construct to be exactly nil. This appears to be a vague allusion to Lindley’s paradox ([Lindley, 1957](#ref-lindley)). That may be probabilistically true, but that is why NHST testing for group differences is testing at an acceptable alpha level or, in the case of equivalence or non-inferiority testing, looking to determine whether an intervention performs the same as Standard of Care or at least does not perform worse given an acceptable level of error ([Mazzolari, Porcelli, Bishop, & Lakens, 2022](#ref-mazzolari2022)). Moreover, this statement ignores the Neyman-Pearson approach of balancing type 1 and type 2 errors. A statistician trained in the Neyman-Pearson approach would know that the alpha level could be lowered in situations where negligible effects could be detected (thereby balancing the type 1 and type 2 error rates), or secondary equivalence testing could be utilized to test whether significant effects of a small effect ([Campbell & Gustafson, 2018](#ref-campbell2018)).

The related statement in The Editorial that “All trials should therefore identify an effect” (Table 1), is simply inaccurate and not justifiable in any case that we can envision; though, it is often unclear if The Editorial is talking about an effect measured by a statistical model/test (which can always be wrong) or a “real” effect which can never be truly known in empirical work. Finally, there is a bit of irony in that while The Editorial states, “it is possible to put a confidence interval around any statistic, regardless of its use, including mean difference, risk, odds, relative risk, odds ratio, hazard ratio, correlation, proportion, absolute risk reduction, relative risk reduction, number needed to treat, sensitivity, specificity, likelihood ratios, diagnostic odds ratios, and difference in medians.”

They omit the fact that SWE or Minimal Clinically Important Difference (MCID) metrics can and should also be reported with confidence intervals. These estimates of “clinical relevance” are subject to the same sampling errors as an estimate of treatment effect.

# 2 Smallest Worthwhile Effect and Minimal Clinically Important Difference Values are Flawed

A failure to recognize the empirical ambiguity in the SWE/MCID metric is a fatal flaw in The Editorial as the primary thesis and remediation for supposed ills of NHST are to examine the overlap between effect estimates and the SWE/MCID. Many researchers, including one author of this manuscript ([Tenan et al., 2020](#ref-tenan2020)), have noted that there are a multitude of issues with SWE/MCID measures reported in the literature.

* Potential Issues with MCID

1. Not all measures have SWE or MCIDs in the literature, something The Editorial overtly recognizes.
2. There is no consensus, accepted calculation for SWE or MCID metrics. To our count, there are at least nine ways that these have been derived in the literature ([Ferreira, 2018](#ref-ferreira2018)).
3. The vast majority or nearly all SWE/MCID metrics reported in the physiotherapy literature do not meet the criteria for SWE conventions set out in by [Ferreira](#ref-ferreira2018) ([2018](#ref-ferreira2018)), which is the SWE manuscript The Editorial cites supporting SWE/MCID use.
4. Whether developing a SWE/MCID via [Ferreira](#ref-ferreira2018) ([2018](#ref-ferreira2018)) criteria or the more common ROC analysis anchoring to another scale, such as the Global Rating of Change scale, this requires dichotomizing an interval or continuous scale. This dichotomization is frequently arbitrary and subject to researcher discretion, making it a substantial source of variance between studies creating SWE/MCID metrics. In general, dichotomization should be avoided at all costs in medical research ([Senn, 2005](#ref-senn2005dichotomania)).
5. For a given SWE/MCID, there are often many different anchors reported in the literature, resulting in different SWE/MCID metrics, even for the same population.
6. Many SWE/MCID are likely biased by regression-to-the-mean, as they use change scores without accounting for the baseline measurement ([Tenan et al., 2020](#ref-tenan2020)).
7. A SWE/MCID, itself, is a point estimate based upon work performed in a sample of the population which is theorized to generalize to that population; as such, all SWE/MCID metrics should be reported with and understood to have confidence intervals around the reported point estimates

Point #7 on the list is what we will primarily discuss throughout the rest of this manuscript, though any one of the above listed issues, in isolation, should give the Editor-in-Chiefs’ who composed The Editorial pause when suggesting that the estimate CI overlap with a SWE/MCID should be used to fully supplant NHST. The Editorial states “If the estimate and the ends of its confidence interval are all more favourable than the smallest worthwhile effect, then the treatment effect can be interpreted as typically considered worthwhile”; however, this “smallest worthwhile effect” (SWE/MCID) is being treated as some sort of immutable ground-truth. In fact, an empirically derived SWE/MCID is, by it’s very nature, going to be derived from a sample of the population and thus have confidence intervals around that point estimate. If we ignore points 1-6 on the previous list, and pretend that the only issue with SWE/MCID measures is that they are not immutable ground-truths, but rather another estimate to compare against, do we have a path forward as The Editorial suggests? Ironically, we do and it is through NHST! If the estimate and the 95% Confidence intervals around both the SWE/MCID and the research study’s effect are reported, then it can be statistically determined if these two estimates are different from each other via the following method articulated by [Altman](#ref-altman2003) ([2003](#ref-altman2003)) where the estimates for the study result is and the SWE/MCID estimate is and their respective standard errors are represented as and .

* Steps to back calculate significance

1. Assume that the 95% CIs are parametric in nature and back-calculate the Standard Errors (SE) for each estimate ([Higgins, Li, & Deeks, 2019](#ref-cochrane6)) using the upper limit (UL) and lower limit (LL) of the CI.
2. Calculate the difference () in estimates
3. Calculate the Standard Error of the Difference ()
4. Calculate the -score
5. The -score can then be used to test of the null hypothesis that, in the population, the difference, , is zero by referencing the calculated -score against the normal distribution -table found in the back of many statistics textbooks.

# 3 Alternative Hypothesis Tests

The Editorial sets out to tell researchers that p-values should never be used but the only statistically valid method which makes their suggested alternative to NHST is, in fact, a p-value. While the above procedure for assessing differences between a research study estimate and an empirical SWE/MCID certainly could be performed, the list of 7 issues we’ve identified with SWE/MCID metrics leads us to believe that the proposed new method is inferior to current standards of practice in their journals, if it has any face validity at all. The authors of The Editorial seem to not understand that there are NHST based tests that could be used in the context they describe. Most of their criticism seems to be aimed at weak hypothesis tests with uninteresting null hypotheses (e.g., a null hypothesis of no effect). However, there are a multitude of hypothesis tests. As pointed out in numerous methodological tutorials, there are superiority, equivalence, non-inferiority, and even minimal effects hypothesis tests ([Caldwell & Cheuvront, 2019](#ref-caldwell2019); [Mazzolari, Porcelli, Bishop, & Lakens, 2022](#ref-mazzolari2022)). Therefore, many of the goals outlined in the The Editorial could very well be accomplished with NHST and p-values.

## 3.1 Vignette on Conditional Equivalence Testing

Let’s suppose we have a study of 2 therapies where there is a standard of care (SC) and a new experimental therapy (EXP) that could improve patient outcomes on variable X. Now, in the early stages of testing EXP we may not have any specific directional hypotheses, and any differences that are caused by EXP would be worthwhile findings. So, in this case, we may want to perform a simple t-test on the mean differences where the null hypothesis is zero *and* perform two one-sided test (TOST) to test for equivalence. This type of test can be accomplished in almost any statistical program (e.g., R, SPSS, SAS, jamovi, JASP, or Stata). However, an author this comment (ARC) has specifically created functions for this purpose in the [TOSTER](https://aaroncaldwell.us/TOSTERpkg) R package.

For the purposes of this vignette, let us assume a difference of less than 1 unit between EXP and SC would be considered practically equivalent. The choice of the equivalence bound is arbitrary and may vary depending on the purpose of the study. Some researchers may use some type of SWE/MCID to set the equivalence, but, as we mentioned above, even these empirically derived equivalence bounds are subject to sampling error. There are many subjective and objective methods of setting an equivalence bound ([Lakens, Scheel, & Isager, 2018](#ref-lakens2018)), and researchers should be careful in describing why and how they set their equivalence bounds.

We can then use the tsum\_TOST function within the package to perform the required statistical tests.

library(TOSTER)  
  
test1 = tsum\_TOST(  
 m1 = 0.01, # mean of SC  
 sd1 = 1.75, # SD of SC  
 n1 = 55, # Sample Size for SC  
 m2 = 0.75, # Mean of EXP  
 sd2 = 2.1, # SD of EXP  
 n2 = 49, # Sample Size for EXP  
 hypothesis = "EQU",  
 low\_eqbound = -1,  
 high\_eqbound = 1  
)  
  
test1$decision$ttest

## [1] "The null hypothesis test was non-significant, t(93.84) = -1.939, p = 5.55e-02"

test1$decision$TOST

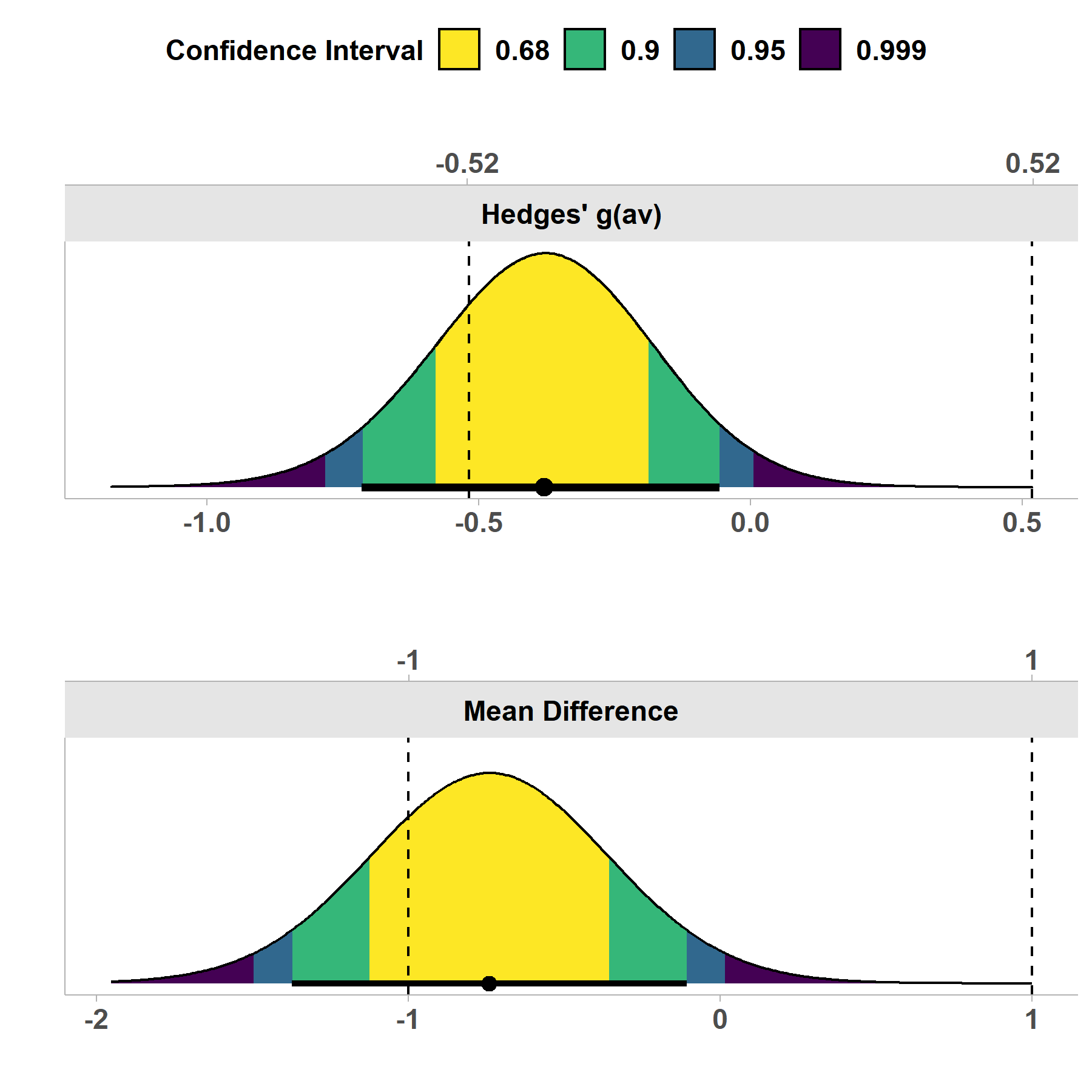
## [1] "The equivalence test was non-significant, t(93.84) = 0.681, p = 2.49e-01"

test1$decision$combined

## [1] "NHST: don't reject null significance hypothesis that the effect is equal to zero \nTOST: don't reject null equivalence hypothesis"

We can also provide a plot of the estimates with multiple confidence intervals.

plot(test1)



*Figure* *1.*  A visualization of the cumulative distribution function with 4 levels of confidence being displayed for the standardized mean difference (top panel) and the mean difference (bottom panel)

The interpretation provided in the output of the function takes a Neyman-Pearson perspective. Both the NHST and TOST tests have an alpha-level of 0.05 and neither test reaches significance. Therefore, an author using this approach would have to conclude that neither null hypothesis is rejected and the collected data is inconclusive about the effects of EXP. The answer may be unsatisfactory but is nonetheless straightforward.

However, those who wish to use an estimation approach may have a different interpretation. Under the approach outlined by [Rafi & Greenland](#ref-rafi2020) ([2020](#ref-rafi2020)), we could instead look at the data and see how “compatible” the data is with each competing hypothesis (i.e., NHST versus TOST). From this perspective, the interpretation is much more fluid, and one could conclude that the data is more incompatible with “no effect” than “equivalence” (p-values of 0.055 and 0.248, respectively).

Either perspective is valid and it is up to researchers to decide how they plan to tests or estimate their effects. If they are in the early stages of testing and evaluating EXP, then an estimation approach with a more flexible interpretation may be more appropriate. However, if EXP has already been well studied and the current study is meant to firmly establish its efficacy against SC then a more strict Neyman-Pearson test.

# 4 Conclusions

We are sad to see yet another example of sport and exercise scientists making claims about statistics beyond their expertise ([Sainani et al., 2020](#ref-sainani2020)). The unfortunate reality is that authoritative papers such as The Editorial can do real damage to the field of physiotherapy. First, the incorrect information provided in The Editorial will undoubtedly mislead physiotherapy researchers towards worse statistical practices and misinformed beliefs about statistics. Second, The Editorial hurts the reputation of the field of physiotherapy by giving the impression that the field is uninformed and has a poor understanding of the very basic concepts of statistics. Misguided commentaries from editorial boards are nothing new within academic publishing ([Mayo, 2021](#ref-mayo2021)). We would caution all non-statisticians to avoid making such sweeping statements about proper statistical practice, such as those made in The Editorial, without the involvement of a variety of statisticians. Even statisticians have diverse viewpoints on how statistics should be applied to the analysis of data (e.g., Frequentism versus Bayesian schools of thought), and editorial commentaries should not be the place for picking philosophical sides. Instead, editorial commentaries should be focused on improving the reporting of statistics within their journals and the publication of didactic papers on statistical practices authored by individuals with formal analytical training, such as the “Statistics Notes” series that the British Medical Journal published 1994-2017 and which remain an invaluable reference series to this day.

# 5 Additional Information

## 5.1 Contributions

If not included in the header included here.

## 5.2 Acknowledgements

People who contributed to the work but do not fit our author criteria should be listed in the acknowledgments, along with their contributions. You must ensure that anyone named in the acknowledgments agrees to being so named. Funding sources should not be included in the acknowledgments, but in the section below.

## 5.3 Funding information

Please provide a list of the sources of funding, as well as the relevant grant numbers, where possible. List the authors associated with specific funding sources. You will also enter this information in a form during the submission process, but it must be repeated here.

## 5.4 Data and Supplementary Material Accessibility

This should list the database(s) and, if appropriate, the respective accession numbers and DOIs for all data or supplementary material for the manuscript that has been made publicly available on a trusted digital repository. If no data, code, or supplementary material are available for this manuscript then the reason for this should be explained here.

# References

Altman, D. G. (2003). Statistics notes: Interaction revisited: The difference between two estimates. *BMJ*, *326*(7382), 219–219. <https://doi.org/10.1136/bmj.326.7382.219>

Altman, D. G., & Bland, J. M. (2011). How to obtain the confidence interval from a p value. *BMJ*, *343*(aug08 1), d2090–d2090. <https://doi.org/10.1136/bmj.d2090>

Caldwell, A. R., & Cheuvront, S. N. (2019). Basic statistical considerations for physiology: The journal temperature toolbox. *Temperature*, *6*(3), 181–210. <https://doi.org/10.1080/23328940.2019.1624131>

Campbell, H., & Gustafson, P. (2018). Conditional equivalence testing: An alternative remedy for publication bias. *PLOS ONE*, *13*(4), e0195145. <https://doi.org/10.1371/journal.pone.0195145>

Elkins, M. R., Pinto, R. Z., Verhagen, A., Grygorowicz, M., Söderlund, A., Guemann, M., … Sheikh, U. (2022). Statistical inference through estimation: Recommendations from the international society of physiotherapy journal editors. *Journal of Physiotherapy*, *68*(1), 1–4. <https://doi.org/10.1016/j.jphys.2021.12.001>

Ferreira, M. (2018). Research note: The smallest worthwhile effect of a health intervention. *Journal of Physiotherapy*, *64*(4), 272–274. <https://doi.org/10.1016/j.jphys.2018.07.008>

Higgins, J. P., Li, T., & Deeks, J. J. (2019). Choosing effect measures and computing estimates of effect. *Cochrane Handbook for Systematic Reviews of Interventions*, 143–176. Retrieved from <https://training.cochrane.org/handbook/current/chapter-06>

Hoekstra, R., Morey, R. D., Rouder, J. N., & Wagenmakers, E.-J. (2014). Robust misinterpretation of confidence intervals. *Psychonomic Bulletin & Review*, *21*(5), 1157–1164. <https://doi.org/10.3758/s13423-013-0572-3>

Lakens, D., Scheel, A. M., & Isager, P. M. (2018). Equivalence testing for psychological research: A tutorial. *Advances in Methods and Practices in Psychological Science*, *1*(2), 259–269. <https://doi.org/10.1177/2515245918770963>

Lindley, D. V. (1957). A statistical paradox. *Biometrika*, *44*(1-2), 187–192. <https://doi.org/10.1093/biomet/44.1-2.187>

Mayo, D. G. (2021). The statistics wars and intellectual conflicts of interest. *Conservation Biology*. <https://doi.org/10.1111/cobi.13861>

Mazzolari, R., Porcelli, S., Bishop, D. J., & Lakens, D. (2022). Myths and methodologies: The use of equivalence and non-inferiority tests for interventional studies in exercise physiology and sport science. *Experimental Physiology*. <https://doi.org/10.1113/ep090171>

Morey, R. D., Hoekstra, R., Rouder, J. N., Lee, M. D., & Wagenmakers, E.-J. (2015). The fallacy of placing confidence in confidence intervals. *Psychonomic Bulletin & Review*, *23*(1), 103–123. <https://doi.org/10.3758/s13423-015-0947-8>

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*, *20*(1). <https://doi.org/10.1186/s12874-020-01105-9>

Sainani, K. L., Borg, D. N., Caldwell, A. R., Butson, M. L., Tenan, M. S., Vickers, A. J., … Bargary, N. (2020). Call to increase statistical collaboration in sports science, sport and exercise medicine and sports physiotherapy. *British Journal of Sports Medicine*, *55*(2), 118–122. <https://doi.org/10.1136/bjsports-2020-102607>

Senn, S. (2005). Dichotomania: An obsessive compulsive disorder that is badly affecting the quality of analysis of pharmaceutical trials. *Proceedings of the International Statistical Institute, 55th Session, Sydney*.

Tenan, M. S., Simon, J. E., Robins, R. J., Lee, I., Sheean, A. J., & Dickens, J. F. (2020). Anchored minimal clinically important difference metrics: Considerations for bias and regression to the mean. *Journal of Athletic Training*, *56*(9), 1042–1049. <https://doi.org/10.4085/1062-6050-0368.20>