Recommendations for statistical analysis involving null hypothesis significance testing

Authors:

Andrew J. Harrison, Biomechanics Research Unit, Department of Physical Education and Sport Sciences, University of Limerick, Limerick, Ireland.

Stuart A. McErlain-Naylor, School of Health and Sports Sciences, University of Suffolk, Ipswich, UK.

Elizabeth J. Bradshaw, Centre for Sport Research, School of Exercise and Nutrition Sciences, Deakin University, Melbourne, Australia. Sport Performance Research Institute New Zealand, Auckland University of Technology, Auckland, New Zealand.

Boyi Dai, Division of Kinesiology and Health, University of Wyoming, Laramie, USA.

Hiroyuki Nunome, Faculty of Sports and Health Science, Fukuoka University, Nanakuma, Jonan-ku, Fukuoka, Japan.

Gerwyn T.G. Hughes, Department of Kinesiology, University of San Francisco, California, USA.

Pui W. Kong, Physical Education and Sports Science Academic Group, National Institute of Education, Nanyang Technological University, Singapore.

Benedicte Vanwanseele, Human Movement Biomechanics Research Group, Department of Movement Sciences, KU Leuven, Leuven, Belgium.

J. Paulo Vilas-Boas, Faculty of Sport, Centre of Research, Educaton, Innovation and Intervention in Sport and Porto Biomechanics Laboratory. University of Porto, Porto, Portugal.

Daniel T.P. Fong*, National Centre for Sport and Exercise Medicine, School of Sport, Exercise and Health Sciences, Loughborough University, Loughborough, UK.

*Corresponding author, email: d.t.fong@lboro.ac.uk

Background

The peer review process of original research articles generally requires authors/researchers to adopt accepted scientific methods, identify testable hypotheses and test those hypotheses using appropriate and established statistical methods. In *Sports Biomechanics*, authors are encouraged to submit original research articles that conform to these norms. Despite widespread use, null hypothesis significance testing (NHST) has received criticism on various counts, especially when there is a reliance on *p*-values alone (as defined below) for NHST. The *p*-value combines sample size, variance and differences in values within the calculation but its meaning is somewhat subtle and difficult to communicate in nontechnical language, leading to over-simplification, distortions of its meaning and misinterpretation.

> Technically a *p*-value is the probability of obtaining an effect at least as extreme as the one in your sample data, assuming the truth of the null hypothesis. (Chang et al., 2019)

Often *p*-values are interpreted as an estimate of the error rate in rejecting the null hypothesis, suggesting that a *p*-value of 0.05 means the probability of a type I error is 5%. However, the true error rate (false discovery of positive cases or effects) in an isolated study may range from 23–50% when $p = 0.05$ (Vidgen & Yasseri, 2016). This has led some to criticise the use of *p*-values in NHST as being more liberal than is generally understood (Sellke et al., 2001; Vidgen & Yasseri, 2016).

Various limitations have been identified when using NHST (*p*-values), including, amongst others:

- Use of *p*-values in isolation without reference to effect sizes and confidence intervals;
- The potential for '*p* hacking' where data and analyses can be deliberately manipulated to reduce *p*-values;
- Simplistic dichotomous interpretations of *p*-values as either significant or nonsignificant;
- Incorrect interpretation of $p > 0.05$ as meaning no effect;
- Whether zero effect is really/always the comparison of clinical or practical interest (*i.e*. the smallest effect size of interest);

- Misinterpreting statistical significance as clinical or practical significance;
- Performing multiple statistical tests without adjusting the criterion *p*-value.

Despite the limitations, critics generally acknowledge that NHST remains in common use and many of its problems stem more from misunderstanding and misuse than from inherent flaws (Miller, 2009).

Magnitude-based inference

As a consequence of the issues highlighted above, the journal *Basic and Applied Social Psychology* moved to ban null hypothesis significance testing (Trafimow & Marks, 2015). This included *p*-values, associated test statistics (*e.g.*, *t*-values and *F*-values), confidence intervals, and statements about 'significant' differences or lack thereof. Instead, the journal encouraged larger sample sizes and interpretation of 'strong descriptive statistics' such as effect sizes. However, an analysis of articles published after the ban found multiple instances of authors overstating their conclusions beyond what would have been supported had statistical significance been considered (Fricker et al., 2019). Of further concern, the information necessary for readers to identify this was not readily available. Many in sport and exercise science have taken a more balanced approach, stating that outcomes should not be evaluated solely on the basis of probabilities (*i.e. p*-values) but should also consider effect sizes and confidence intervals in relation to minimum clinically or practically important effects (Winter et al., 2014; Zhu, 2012). These same recommendations have been made within previous *Sports Biomechanics* articles (Knudson, 2009, 2017).

An alternative method of inference, named 'magnitude-based inference' (MBI) (Hopkins et al., 2009) or more recently 'magnitude-based decisions' (Hopkins, 2017), was proposed and claimed to address many of the concerns already mentioned (Batterham & Hopkins, 2006). Specifically, MBI aimed to address an over-reliance on probability and the inattention to both magnitude (*i.e*. effect size) and certainty (*i.e*. confidence interval) of reported effects (Batterham & Hopkins, 2006). MBI classifies effect sizes as harmful/negative, trivial (commonly defined as between −0.2 and +0.2 standardised effect size units), or beneficial/positive (Batterham & Hopkins, 2006; Hopkins et al., 2009). The probability that the true effect falls within one of these three categories is then assigned, based upon an interpretation of confidence intervals and *p*-values.

Critique of magnitude-based inference

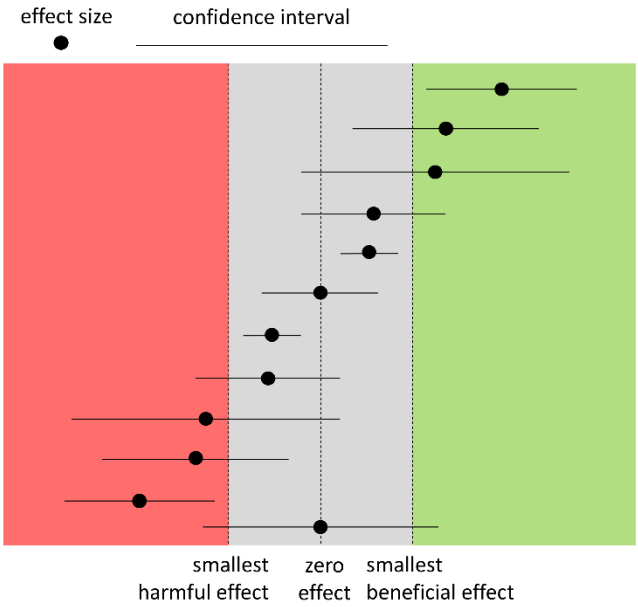
MBI calculates one-sided *p*-values for benefit and harm (Welsh & Knight, 2015). As defined above, a one-sided *p*-value for benefit of 0.10 means that there is a 10% chance of an effect at least as large as that observed arising if the intervention was not beneficial (*i.e*. if the true effect is ≤ 0.2). Recall that the *p*-value *does not* report the probability that the true effect is beneficial or not beneficial given the observed data. MBI incorrectly interprets this example as a 10% chance that the intervention is not beneficial and a 90% chance that the intervention is beneficial (Sainani et al., 2019). A probability greater than 75% (*i.e*. one-sided *p*-value for benefit < 0.25) is qualitatively interpreted as a 'likely' benefit, with a 25% probability threshold (*i.e*. one-sided *p*-value for benefit < 0.75) for 'possible' benefit. Effects are typically considered unclear if the probabilities of harm and benefit are both ≥ 5% (Batterham & Hopkins, 2006; Sainani et al., 2019), equivalent to a 90% confidence interval overlapping both harmful and beneficial effect sizes.

The MBI interpretation detailed above has been shown via simulation and mathematics to create peaks of false positives (type I errors: falsely rejecting the null hypothesis) at specific small-to-moderate sample sizes (Sainani, 2018). At these typically used sample sizes, MBI reduces the type II error rate (false negatives: falsely accepting the null hypothesis) but increases the type I error rate by two to six times compared to standard null hypothesis significance testing (Sainani, 2018). Further critiques by statisticians have highlighted that MBI sample size calculators may underestimate required sample sizes (Welsh & Knight, 2015), and that MBI conflates frequentist and Bayesian methods of statistical inference (Sainani et al., 2019).

Bayesian inference does not calculate a single probability (*i.e*. a *p*-value), but rather calculates a full probability distribution (see Kruschke & Liddell, 2018 for an introduction to Bayesian inference). Using this posterior probability distribution (our updated knowledge regarding the uncertain parameter estimate), 'credible intervals' can provide the probability that the true effect falls within a certain range. For example, a 95% Bayesian credible interval of 0.05 to 0.3 represents a 95% chance the true effect size lies between 0.05 and 0.3. In contrast, a 95% frequentist confidence interval represents the range of effect sizes that would not be rejected at $p < 0.05$. MBI, however, interprets frequentist confidence intervals as if they were Bayesian credible intervals (Sainani et al., 2019). This equivalence is only true when one assumes a 'flat'

or ‘uniform’ prior distribution of possible parameter estimates (Van Zwet, 2019). This would mean that all effect sizes were equally likely - unrealistic for most applications in the sport and exercise sciences. A ‘flat’ prior will lead to overly optimistic inferences due to the starting assumption of an almost certainly non-trivial effect. It is further noted that the MBI method has never been published in a journal following peer review by statisticians. These critiques have led some sport and exercise science journals to caution against the use of MBI (Nevill et al., 2018) or stop accepting manuscripts using MBI altogether (Gladden, 2019).

To correctly interpret effects in relation to practically or clinically meaningful thresholds, researchers must first specify the range of effect sizes which are considered to be *not* meaningful. Where possible, this decision should be informed by existing literature and not arbitrarily chosen as representative of a ‘small’ effect. For example, evidence may or may not support a ‘smallest effect size of interest’ of +0.2, similar to that commonly used in MBI (Lohse et al., 2020). A 95% confidence interval that excludes a value (*e.g.*, the smallest effect size of interest) implies that value can be rejected in a two-sided hypothesis test using an alpha of 0.05 (likewise a 90% confidence interval for a one-sided hypothesis test). Therefore, an effect could be considered meaningful if the entire confidence interval was greater than the smallest effect size of interest. Combining effect sizes and confidence intervals in this way allows interpretation of both the magnitude of an effect and the level of (un)certainty around that parameter estimate. With an increase in sample size for any given effect, the confidence interval would narrow around the parameter estimate (increased certainty/decreased uncertainty) just as a *p*-value would decrease. Unlike isolated null hypothesis significance testing or magnitude-based inference, this approach ensures that effects smaller than the smallest effect size of interest cannot be considered ‘significant’ or ‘beneficial’ regardless of an over-inflated sample size (null hypothesis significance testing) or a ‘beneficial’ central parameter estimate (see Figure 1). Such an approach has been frequently recommended in the sport and exercise sciences in recent years (Knudson, 2009, 2017; Zhu, 2012) and the reporting of effect sizes and confidence intervals has been made compulsory in certain journals (Winter et al., 2014).



recommended interpretation	NHST *p*-values alone	typical MBI interpretation
beneficial	significantly beneficial	very likely beneficial
not harmful	significantly beneficial	likely beneficial
not harmful	not significant	likely beneficial
not harmful	not significant	possibly beneficial
trivial	significantly beneficial	trivial
trivial	not significant	trivial
trivial	significantly harmful	trivial
not beneficial	not significant	possibly harmful
not beneficial	not significant	likely harmful
not beneficial	significantly harmful	likely harmful
harmful	significantly harmful	very likely harmful
equivocal	not significant	unclear

Figure 1. Recommended interpretation, NHST *p*-values alone, and typical MBI nterpretation. In this case NHST refers to one-sided null hypothesis significance testing. Figure is based on Barker and Schofield (2008) and is an illustrative guide only.

Sports Biomechanics recommendations

Given the current status in sport and exercise science and in statistical science, where polarised arguments have been presented for and against the use of NHST and/or MBI, it is now important that *Sports Biomechanics* presents clear recommendations on the acceptability of research methods and procedures for statistical analysis. Therefore, the Editorial Board wish to reiterate that original research articles submitted to *Sports Biomechanics* should provide clear statements of the rationale for the research, hypotheses or research questions under investigation, and robust and accepted procedures for testing hypotheses and making inferences.

Authors are advised to ensure the following recommendations are satisfied in original research submissions to *Sports Biomechanics*:

- Clearly state the testable hypotheses of the investigation, and distinguish between primary and secondary outcomes where appropriate;
- Match the analysis techniques to the purpose of the research;
- Justify the sample size used (Knudson, 2017) — power analysis (*e.g.*, G*Power: Faul et al., 2007) is one acceptable way to do this, but some research (*e.g.*, case studies or

using elite athletes, etc.) may justifiably involve small samples where large samples may not be possible in specialist populations. This should be considered when interpreting the results;

- Ensure that hypotheses are tested using robust and accepted statistical methods;
- Cite the sources of statistical methods used and the version numbers of statistical analysis software (we encourage use of methods that are peer-reviewed in statistical journals);
- Ensure all assumptions of the statistical tests are satisfied; use non-parametric tests when the assumptions for parametric tests are not satisfied;
- Control for multiple comparisons (multiple hypotheses) with a justified correction using robust procedures such as the Bonferroni, Benjamini and Hochberg, or Holm's step-down adjustment methods;
- NHST *p*-values are acceptable but should not be used in isolation for inference (Wasserstein & Lazar, 2016);
- Report exact *p*-values to 3 decimal places rather than for example, $p < 0.05$;
- When exact *p*-values are less than 0.001 ensure $p < 0.001$ is reported but never $p = 0.000$;
- Not interpret $p > 0.05$ in isolation as evidence of no effect;
- Use appropriate effect size indices and/or confidence intervals around effect sizes; consider Cohen's *d* and difference in means for comparisons, *r* or R2 for correlation, and beta coefficients for regression;
- Define the levels of effect sizes for clinical or practical significance;
- Interpret effect sizes in context to provide information on the practical or clinical importance of the effects observed.

Additionally, *Sports Biomechanics* will consider studies where justifiable replication of an experiment is used to verify findings (Knudson, 2017). Replication of experimental phenomena is a cornerstone of scientific method.

> No isolated experiment, however significant in itself, can suffice for the experimental demonstration of any natural phenomenon. (Fisher, 1937).

Lack of experimental replication has been identified as a problem in published research generally, since there can be a tendency to consider studies that replicate an experiment as not meeting the threshold of original research. Experimental replication should be an important part of the scientific process and be encouraged where necessary.

Recommendations on magnitude-based inference

Having considered the balance of arguments related to the MBI approach, the Editorial Board of *Sports Biomechanics* concludes that magnitude-based inference has not achieved general acceptance in the domain of Sport and Exercise Science and has been increasingly criticised by statisticians because of its tendency to inflate the rate of false positive outcomes. Consequently, the Editorial Board discourages the use of MBI methods, and will reject submissions using MBI without review until such time that the method is demonstrated as robust and gains general recognition within science and statistics.

Acknowledgements

We would like to thank Dr Kristin Sainani from Stanford University for her advice on this article.

Disclosure statement

No potential conflict of interest was reported by the author(s).

References

Barker, R., & Schofield, M. (2008). Inference about magnitudes of effects. *International Journal of Sports Physiology and Performance*, *3*, 547-557. doi:10.1123/ijspp.3.4.547

Batterham, A. M., & Hopkins, W. G. (2006). Making meaningful inferences about magnitudes. *International Journal of Sports Physiology and Performance*, *1*, 50–57. doi:10.1123/ijspp.1.1.50

Chang, M., Balser, J., Roach, J., & Bliss, R. (2019). *Innovative strategies, statistical solutions and simulations for modern clinical trials*. CRC Press, Taylor & Francis Group.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*, 175–191. doi:10.3758/bf03193146

Fisher, R. A. (1937). *The design of experiments* (2nd ed.). Oliver and Boyd.

Fricker, R. D., 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, 73*, 374–384. doi:10.1080/00031305.2018.1537892

Gladden, L. B. (2019). Editorial note to Batterham and Hopkins letter and Sainani response. *Medicine and Science in Sports and Exercise*, *51*, 601. doi:10.1249/MSS.0000000000001825

Hopkins, W. G., Marshall, S. W., Batterham, A. M., & Hanin, J. (2009). Progressive statistics for studies in sports medicine and exercise science. *Medicine and Science in Sports and Exercise*, *41*, 3–13. doi:10.1249/MSS.0b013e31818cb278

Hopkins, W. G. (2017). *Spreadsheets for analysis of controlled trials, crossovers and time series*. Retrieved May 15, 2020, from Sportscience website: sportsci.org/2017/wghxls.htm

Knudson, D. (2009). Significant and meaningful effects in sports biomechanics research. *Sports Biomechanics*, *8*, 96–104. doi:10.1080/14763140802629966

Knudson, D. (2017). Confidence crisis of results in biomechanics research. *Sports Biomechanics*, *16*, 425–433. doi:10.1080/14763141.2016.1246603

Kruschke, J. K., & Liddell, T. M. (2018). Bayesian data analysis for newcomers. *Psychonomic Bulletin & Review*, *25*, 155–177. doi:10.3758/s13423-017-1272-1

Lohse, K.R., Sainani, K.L., Andrew Taylor, J., Butson, M.L., Knight, E.J., & Vickers, A.J. (2020). Systematic review of the use of "magnitude-based inference" in sports science and medicine. PLoS ONE, 15, e0235318. doi:10.1371/journal.pone.0235318.

Miller, J. (2009). What is the probability of replicating a statistically significant effect. *Psychonomic Bulletin & Review*, *16*, 617–640. doi:10.3758/PBR.16.4.617

Nevill, A. M., Williams, A. M., Boreham, C., Wallace, E. S., Davison, G. W., Abt, G., Lane, A. M., Winter, E. M., & Board, E. (2018). Can we trust "Magnitude-based inference"? *Journal of Sports Sciences, 36*, 2769–2770. doi:10.1080/02640414.2018.1516004

Sainani, K. L. (2018). The problem with "Magnitude-based Inference". *Medicine and Science in Sports and Exercise, 50*, 2166–2176. doi:10.1249/MSS.0000000000001645

Sainani, K. L., Lohse, K. R., Jones, P. R., & Vickers, A. (2019). Magnitude-based inference is not Bayesian and is not a valid method of inference. *Scandinavian Journal of Medicine & Science in Sports*, *29*, 1428–1436. doi:10.1111/sms.13491

Sellke, Y., Bayarri, M. J., & Berger, J. O. (2001). Calibration of p-values for testing precise null hypotheses. *The American Statistician*, *55*, 62–71. doi:10.1198/000313001300339950

Trafimow, D., & Marks, M. (2015). Editorial. *Basic and Applied Social Psychology*, *37*, 1–2. doi:10.1080/01973533.2015.1012991

van Zwet, E. (2019). A default prior for regression coefficients. *Statistical Methods in Medical Research, 28,* 3799–3807. doi:10.1177/0962280218817792

Vidgen, B., & Yasseri, T. (2016). P-values: Misunderstood and misused. *Frontiers in Physics*, *4*, 6. doi:10.3389/fphy.2016.00006

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA statement on p-values: Context, process, and purpose. *The American Statistician, 70,* 129–133. doi:10.1080/00031305.2016.1154108

Welsh, A. H., & Knight, E. J. (2015). "Magnitude-based inference": A statistical review. *Medicine and Science in Sports and Exercise*, *47*, 874–884. doi:10.1249/MSS.0000000000000451

Winter, E. M., Abt, G. A., & Nevill, A. M. (2014). Metrics of meaningfulness as opposed to sleights of significance. *Journal of Sports Sciences*, *32*, 901–902. doi:10.1080/02640414.2014.895118

Zhu, W. (2012). Sadly, the earth is still round ($p < 0.05$). *Journal of Sport and Health Science*, *1*, 9–11. doi:10.1016/j.jshs.2012.02.002