SPORTSCIENCE · sportsci.org Latest issue
Original Research / Research Resources This issue

Discussion Paper on Sampling Uncertainty for Editorial Boards

Will G Hopkins

Sportscience 25, 10-16, 2021 (sportsci.org/2021/SamplingPaper.htm)
Institute for Health and Sport, Victoria University, Melbourne, Australia. Email. Reviewer: Dan Cleather, School of Sport, Health and Applied Science, St Mary's University, Twickenham, UK.

The compatibility (or confidence) interval is probably the best measure of uncertainty in the value of an effect derived from a sample. The interval, or the sampling distribution from which it is derived, is the basis of the following methods for assessing acceptable sampling uncertainty: informal assessment of the interval as precision of estimation; Bayesian assessment of the interval or sampling distribution with minimally or other informative priors; the nil-hypothesis significance test; and tests of substantial and non-substantial hypotheses. Editorial boards should decide which of these methods they would prefer to see in submitted manuscripts. For theoretical and practical reasons, I recommend Bayesian assessment and tests of substantial and non-substantial hypotheses; I also recommend magnitude-based decisions, which is consistent with both these methods.

Keywords: Bayesian inference; compatibility interval; confidence interval; inferiority, superiority and equivalence testing; magnitude-based inference; magnitude-based decisions; nil-hypothesis significance test; p value; precision of estimation.

Reprint pdf · Reprint docx

Update 22 Dec. I have added a paragraph arguing against the use of a meta-analyzed mean as an informative prior in a Bayesian analysis, except in a special case. And at the end of the recommendation section I now correctly assert that statistical significance is not necessary for an effect to be decisively substantial.

Update 8-10 Sept. Evidence provided by significance and non-significance is now described in terms of necessary and/or sufficient for substantial and non-substantial at the end of the recommendations section.

Any effect in a study of a sample suffers from sampling uncertainty, whereby another sample would produce a different value of the effect. Journal editors have to make decisions about acceptable sampling uncertainty of effects in manuscripts submitted for publication. For example, if the sample size is too small, the uncertainty will be too large for the effect to be in any sense decisive, so the manuscript would usually be rejected. The best measure of sampling uncertainty is probably the compatibility (or confidence) interval, which shows a range of values of the effect that are most compatible with the data and statistical model used to derive the effect (Rafi & Greenland, 2020). Several methods of assessing acceptability of the uncertainty are available, all based on coverage (width and disposition) of the compatibility interval or coverage of the sampling distribution used to derive the compatibility interval. What follows is my summary of the methods and my recommendations.

Informal assessment of the compatibility interval as precision of estimation

This method is promoted by Ken Rothman in his epidemiological texts (e.g., Rothman, 2012) and by psychologist Geoff Cumming in his "new statistics" (e.g., Cumming, 2014). Narrower compatibility intervals obviously represent more precise estimates. Unfortunately, these authors offer little guidance on what level of compatibility is appropriate (e.g., 90%, 95%, 99%), or on how narrow the interval needs to be for a given effect in a given setting. Even in an article on estimating sample size for adequate precision, Rothman and Greenland (2018) offer no advice about desired level and width of the compatibility interval, but where the interval sits in relation to certain values (e.g., beneficial, trivial, harmful) can be important: one should aim for an interval that does or does not include such values, and the resulting conclusion about the effect is phrased in terms of level of compatibility or incompatibility of those values with the data and model. The interval is not regarded as possible values of the effect in the population from which the sample is drawn; the method would thereby be Bayesian, and it would attract criticism from strict Bayesians (discussed next).

Importantly, Rothman and Cumming regard their method as a replacement for statistical significance. For example, Rothman (2012) states "Estimation using confidence intervals allows the investigator to quantify separately the strength of a relation and the precision of an estimate and to reach a more reasonable interpretation… In most instances, there is no need for any test of statistical significance to be calculated, reported, or relied on, and we are much better off without them."

Quantitative assessment of the interval or sampling distribution

Here the compatibility interval or the sampling distribution centered on the observed value of the effect is interpreted overtly as probabilistic values of the true effect: the effect in the population from which the sample is drawn. Formally, this method is Bayesian assessment with a minimally informative prior, and it has been promoted by various authors (Albers et al., 2018; Burton, 1994; Shakespeare et al., 2001), including the progenitors of magnitude-based inference (Batterham & Hopkins, 2006; Hopkins & Batterham, 2016). Some authors (Barker & Schofield, 2008; Sainani et al., 2019; Welsh & Knight, 2015) have claimed that the method is not Bayesian, because a minimally informative prior is not "proper" (it cannot be included in Bayesian computations, because it has zero likelihood for all values of the effect) and because such a prior implies belief that unrealistically large values of the effect have the same likelihood (albeit zero) as realistic small values. These criticisms have been addressed (Batterham & Hopkins, 2015; Batterham & Hopkins, 2019; Hopkins & Batterham, 2008; Hopkins & Batterham, 2016), most recently by using Greenland's (2006) simplified approach for Bayesian computations to show that a realistic weakly informative normally distributed prior (excluding extremely large effects at the 90% level of confidence) makes no practical difference to the posterior probabilities of the true effect for any reasonable sample size (Hopkins, 2019); link to the article.

Quantitative assessment accounting for prior belief in the magnitude of the effect

This method is the traditional Bayesian assessment with informative priors. It is very rarely used in exercise and sport studies. The full Bayesian implementation is challenging, since a prior distribution has to be found and justified for every parameter in the statistical model. Greenland's (2006) simplified Bayesian method is more intuitive and easily implemented with a spreadsheet; link to the article. Whatever method is used, the problem with informative priors based on belief is that they are difficult to justify and quantify, and the more informative they are, the more they bias the effect. They therefore offer the researcher an opportunity to bias the effect towards a *desired* magnitude, by stating a prior centered on that magnitude. I would distrust any effect that was modified by prior belief.

A prior provided by a published meta-analysis is more trustworthy than a belief, but it will generally be problematic. The meta-analysis needs to have been performed with a random effect to allow for and estimate real differences in the effect between settings. When the resulting meta-analyzed mean effect is applied as a prior, it must include the uncertainty due to the real differences, because the researcher's setting is inevitably different from the mean of all settings. Inclusion of this uncertainty will make the prior more diffuse; if it then has any effect at all, it will improve the precision of the researcher's effect, but at the arguably unacceptable cost of biasing it by shrinking it towards the meta-analyzed mean. It is only when the meta-analysis shows that the real differences between settings are trivial that it makes sense to use a meta-analyzed prior; the researcher's effect will then differ from the meta-analyzed mean effect only because of sampling variation, and the Bayesian analysis will produce an unbiased estimate with precision equivalent to an increase in the researcher's sample size. (Note that it is not sufficient for the meta-analyst to omit the random effect on the basis of non-significance in a nil-hypothesis test for heterogeneity; the standard error of the random effect must be estimated and evaluated.)

The nil-hypothesis significance test

Researchers interpret statistical significance and non-significance as sufficient evidence that an effect is present (substantial) or absent (trivial or even zero). To this extent, NHST is a method for assessing sampling uncertainty, acceptable when the effect is significant. It continues to be used exclusively by almost all authors, in spite of concerns expressed by generations of statisticians, most prominently by Amrhein, Greenland and McShane (2019) exhorting us to "retire statistical significance". A task force convened by the president of the American Statistical Association has called for proper application and interpretation of statistical significance, but gives no guidance as to what is proper (Benjamini et al., 2021). See the Critical Inference blog for a critique of the ASA statement.

Better evidence about magnitudes is provided by tests of substantial and non-substantial hypotheses (discussed next). When such tests were applied to effects presented at a recent sport-science conference, NHST resulted in an unacceptable prevalence of misinterpretations (Hopkins, 2021); link to the article. The prevalence of misinterpretations depends on sample size, but "researchers should understand that the problem of misinterpretations with significance and non-significance is not solved by using the sample size estimated with a power calculation" (Hopkins, 2021).

Tests of substantial and non-substantial hypotheses

If researchers want to determine whether an effect is substantial and positive or negative, the obvious and only hypotheses to test (and to hope to reject) are the hypothesis that the effect is not substantially positive or not substantially negative, constituting so-called superiority and inferiority testing. (The names arise from effects representing the difference in two treatments, where a substantial difference represents inferiority or superiority of one treatment relative to the other.) Similarly, to determine whether an effect is trivial, the researcher should test (and hope to reject) the hypothesis that the effect is substantially positive and the hypothesis that the effect is substantially negative, which together constitute a so-called equivalence test. Although this approach has been available for decades, it has been used very little, presumably because it is much easier to use (and misinterpret) the p value for the nil-hypothesis test. Recently Lakens et al. (2018) have brought tests of substantial and non-substantial hypotheses to the attention of sport psychologists.

Magnitude-based inference

MBI is essentially a Bayesian method, in which the prior is minimally informative and the posterior probabilities of the true effect are calculated for substantial and trivial magnitudes demarcated by smallest important positive and negative (or beneficial and harmful) magnitudes of the effect. The quantitative probabilities are expressed with accessible qualitative terms ranging from most unlikely to most likely (Hopkins et al., 2009) using a scale similar to that of climate scientists (Mastrandrea et al., 2010). Guidelines are provided for what constitutes adequate precision for effects with and without clinical or practical relevance.

This method was gaining ground in exercise and sport studies, until it was criticized by several authors (Sainani, 2018; Welsh & Knight, 2015). In response to the criticisms, the developers of MBI showed that, as an empirical method, it has desirable characteristics, including well-defined and acceptably low error rates, potentially higher publication rates, and negligible publication bias compared with NHST (Hopkins & Batterham, 2016). More recently, MBI has been shown to be equivalent to substantial and non-substantial hypothesis testing, and it was renamed as magnitude-based decisions (MBD) (Hopkins, 2020); link to the article. Concerns that Bayesian assessment with a minimally informative prior is invalid have also been addressed (Hopkins, 2019; Hopkins & Batterham, 2008; Hopkins & Batterham, 2016). The most recent criticism of MBI or MBD is that it is misused, mainly by authors interpreting *possibly* and *likely* substantial as *decisively* substantial (Lohse et al., 2020). The claims of misuse were shown to be grossly exaggerated (Aisbett, 2020). In any case, such misuse is easily corrected during peer review.

Recommendations

In my experience, some statisticians defend their preferred method and criticize others with flawed logic and ad hominem arguments. An editorial board receiving advice on this discussion document from other statisticians should therefore inspect the advice carefully for well-reasoned and objective evidence supporting or refuting my assertions. It is not sufficient for a statistician to claim merely that the evidence supporting MBI or MBD has not been published in statistics journals; any statisticians making this claim must themselves provide convincing evidence that the evidence is flawed. The board should also be wary of such statisticians adopting privileged and mistaken views, the concern of a group (Cleather et al., 2021) responding to a call by Sainani et al. (2021) to increase statistical collaboration in the disciplines of sport science.

My advice to an editorial board for policy in respect of sampling uncertainty in manuscripts is as follows…

- Forbid use of nil-hypothesis testing, including the traditional p value, thereby preventing authors and readers from making misleading conclusions based on significance and non-significance, irrespective of whatever other methods they include for assessing sampling uncertainty. An editorial board should obviously think carefully before banning a specific method. As Greenland (2017) stated, "do not reject out of hand any methodology because it is flawed or limited, for a methodology may perform adequately for some purposes despite its flaws." Evidently NHST does not pass muster in the view of Greenland and others.
- Mandate use of compatibility intervals, preferably at the 90% level, which is consistent with alphas of 0.05 for substantial and non-substantial hypothesis tests. Showing 95% intervals would allow readers to see easily whether the interval excluded zero and therefore whether the effect was statistically significant.
- NHST has to be replaced by something. The choices are tests of substantial and non-substantial hypotheses with alphas appropriate for the setting and/or Bayesian estimation of probabilities of effect magnitudes. Bayes factors can be used to combine informative priors with tests of substantial and non-substantial hypotheses, but the computations are difficult (e.g., van Ravenzwaaij et al., 2019). The only other option is qualitative interpretation of the compatibility interval promoted by Rothman (2012) and Cumming (2014), but this method lacks practical guidelines.
- There will be readers of a frequentist persuasion who prefer hypothesis testing, and readers who prefer Bayesian assessment. It is therefore reasonable to ask authors to provide both in their manuscripts. MBD can be presented in a manner that achieves both. The latest iteration of a methods section describing both is shown below. A journal could provide something similar in its instructions to authors or in an opinion piece, which authors could refer to and thereby save ~500 words in manuscripts.
- Bayesian analysis with a minimally informative prior is preferable, but authors should use Greenland's (2006) Bayesian method to determine whether a realistic weakly informative prior produces substantial shrinkage (reduction in the magnitude of the observed value and width of the compatibility interval) of any effects, and they should report the outcomes accordingly. If a full Bayesian analysis is employed, the authors should convincingly justify the quantification of the uncertainty representing their prior beliefs. The original compatibility interval should always be shown, since informative priors may bias the outcome.
- Author guidelines should communicate the essence of the following advice… "Whichever approach researchers use, they should state clearly that a conclusion, decision or probabilistic statement about the magnitude of an effect is based on the uncertainty arising from sampling variation and is conditioned on assumptions about the data and the statistical model used to derive the compatibility interval and associated p values. The way in which violation of these assumptions could bias the outcome should be discussed and, where possible, investigated quantitatively. A straightforward method is sensitivity analyses, in which the width and disposition of the compatibility interval relative to smallest importants are determined for realistic worst-case violations." (Hopkins, 2021); link to the article.
- Given the inevitable additional uncertainties arising from violation of assumptions, we should ask authors to follow the advice of Greenland, Rothman, and no doubt others, to avoid the dichotomization of all hypothesis testing and instead to interpret qualitatively the strength of evidence for and/or against magnitudes.
- We cannot avoid the dichotomization implicit in deciding whether sampling uncertainty is acceptable for publication, and it seems to me that the least we should expect from authors for their primary outcome (in anything other than a pilot study) is a 90% compatibility interval that does not include substantial positive and negative values (equivalent to rejection of at least one substantial hypothesis, p<0.05, and a very unlikely substantial effect in a Bayesian analysis with a minimally in-

formative prior). Thereafter, Bayesian assessments provide qualitative evidence *for* a magnitude of interest (e.g., high likelihood of benefit), whereas hypothesis tests work by providing evidence *against* a magnitude (e.g., low compatibility with non-benefit) and are thereby less intuitive. Greenland (2019) promotes an example of evidence of the latter kind: assessment of the p value of the hypothesis test transformed into a "surprisal" (S) value, which is the number of consecutive head tosses of a fair coin that has the probability p. I do not recommend S values for assessing sampling uncertainty.

If an editorial board bans NHST from its journal, I foresee a problem. Researchers who get statistical significance for their main effect, and who do not want to bother with the more insightful but challenging methods of dealing with sampling uncertainty, will submit their manuscripts to a journal that welcomes statistically significant effects. Such effects are more likely with larger sample sizes (indeed, all effects become significant with large-enough sample sizes), hence a journal banning NHST might be at risk of becoming known for publishing small-scale studies. This problem would be solved if all journals banned NHST, but that won't happen anytime soon: no-one wants to admit they've been supporting bad methodology, overwhelming evidence does not always disabuse people of mistaken beliefs, and the issue of sampling uncertainty may be considered secondary to improving the impact factor and increasing the proportion of rejected manuscripts.

I therefore expect most journals will allow authors to continue to use the traditional p value, while only encouraging use of better methods. Unfortunately, exhorting authors to show but not to interpret the traditional p value will not stop many readers from misinterpreting $p<0.05$ as evidence of a real effect and $p>0.05$ as evidence of no effect. In fact the evidence is neither necessary nor sufficient: effects can be significant but not decisively substantial, decisively substantial but not significant, non-significant but not decisively trivial, and decisively trivial but significant (not non-significant). Rejection of substantial and non-substantial hypothesis, and equivalent decisions based on Bayesian probabilities, provide necessary and sufficient evidence, at least as far as sampling uncertainty is concerned. If these better methods are used, inclusion of NHST will serve only to muddy the waters.

Example of a statistical methods section with frequentist and Bayesian approaches

The following is taken from the methods section of a manuscript in preparation. Material in brackets […] is specific to the design and data of the study. Mention of MBD is avoided deliberately (for submission to journals currently banning MBD), but the descriptions of hypothesis testing and Bayesian assessment are consistent with MBD.

[Physical test scores are shown as means and standard deviations (SDs). Means of the dependent variables are shown as the back-transformed least-squares means with SDs in back-transformed ± percent units (when <30%) or ×/÷ factor units (when >30%) derived from a model without a physical-test predictor. Effects are shown in percent units with uncertainty expressed as ±90% compatibility limits (CL), when the effect and its ±90% CL are <30%; otherwise, factor effects with ×/÷90% CL are shown.]

For those who prefer a frequentist interpretation of sampling uncertainty, decisions about magnitudes accounting for the uncertainty were based on one-sided interval hypothesis tests, according to which a hypothesis of a given magnitude (substantial, non-substantial) is rejected if the 90% compatibility interval falls outside that magnitude (Hopkins, 2020). P values for the tests were therefore the areas of the sampling t distribution of the effect falling in the hypothesized magnitude, with the distribution centred on the observed effect. Hypotheses of inferiority (substantial negative) and superiority (substantial positive) were rejected if their respective p values ($p_–$ and p_+) were <0.05; rejection of both hypotheses represents a decisively trivial effect in equivalence testing. The hypothesis of non-inferiority (non-substantial-negative) or non-superiority (non-substantial-positive) was rejected if its p value ($p_{N–}=1–p_–$ or $p_{N+}=1–p_+$) was <0.05, representing a decisively substantial effect in minimal-effects testing.

A complementary Bayesian interpretation of sampling uncertainty is also provided, when at least one substantial hypotheses was rejected: the p value for the other hypothesis is the posterior probability of a substantial true magnitude of the effect in a reference-Bayesian analysis with a minimally informative prior (Hopkins, 2019), and it was interpreted qualitatively using the following scale: >0.25, possibly; >0.75, likely; >0.95, very likely; >0.995, most likely

(Hopkins et al., 2009). The probability of a trivial true magnitude ($1–p_-–p_+$) was also interpreted with the same scale. Possible or likely magnitudes are summarized as *some evidence* for those magnitudes; very likely and most likely are summarized as *good evidence*. Probabilities were not interpreted for unclear effects: those with inadequate precision at the 90% level, defined by failure to reject both substantial hypotheses ($p_->0.05$ and $p_+>0.05$). Effects on magnitudes and probabilities of a weakly informative normally distributed prior centered on the nil effect and excluding extremely large effects at the 90% level were also investigated (Greenland, 2006; Hopkins, 2019).

Effects with adequate precision at the 99% level ($p_-<0.005$ or $p_+<0.005$) are highlighted in bold in tables, since these represent stronger evidence against substantial hypotheses than the 90% level and therefore incur lower inflation of error with multiple hypothesis tests. For practitioners considering implementation of a treatment based on an effect in this study [(e.g., training to improve try scoring by increasing jump height)], the effect needs only a modest chance of benefit (at least possibly increased try scoring, $p_+>0.25$) but a low risk of harm (most unlikely impaired try scoring, $p_-<0.005$). Substantial effects highlighted in bold therefore represent potentially implementable effects. However, it is only for [effects on tries scored assessed via match winning] that the outcomes have direct relevance to benefit and harm (winning and losing matches); these effects were therefore also deemed potentially implementable when the chance of benefit outweighed an otherwise unacceptable risk of harm (the odds ratio of benefit to harm >66.3) (Hopkins & Batterham, 2016). For these effects, the potential for benefit and harm was also investigated for realistic changes in physical-test measures (less than 2 SD).

References

Aisbett J. (2020). Conclusions largely unrelated to findings of the systematic review: Comment on "Systematic review of the use of “magnitude-based inference” in sports science and medicine". PLoS ONE 15 https://journals.plos.org/plosone/article/comment?id=10.1371/annotation/330eb883-4de3-4261-b677-ec6f1efe2581

Albers CJ, Kiers HA, van Ravenzwaaij D. (2018). Credible confidence: a pragmatic view on the frequentist vs Bayesian debate. Collabra: Psychology 4, 31.

Amrhein V, Greenland S, McShane B. (2019). Retire statistical significance. Nature 567, 305-307.

Barker RJ, Schofield MR. (2008). Inference about magnitudes of effects. International Journal of Sports Physiology and Performance 3, 547-557.

Batterham AM, Hopkins WG. (2006). Making meaningful inferences about magnitudes. International Journal of Sports Physiology and Performance 1, 50-57.

Batterham AM, Hopkins WG. (2015). The case for magnitude-based inference. Medicine and Science in Sports and Exercise 47, 885.

Batterham AM, Hopkins WG. (2019). The problems with The Problem with 'Magnitude-based Inference'". Medicine and Science in Sports and Exercise 51, 599.

Benjamini Y, De Veaux R, Efron B, Evans S, Glickman M, Graubard BI, He X, Meng X-L, Reid N, et al. (2021). ASA president’s task force statement on statistical significance and replicability. Harvard Data Science Review 3 https://doi.org/10.1162/99608f92.f0ad0287

Burton PR. (1994). Helping doctors to draw appropriate inferences from the analysis of medical studies. Statistics in Medicine 13, 1699-1713.

Cleather DJ, Hopkins WG, Drinkwater EJ, Stastny P, Aisbett J. (2021). Improving collaboration between statisticians and sports scientists. British Journal of Sports Medicine 55, 118-122. https://bjsm.bmj.com/content/55/2/118.responses#improving-collaboration-between-statisticians-and-sports-scientists

Cumming G. (2014). The new statistics: why and how. Psychological Science 25, 7-29.

Greenland S. (2006). Bayesian perspectives for epidemiological research: I. Foundations and basic methods. International Journal of Epidemiology 35, 765-775.

Greenland S. (2017). For and against methodologies: some perspectives on recent causal and statistical inference debates. European Journal of Epidemiology 32, 3-20.

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 73, 106-114.

Hopkins WG. (2019). A spreadsheet for Bayesian posterior compatibility intervals and magnitude-based decisions. Sportscience 23, 5-7. https://www.sportsci.org/2019/bayes.htm

Hopkins WG. (2020). Magnitude-based decisions as hypothesis tests. Sportscience 24, 1-16. https://www.sportsci.org/2020/MBDtests.htm

Hopkins WG. (2021). Misleading conclusions based on statistical significance and non-significance at a recent international conference. Sportscience 25, 1-5. https://www.sportsci.org/2021/NHSTmisuse.htm

Hopkins WG, Batterham AM. (2008). An imaginary Bayesian monster [letter]. International Journal of Sports Physiology and Performance 3, 411-412.

Hopkins WG, Batterham AM. (2016). Error rates, decisive outcomes and publication bias with several inferential methods. Sports Medicine 46, 1563-1573.

Hopkins WG, Marshall SW, Batterham AM, Hanin J. (2009). Progressive statistics for studies in sports medicine and exercise science. Medicine and Science in Sports and Exercise 41, 3-12.

Lakens D, Scheel AM, Isager PM. (2018). Equivalence testing for psychological research: a tutorial. Advances in Methods and Practices in Psychological Science 1, 259-269.

Lohse K, Sainani K, Taylor JA, Butson ML, Knight E, Vickers A. (2020). Systematic review of the use of "Magnitude-Based Inference" in sports science and medicine. PLoS ONE, https://doi.org/10.1371/journal.pone.0235318.

Mastrandrea MD, Field CB, Stocker TF, Edenhofer O, Ebi KL, Frame DJ, Held H, Kriegler E, Mach KJ, et al. (2010). Guidance note for lead authors of the IPCC fifth assessment report on consistent treatment of uncertainties. Intergovernmental Panel on Climate Change (IPCC), https://pure.mpg.de/rest/items/item_2147184/component/file_2147185/content.

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, 1-13.

Rothman KJ. (2012). Epidemiology: an Introduction (2nd ed.). New York: OUP.

Rothman KJ, Greenland S. (2018). Planning study size based on precision rather than power. Epidemiology 29, 599-603.

Sainani KL. (2018). The problem with "magnitude-based inference". Medicine and Science in Sports and Exercise 50, 2166-2176.

Sainani KL, Borg DN, Caldwell AR, Butson ML, Tenan MS, Vickers AJ, Vigotsky AD, Warmenhoven J, Nguyen R, et al. (2021). Call to increase statistical collaboration in sports science, sport and exercise medicine and sports physiotherapy. British Journal of Sports Medicine 55, 118-122. https://bjsm.bmj.com/content/55/2/118

Sainani KL, Lohse KR, Jones PR, Vickers A. (2019). Magnitude-Based Inference is not Bayesian and is not a valid method of inference. Scandinavian Journal of Medicine and Science in Sports 29, 1428-1436.

Shakespeare TP, Gebski VJ, Veness MJ, Simes J. (2001). Improving interpretation of clinical studies by use of confidence levels, clinical significance curves, and risk-benefit contours. Lancet 357, 1349-1353.

van Ravenzwaaij D, Monden R, Tendeiro JN, Ioannidis JP. (2019). Bayes factors for superiority, non-inferiority, and equivalence designs. BMC Medical Research Methodology 19, 1-12.

Welsh AH, Knight EJ. (2015). "Magnitude-based Inference": A statistical review. Medicine and Science in Sports and Exercise 47, 874-884.

Acknowledgements: Thanks to the reviewer (Dan Cleather), Ken Quarrie and Sander Greenland for useful feedback and corrections.

Published September 2021