This is a PREPRINT and has not been peer reviewed

Magnitude Based Inference in Relation to One-sided Hypotheses Testing Procedures

Janet Aisbett

Meraglim Holdings Corporation

Daniël Lakens

Technical University of Eindhoven

Kristin Sainani

Department of Epidemiology and Population Health, Stanford University

Correspondence concerning this article should be addressed to Janet Aisbett, Meraglim Holdings, West Palm, Beach Florida. Email: janet.aisbett@gmail.com. DL was funded by a VIDI Grant 452-17-013 from the Netherlands Organisation for Scientific Research.

All authors have read and approved this version of the manuscript. This article was last modified on August 21, 2020. Please cite as: Aisbett, J., Lakens, D., & Sainani, K. (2020).

Magnitude Based Inference in Relation to One-sided Hypotheses Testing Procedures.

SportRxiv. doi: 10.31236/osf.io/pn9s3

Abstract

Magnitude based inference (MBI) was widely adopted by sport science researchers as an alternative to null hypothesis significance tests. It has been criticized for lacking a theoretical framework, mixing Bayesian and frequentist thinking, and encouraging researchers to run small studies with high Type 1 error rates. MBI terminology describes the position of confidence intervals in relation to smallest meaningful effect sizes. We show these positions correspond to combinations of one-sided tests of hypotheses about the presence or absence of meaningful effects, and formally describe MBI as a multiple decision procedure. MBI terminology operates as if tests are conducted at multiple alpha levels. We illustrate how error rates can be controlled by limiting each one-sided hypothesis test to a single alpha level. To provide transparent error control in a Neyman-Pearson framework and encourage the use of standard statistical software, we recommend replacing MBI with one-sided tests against smallest meaningful effects, or pairs of such tests as in equivalence testing. Researchers should pre-specify their hypotheses and alpha levels, perform *a priori* sample size calculations, and justify all assumptions. Our recommendations show researchers what tests to use and how to design and report their statistical analyses to accord with standard frequentist practice.

1. Introduction

Magnitude Based Inference (MBI¹; Batterham & Hopkins, 2006, 2015; Hopkins & Batterham, 2016) arose out of a recognition of the limitations of conventional null hypothesis

¹ In 2019 MBI was renamed Magnitude Based Decisions (MBD; see Hopkins, 2019a) to acknowledge that it is a decision-making rather than inference procedure. We retain the original name for consistency with the literature we cite.

significance testing (NHST). It was invented in the context of sports science, where sample sizes may be small (Bernards et al., 2017; Speed & Andersen, 2000), and where, unlike fields such as medicine, false positives may be preferred over false negatives provided the risk of adverse effects is low (Drinkwater, 2008). With conventional hypothesis testing, researchers often pay too much attention to p -values and ignore effect sizes, which leads to common errors such as the confusion of practical and statistical significance, and the misinterpretation of $p > .05$ as evidence of no effect. MBI attempts to redirect researchers' focus from p -values to statistical tests that compare observed effects against meaningful effect sizes.

Statistical tests that explicitly take into account the practical relevance of effect sizes have been suggested as an improvement on null hypothesis significance testing for more than half a century (e.g., Hodges & Lehmann, 1954; Nunnally, 1960; Berger & Hsu, 1996). Although equivalence and non-inferiority tests are common in pharmacokinetics and biostatistics, such approaches remain relatively rare in many fields. MBI was an independent attempt to develop a statistical procedure where researchers test against the smallest effects that are deemed important, and it was widely adopted by sports scientists through Excel spreadsheet implementations. This is an important accomplishment: in a time when most statistical inferences rely on NHST, and there is an intensive debate about how researchers can move beyond $p < .05$ (Wasserstein et al., 2019), MBI helped sports scientists recognize the existence of statistical tools other than null hypothesis significance testing.

However, for over a decade, statisticians have identified problems with MBI, including that it lacks a properly developed theoretical framework, mixes Bayesian and frequentist language, promotes studies with small sample sizes, and leads to high Type I error rates (Baker & Schofield, 2008; Welsh & Knight, 2015; Sainani 2018; Curran-Everett, 2018; Lohse et al., 2020). Poor software development practices have been noted in its Excel implementations (Sainani et al., 2019). Interestingly, until very recently none of the critiques

mentioned established interval testing approaches such as equivalence testing (e.g. Lakens, 2017) despite the obvious similarity in their identification of an interval of effect sizes of practical equivalence.

In this paper we describe confidence intervals in relation to one-sided tests of meaningful effects and the associated interval testing approach of equivalence testing (e.g., Berger & Hsu, 1996; Piaggio et al., 2006). This is used to formally show what MBI does. MBI needs to be stripped back to the underlying one-sided tests to place it in a Neyman-Pearson frequentist setting in which errors are controlled.² We give detailed recommendations on not only the tests required but the way that trial analyses should be designed and how sample sizes are calculated.

The remainder of the paper is organised as follows. In section 2, we describe one-sided tests and their combinations in equivalence and minimal effects testing. In section 3, we provide examples of analyses using one-sided tests and illustrate how error rates are controlled. In section 4, we describe the clinical and mechanistic forms of MBI developed by Hopkins and Batterham (Batterham & Hopkins, 2006; 2010, Hopkins & Batterham, 2006, Hopkins, 2019) in terms of such one-sided hypothesis tests. In section 5, we present recommendations for using one-sided tests in well-designed analyses that follow standard reporting practices. In section 6, we discuss these recommendations and the contributions of MBI, and look at how past reports using MBI analyses can be interpreted. Supplemental Appendix 1 briefly describes Excel implementations of MBI, Supplemental Appendix 2 details MBI formulas that need changing for our description to provide it with a theoretical foundation, and Supplemental Appendix 3 contains error rate and sample size calculations for (combinations of) one-sided test procedures.

2. Smallest meaningful effect sizes and one-sided test procedures

A two-sided null hypothesis significance test (NHST) at alpha level .05 corresponds to examining whether a 95% confidence interval around the point estimate of effect size excludes the test value of zero. On the other hand, a one-sided hypothesis test of effects greater than zero at this alpha level will be statistically significant if the lower bound of a 90% confidence interval is larger than zero. This correspondence between hypothesis tests and the position of confidence intervals can be easily extended to tests of any ranges of effect sizes. In this section we first look at the positioning of confidence intervals in relation to smallest meaningful effect magnitudes, which is the basis of MBI. Then we describe one-sided tests that seek to reject meaningful or non-meaningful effects, and procedures that combine these tests.

2.1 Confidence intervals in relation to smallest meaningful effect sizes

To interpret observed effects in relation to effects that are deemed meaningful, researchers need to specify which effects are *not* deemed meaningful. One commonly used approach is to specify an interval of effects equivalent to the absence of a meaningful effect (which we will call an equivalence range). Figure 1 illustrates how such an equivalence range subdivides effect sizes (depicted on the x -axis) into three subranges determined by the lower and upper bounds of the equivalence range. These bounds, indicated by the vertical dotted lines in the figure and labelled Δ_L and Δ_U , should be set by researchers based on which effect sizes are deemed theoretically or practically important. MBI calls these bounds ‘thresholds’ and refers to effects that fall within the equivalence range as ‘trivial’, to effects that are larger than the

equivalence range as ‘positive’, and to effects smaller than the equivalence range as ‘negative’. We will adopt this terminology here. Note that MBI assumes bounds are symmetrical around zero, but this is not necessary.

MBI uses words to describe six decisions about the effect size that are based on six positions of confidence intervals (CIs) in relation to the equivalence range, as illustrated in Figure 1. These six MBI decisions can be described using a combination of one-sided hypothesis tests that correspond to asking different research questions. The tests are variously referred to as tests of superiority, inferiority, non-superiority, non-inferiority, equivalence, and minimal effects, as we now describe.

Figure 1. The horizontal axis represents the effect size, with the upper and lower bounds of the equivalence range at Δ_L and Δ_U . The horizontal lines represent confidence intervals. Six unique decisions about the effect size can be made depending on the position of the CI in relation to the equivalence bounds. These positions are illustrated by the lines ‘A’ to ‘F’.

2.2 One-sided, equivalence and minimal effects testing

Equivalence tests, of which the best-known approach is TOST (Rogers, Howard, & Vessey, 1993), provide an important tool to interpret non-significant null hypothesis tests, and prevent the fallacy of interpreting $p > .05$ in NHST as the absence of an effect (Parkhurst, 2001). The TOST procedure is designed to test whether it is possible to reject the presence of meaningful effects. If the presence of meaningful effects can be rejected, one can act as if the effect is too small to be meaningful.

Figure 2. Comparison of the hypothesis of interest H_1 and the null hypothesis H_0 in procedures using one-sided hypothesis tests. In all these tests the goal is to reject effects in the shaded area (H_0) using one-sided tests of effects either larger or smaller than the equivalence range bounds Δ_L and Δ_U .

An equivalence test has as null hypothesis that the effects are either positive (larger than Δ_U) or negative (smaller than Δ_L). Such a composite hypothesis is illustrated in Figure 2a.

Rephrasing equivalence testing in terms of CIs, the null hypothesis is rejected only in case ‘B’ in Figure 1, when the CI lies completely in the equivalence range. Because both one-sided tests need to be statistically significant to conclude equivalence, a $100(1-2\alpha)\%$ confidence interval corresponds to one-sided testing at level α (e.g., a 90% CI for a test with $\alpha = 0.05$).

In some cases, researchers are only interested in equivalence in one direction; this is commonly referred to as a *non-inferiority test* and is illustrated by Figures 2e and 2f. A finding is considered non-negative (or non-inferior in the positive direction) if effects more negative than Δ_L can be rejected. In Figure 1, case ‘E’ would be considered non-negative, as the confidence interval shows we can reject effects smaller than Δ_L .

Minimal effect tests (Murphy & Myers, 1999), on the other hand, seek to determine whether an effect is large enough to be meaningful, and aim to reject effects in the equivalence range. Whereas a one-sided null hypothesis test aims to reject a difference of at most zero, a minimal effects test has the more ambitious goal of rejecting effects of at most Δ_U and at least Δ_L (Figure 2b). If we can reject such effects, we can act as if there is an effect large enough to be deemed practically or theoretically meaningful.

The two one-sided tests used in minimal effects testing test for effects at least as large as the lower bound (that is, effects that are non-negative), as illustrated in Figure 2c, and for effect at most as large as the upper bound Δ_U (that is, non-positive effects) as illustrated in Figure 2d. These are sometimes called superiority tests, although this term is often reserved for testing against an upper bound of zero.³ In terms of CIs, the null hypothesis for a

³ The different terms to describe tests might be confusing. It helps to keep in mind all these tests aim to reject the presence of effects in specific ranges—which are bounded by the

directional test predicting a positive effect is rejected if the $100(1-2\alpha)\%$ CI falls completely above the upper bound Δ_U . It would therefore be rejected by the first CI position in Figure 1, case ‘A’, and by the third position, case ‘C’.

Figure 3 visualises regions where the tests of positive, negative, non-positive and non-negative effects are rejected on the two-dimensional space generated by a sample mean difference (i.e., the size of the effect) on the x-axis, and the standard error on the y-axis. Since confidence intervals are defined by a sample mean and the standard error, each CI maps to a unique position on this space. Remember that the standard error (SE) decreases as the sample size increases, and that the standard error determines the width of the confidence interval. Therefore, the smaller the SE, the smaller the deviation from the test value that will be statistically significant in a hypothesis test.

In Figure 3, the area outside the V shape formed by the grey lines is the rejection region of a two-sided NHST at level .10—this presentation is similar to a funnel plot often used in meta-analyses. The rejection region for the one-sided test that aims to reject non-negative effects in a minimal effects test is shown in red, and that for the one-sided test that aims to reject non-positive effects in a minimal effect test is shown in green. Because these tests are performed against an effect larger than zero (specifically, $\Delta_L = -\Delta$ and $\Delta_U = \Delta$) the rejection region is entirely contained within the NHST rejection region (indicated by the grey V centred at zero). The equivalence test requires rejection of both the test that the effect is negative and the test that it is positive and so its rejection region is the central grey triangle.

As the figure shows, when the standard error is large enough—for example, when the sample size is very small—it is not possible to find statistical equivalence, because the CI

smallest effect magnitudes that are deemed meaningful—with a pre-specified error rate determined by the test alpha level.

MBI AND ONE-SIDED TESTS

will be so wide it never fits between the equivalence bounds. See case ‘F’ in Figure 1, which can only occur when the SE is larger than the peak of the grey triangle in Figure 3. When the standard error is large relative to the smallest meaningful effect magnitude, the observed effect size must also be large in order to fall in a rejection region. In other words, with small sample sizes, only relatively large effects can be statistically rejected.

Figure 3. Rejection regions for the tests shown in Figure 2 as a function of sample mean and standard error (after assuming equivalence range bounds $-\Delta$ and Δ then standardising for Δ). All one-sided tests are at an alpha level of .05. Rejection region for equivalence testing is the grey triangle. The rejection regions for tests of the hypotheses that effects are non-negative and non-positive (the minimal effects tests) are the red and green shaded areas. ‘Inconclusive’ is where neither the test for positive effects nor the test for negative effects can be rejected. The grey V-shaped line is the boundary of the NHST rejection region at level .10. Normal distributions approximate t -distributions. The confidence interval labelled ‘C’ in Figure 1 is represented by the position shown as a blue dot

3. Examples of how to use one-sided tests in sports science

In this section, we present two hypothetical examples to demonstrate how one-sided tests can be used to answer specific research questions. These examples also illustrate factors that need to be considered in the design stage. First, imagine that our research question of interest is whether beta-alanine supplements can improve swimming performance. We plan a randomized trial to compare a supplemented group to a placebo group. What is the hypothesis of interest? We are interested in testing the null hypothesis of non-benefit. Since there are costs associated with taking a useless supplement (time, money, etc.), we don't want to recommend the supplement unless we can reject the null hypothesis of non-benefit at a sufficiently strict alpha. Thus, we will use a one-sided minimal effects test aimed at rejecting non-positive effects.

We need to further refine the hypothesis by specifying a primary outcome, such as the baseline-to-4 week improvement in 200-meter freestyle times. If we plan to measure additional endpoints, these either need to be clearly designated as secondary endpoints or must be accounted for in sample size planning, as discussed in section 5. We next need to pick a smallest meaningful effect magnitude, and Type I and Type II error levels. Perhaps, after careful consideration⁴, we decide that a 1-second improvement would be a meaningful improvement for our population, and that Type I and Type II error rates of .05 and .30, respectively, are acceptable. We also believe from previous studies that the standard deviation in our population is 5 seconds and that a true effect size of 3 seconds is realistic. To calculate the one-sided test sample size we use a standard NHST sample size calculator for a

⁴ Both these choices require careful thought. But a discussion of how these choices are made is beyond the scope of this article. Researchers need to consider prior literature, the population of interest, and the costs and benefits of false negatives and false positives.

MBI AND ONE-SIDED TESTS

comparison of two means (details are in Appendix 3): Choose a one-sided test and enter the standard deviation and desired Type I and Type II error rates as usual, but instead of providing the anticipated effect size of 3 seconds, enter the absolute difference between the anticipated effect size and the smallest meaningful effect. This difference is 2 seconds. Plugging in these values yields 59 per group, as we show in Table 1.

To analyse the resulting data, we would use a minimal effects test of the hypothesis that effects are smaller than Δ_U , which can be implemented in standard software packages such as SAS or JASP, through Excel add-ins or using R-libraries. Suppose we run this test and obtain a p -value of .04. This would meet the pre-specified alpha of .05 and would yield a decision to act as if there is a positive effect (i.e., to assume beta-alanine supplements bring an average improvement of at least 1 second). In addition to the p -value, we should report the mean improvement in swimming times, as well as the standard deviation in swimming times, and the 95% confidence interval for the mean improvement. Graphical displays that show individual-level data should also be included.

Table 1. Worked sample size calculation for example 1, a randomized trial of beta-alanine versus placebo pill in swimmers. The primary dependent variable is the 4-week improvement in 200 M Freestyle times (in seconds). Formulas are for a two-group comparison of means with equal sample sizes.

Parameters	Smallest meaningful effect magnitude (Δ): ± 1 second Expected difference in means (ES, or effect size): 3 seconds Standard deviation (σ): 5 seconds Type I error rate (α): 5% ($Z_{1-\alpha} = Z_{0.95} = 1.645$) Type II error rate (β): 30% ($Z_{1-\beta} = Z_{0.70} = 0.524$)
How to use an NHST sample size calculator to perform the calculation	Choose a one-sided test. Use $ \Delta - ES $ for the effect size value, e.g., $3-1=2$.

NHST formula, one-sided test	$n \text{ per group} = 2\sigma^2 \left(\frac{Z_{1-\alpha} + Z_{1-\beta}}{\delta} \right)^2; \text{ set } \delta = \Delta - ES $
Worked calculation	$n \text{ per group} = 2(5^2) \left(\frac{1.645 + 0.524}{ 1 - 3 } \right)^2 = 59$

246

247 As a second example, suppose we want to gauge the effect of ice baths on post-exercise
 248 muscle inflammation. Ice baths are inexpensive, widely used, and reduce immediate pain in
 249 many athletes; however, some recent evidence suggests that they may paradoxically increase
 250 muscle inflammation. In this case, we may be interested in designing a trial focused on
 251 answering the question of whether ice baths are harmful or not. We might only recommend
 252 continued use of ice baths if we can reject harm with a sufficiently stringent alpha level. In
 253 this case, our null hypothesis of interest is that ice baths are harmful. Thus, we will use a one-
 254 sided test of harm (effect sizes of Δ_L or smaller) with an alpha level of .005.

255 As before, we must choose the primary outcome, smallest meaningful effect bounds, and
 256 Type II error rates *a priori*. Suppose we plan to compare runners who take an ice bath
 257 immediately after a strenuous workout to runners who only stretch; the primary outcome is
 258 Ikb α phosphorylation, a marker of muscle inflammation, measured in units of mean
 259 fluorescence intensity (MFI) at 24 hours post-workout. We choose a bound of 10 MFI. The
 260 standard deviation for MFI is 40 and we expect a 20-MFI difference (or half standard
 261 deviation difference) between groups.

262 What sample size do we need if we are willing to tolerate an overall Type II error of
 263 20%? One can again use a standard NHST sample size calculator for a comparison of two
 264 means: Choose a one-sided test and enter the standard deviation and desired Type I and Type
 265 II error rates as usual, but instead of providing the anticipated effect size of 20 seconds, enter
 266 the absolute difference between the anticipated effect size and the smallest meaningful effect.

This difference is $20 - (-10) = 30$ seconds. Table 2 shows that for a test of harm with an alpha of .005 and 20% Type II error, we will need a sample size of 42.

Again, the data can be analyzed in any standard statistical package. Suppose we obtain a p -value of .001. This meets the pre-specified alpha level of $<.005$, which means that we can reject harmful effects with a 0.5% error rate. We would report the results of the study as one would for a non-inferiority study, meaning that our conclusion would be that we can act as if ice baths are not harmful, but cannot make any claims about benefit.

Table 2. Worked sample size calculation for example 2, a randomized trial of ice baths versus stretching. The primary dependent variable is Ikbα phosphorylation, measured in units of mean fluorescence intensity (MFI) at 24 hours post-workout. Formulas are for a two-group comparison of means with equal sample sizes.

Parameters	<p>Smallest meaningful effect bound (Δ): ± 10 MFI</p> <p>Expected difference in means between the ice bath and control groups (ES, or effect size): 20 MFI</p> <p>Standard deviation (σ): 40 MFI</p> <p>Type II error rate (β): 20% ($Z_{1-\beta} = Z_{0.80} = 0.842$)</p> <p>alpha level for the harm test (h): 0.5% ($Z_{1-h} = Z_{0.995} = 2.576$)</p>
Sample size for a one-sided test of harm	$n = 2\sigma^2 \left(\frac{Z_{1-h} + Z_{1-\beta}}{ES - \Delta_L} \right)^2 = 2(40)^2 \left(\frac{2.576 + 0.842}{20 - (-10)} \right)^2 = 42$

4. Explaining MBI Through One-Sided Hypothesis Tests

We now describe MBI in terms of the four one-sided hypothesis tests against the smallest meaningful effect magnitudes—namely, the tests of positive, negative, non-positive, or non-negative effects shown in Figure 2. MBI builds on what we have described so far by effectively testing at different alpha levels and by assigning verbal labels to indicate

combinations of test results. Various spreadsheet implementations of MBI are available on the site sportsci.org and are briefly reviewed in Appendix 1.

4.1 Relating CIs, hypothesis tests, and verbal descriptors

Table 3 translates the CIs in Figure 1 into the results of four one-sided hypothesis tests. It also lists how these decisions would be described using MBI terminology versus typical terminology found in software documentation (e.g. Costelloe & Watts, 2015) and drug and clinical trial analyses (e.g., CMPM, 2001; Goeman, Solari & Stejles, 2010). These six combinations of hypotheses were described as decisions in Lehmann's (1957) development of multiple decision procedures.

Table 3. Figure 1 decisions in terms of placement of CIs with respect to the bounds of meaningful effect sizes and in terms of outcomes of hypothesis tests.

DECISION: CI PLACEMENT	DEFINITION IN TERMS OF HYPOTHESIS TESTS	TERMINOLOGY IN THE LITERATURE	MBI SUMMARY DECISION (Hopkins & Batterham, 2016)
A: Below Δ_L	Reject non-negative effects	Inferior (or superior: negative direction)	Negative
B: In between Δ_L and Δ_U	Reject positive and reject negative effects	Equivalent	Trivial
C: Above Δ_U	Reject non-positive effects	Superior (positive direction)	Positive
D: Contains only Δ_L	Reject positive effects but can't reject that effect is Δ_L	Not inferior but not superior (negative direction)	Negative-Trivial
E: Contains only Δ_U	Reject negative effects but can't reject that effect is Δ_U	Not inferior but not superior (positive direction)	Trivial-Positive
F: Contains Δ_L and Δ_U	Can't reject positive effects and can't reject negative effects	Inconclusive	Unclear

4.2. MBI described in terms of one-sided hypothesis tests

While MBI is based on Figure 1—and thus its decisions are based on the position of a confidence interval that can be reformulated as in Table 3 — MBI is more complex because it performs the one-sided tests at three different alpha levels.⁵ One level is stringent (default alpha of .005), one level is moderate (the standard alpha of .05) and the third is very lenient (alpha of .25). The results of all the tests at the three levels are combined to give verbal descriptions of the decisions, as shown in Table 4. Here, each MBI decision is associated with hypothesis tests at the alpha level shown at the top of the relevant column. Decisions are made at the smallest alpha for which the decision conditions hold. For example, if non-negative effects are rejected at levels .25 and .05, say, the verbal description for .05 is used and the MBI decision is ‘Very likely negative’. The table shows how to map the verbal labels used by MBI into test alpha levels and onto the comparative terminology of superior, inferior and non-inferior treatments with respect to a standard.

We note that there are differences between our description of MBI in terms of the outcomes of hypothesis tests and the implementation of the tests in the spreadsheets. Our reinterpretation of MBI brings the theory into line with the definitions in terms of CIs in Figure 1 and allows all the MBI decisions to be defined purely in terms of the outcomes of the four one-sided hypotheses tests determined at the three different alpha levels.⁶

⁵MBI implementations don’t repeat tests, but the effect of the calculations is as if they did.

⁶ Although MBI labels are theoretically based on the position of CIs at different alpha levels (Batterham & Hopkins, 2006), as explained by Table 4, spreadsheet implementations of MBI differ in their determination of the ‘Possibly’ and ‘Trivial’ decisions. Appendix 2 describes the MBI definitions, including its ‘Possibly trivial’ decision, in relation to our definitions.

MBI AND ONE-SIDED TESTS

Table 4. MBI decisions in relation to test outcomes at three alpha levels for the one-sided tests. A decision at a smaller alpha overrides a decision at a larger alpha. The values .25, .05 and .005 are defaults.*

DEFINITION IN TERMS OF HYPOTHESIS TESTS	MBI DECISION		
	.25	.05	.005
Reject non-negative effects	Likely negative	Very likely negative	Most likely negative
Reject positive and reject negative effects	Likely trivial	Very likely trivial	Most likely trivial
Reject non-positive effects	Likely positive	Very likely positive	Most likely positive
Negative-trivial: Reject positive effects at $\alpha=.05$ but can't reject that effect is Δ_L	Possibly negative		
Positive-trivial: Reject negative effects at $\alpha=.05$ but can't reject that effect is Δ_U	Possibly positive		
Inconclusive: Can't reject positive effects and can't reject negative effects		Unclear (more data are needed)	

*Some MBI spreadsheet implementations on sportsci.org only allow users to change the alpha level used in determining 'Unclear'. See Appendix 1.

Figure 4 illustrates the relation between the decisions of Table 4 and equivalence and minimum effects tests in terms of rejection regions that are functions of the standardised sample mean and the standard error. Figure 4 is effectively multiple versions of the depiction in Figure 3 of rejection regions for these tests at a single alpha level, with the more saturated shades in Figure 4 corresponding to rejection of tests at lower alpha levels. Mechanistic MBI can be reduced to the five types of tests listed in Table 3 provided the following changes are made: First, a single alpha level is chosen *a priori* (which, of course, may take values other than .005, .05, or .25), and second, 'Unclear' results are calculated at this level only, and correspond to Inconclusive results in an equivalence test.⁷ With these changes, Figure 4 reduces to Figure 3 and Table 4 to Table 3.

⁷ In mechanistic MBI, 'Unclear' effects can sometimes override otherwise positive or negative decisions, as pictured in Figure 4. For example, whereas a positive result is declared

In addition to returning decisions, MBI also returns probabilities, described as, for example, a “55% chance that the true value of the effect statistic is beneficial” which gives a Bayesian interpretation of frequentist statistics. See Appendix 1.

Comparing an observed p -value against different alpha levels is an example of the improper use of a hybrid approach to statistical tests (Gigerenzer, 2018). In Fisherian significance tests one reports a p -value and interprets smaller p -values as stronger evidence. In a Neyman-Pearson approach, which we follow in this manuscript, an alpha level is specified before the data are collected with the goal to control error rates when acting as if effects in certain ranges can be rejected, or not. This is discussed further in section 5.

When users draw conclusions, they are consciously or unconsciously selecting a single alpha level in a Neyman-Pearson approach. A concern with offering results from tests at multiple alpha levels is that users may make conclusions at a higher alpha level than they would normally consider, and MBI default settings offer unusually high levels. If ‘Likely positive’ effects are treated as positive, the test level is .25, much higher than researchers would typically consider desirable because it allows a Type I error rate of up to 25%. If ‘Possibly positive’ effects are treated as positive, the probability of an erroneous inference is as high as 75% because failing to reject a test of positive at 0.25 is equivalent to rejecting a test of non-benefit at 0.75. A concern with multiple alpha levels is that in practice users may default to the highest alpha level when drawing conclusions.

in Table 1 when negative and trivial effects are rejected, a ‘Likely positive’ result is declared in MBI (Table 4) only when non-positive effects are rejected AND negative effects are rejected at the default alpha of .05. In reducing MBI to tests at a single alpha level, the unclear/inconclusive decision (CI position ‘F’ in Figure 1) should be determined at the same level as the other decisions.

Figure 4. MBI decisions as regions which are functions of the sample mean and standard error, standardised by the meaningful effect magnitude bound Δ . The red regions labelled ‘Negative’ are where tests of non-negative effects are rejected, the green regions labelled ‘Positive’ are where tests of non-positive effects are rejected, and the monochrome regions labelled ‘Trivial’ are where tests of positive and negative effects are rejected. More saturated shades indicate tests at lower alpha levels. The area outside the blue dashed-line V shape is the rejection area for standard NHST at alpha .05.

4.3 Clinical MBI

MBI is presented as two versions: mechanistic (or non-clinical) MBI and clinical MBI. Table 4 and Figure 4 describe mechanistic MBI, which was the original formulation. Clinical MBI was introduced as a testing procedure to address situations where there is a strong asymmetry between negative and positive findings (Hopkins, 2007). These situations are typically trials in which the effect is a surrogate for a performance indicator such as strength

MBI AND ONE-SIDED TESTS

or speed, where the effect is defined so that positive effects are considered beneficial, and negative effects are harmful. Thus, the lower bound Δ_L is the largest value considered harmful and the upper bound Δ_U is the smallest value considered beneficial.

Clinical MBI involves two one-sided tests of harm and benefit (that is, of negative and positive effects) but these tests are by default conducted at different alpha levels, reflecting the asymmetry between harmful and beneficial effects. Although error rates in a two-sided test are typically distributed equally over the two directions, this symmetry is not required. For example, if the cost of incorrectly concluding an effect is positive is greater than that of incorrectly concluding an effect is negative, researchers may recommend asymmetrical error rates, where, say, an alpha level of .01 is used for the test of effects in the negative direction, and a level of .04 is used for the test of effects in the positive direction (see, e.g., Rice & Gaines, 1994).

Figure 5. Decision regions for clinical MBI when the SESOIs are symmetric, and with default alpha settings. Normal distributions approximate t -distributions. The rejection region for benefit (positive) is shaded dark grey. The white region is where neither benefit nor harm (negative) is rejected. These regions comprise the 'Don't use' decision. The remaining area

(light grey and green shades) is the ‘Consider using’ decision and includes the decisions ‘Possibly positive’ plus ‘Likely’ through to ‘Most likely’ positive. The areas to the right of the blue dashed lines are the rejection areas for traditional one-sided NHST at alpha levels of .025 (long dash) and .05.

Figure 5 illustrates how clinical MBI returns decisions ‘Consider Using’ and ‘Don’t use’ as a function of standardised sample mean and standard error. The dark grey region in the figure groups all the ‘negative’ (i.e., ‘harmful’) and ‘trivial’ decisions from Table 4 under the label ‘Don’t use’, which annotates these decisions and the ‘Unclear’ decision in some MBI spreadsheet implementations. Remaining decisions are annotated ‘Use’ in older versions of these spreadsheets, and ‘Consider using’ in recent amendments. Figure 5 shows that ‘Consider using’ is defined by rejection of the harm test and failure to reject benefit, and ‘Don’t use’ is defined by failure to reject harm or rejection of benefit. A decision is either ‘Consider using’ or ‘Don’t use’; there are no inconclusive results. Failing to reject benefit provides no basis to act as if the effect is beneficial rather than trivial, and hence ‘Consider using’ provides no basis either.

With small-to-moderate sample sizes (which are common in sports science) and using the default settings, clinical MBI’s final decision—to use or not use the intervention—is predominantly determined by the test for harm, not the test for benefit. Figure 5 illustrates this: Above the horizontal dotted line, the benefit test is almost never rejected when the harm test is rejected, meaning that the harm test dominates; below this line, the benefit test dominates. The figure also shows that at the default alpha level of .005 this test of harm will be inconclusive (‘Unclear’) with large SEs (small sample sizes) unless the effect size is large relative to the equivalence bound Δ .

An alternative to clinical MBI is simply a one-sided test of harm, that is, a non-inferiority test. To draw conclusions about benefit, a test of non-benefit could be applied after the test of harm.⁸ Since clinical MBI did not allow conclusions about benefit, a smallest meaningful benefit of zero may be justifiable. The test for non-benefit is then a one-sided NHST of effect sizes less than zero. Rejection regions for this test are illustrated in Figure 5.

5. Recommendations

In the previous sections, we explained MBI in relation to one-sided tests and we identified issues that need to be addressed. In this section, we give recommendations for statistical analyses that take account of meaningful effects and we provide guidance on how to calculate sample sizes to achieve target error rates.

5.1 Overall recommendations

In section 4, we described how MBI can be reduced to the one-sided tests that underlie it. If the hypothesis of interest is to determine whether effects are positive (beneficial) or negative (harmful), MBI can be replaced with minimal effects tests of non-positive or non-negative respectively. If the hypothesis of interest is to determine whether effects are trivial, MBI can be replaced with equivalence testing. If the hypothesis of interest is to determine whether effects are non-positive or non-negative, MBI can be replaced with tests of positive

⁸ If both tests are at the same alpha level, rejection of a test of non-benefit necessarily means that the test of harm is rejected. This is not always true when the harm test is at a more stringent level. Note that the authors disagree on the use of two tests—DL believes that researchers interested in concluding benefit should simply use a test of non-benefit.

or negative. We also considered the situation when, having determined that an effect is not harmful, the hypothesis of interest switches to whether the effect is greater than zero (tested by one-sided NHST), or greater than the smallest meaningful effect (minimal effects test). These tests can all be implemented using standard statistical software.

A simple summary of our recommendation is therefore this: replace MBI with the one-sided tests that underlie it. However, we also make recommendations on how to correctly plan and report analyses using these tests, as summarized in Table 5.

Table 5. Recommendations for testing against meaningful effects in a Neyman-Pearson framework.

OVERALL RECOMMENDATION	
Replace MBI with minimal effects tests, non-inferiority tests or equivalence tests depending on the hypothesis of interest. These tests can be conducted sequentially or simultaneously without adjustment of the alpha levels for multiplicity.	
SPECIFIC RECOMMENDATIONS FOR RESEARCH DESIGN AND REPORTING	
Registration	Pre-register your research plan, following directions at repositories such as osf.io and aspredicted.org
Choice of a specific hypothesis	Pre-register and justify a single hypothesis of interest, including a primary dependent variable.
Choice of smallest effect size of interest	Pre-register and justify the smallest effect of interest; use of defaults is not recommended.
Interpretation of p -values	Report p -values numerically; do not interpret them as Bayesian posterior probabilities within a frequentist analysis.
Interpretation of what a test says about effect sizes	When a test rejects effects in a range, don't say effects are <i>not</i> in that range. Thus if non-positive is rejected, say effects smaller than Δ_L have been rejected or that <i>we can act as if</i> effects are larger than Δ_L or that findings are <i>compatible with</i> effects larger than Δ_L .
Control of Type I errors	Choose and justify a single alpha level <i>a priori</i> for each test. Different tests may be at different alpha levels if this is justified.
Control of Type II errors and sample size	See Table 6 and Appendix 3

calculators	
Reporting	Report as minimal effects testing, non-inferiority testing or an equivalence test, including the specific hypotheses tested and exact p -values returned. If more than one of these tests is conducted, report results of all tests. ⁹ Report and justify the smallest effect size of interest, alpha levels, Type I error, and Type II error/power calculated for each test conducted.

444

445 5.2. Recommendations for design, conduct and reporting of statistical analyses

446 *Design stage*

447 Several design steps occur prior to data collection. First, the researcher must define their
448 hypothesis of interest, which includes specifying a statistical test (one of the tests discussed in
449 this manuscript), a primary dependent variable, and a smallest effect size of interest. Picking
450 a single primary dependent variable prevents inflation of Type I error due to multiplicity. If
451 multiple endpoints are of interest, then this must be factored into the choice of alpha level.
452 Guidance on adjusting alpha levels for multiple tests can be found in texts such as Bretz,
453 Hothorn & Westfall (2011). Note that it is not necessary to adjust the alpha level for
454 multiplicity when testing more than one hypothesis—e.g., both equivalence and minimal
455 effects—on the same endpoint. These hypotheses also do not need to be tested at the same
456 alpha level, if this can be justified.

457 The choice of smallest effect size of interest should be informed by the question at hand
458 and clearly justified (Lakens et al., 2018b; Cook et al., 2017; Hopkins, Hawley & Burke,
459 1999; Hopkins, 2010). We do not recommend the use of default choices for the equivalence
460 bounds, such as 0.2 standard deviations.

⁹ If for example using $\Delta_L = -0.2, \Delta_U = 0.5$, a minimum effects test (test for non-positive) returns $p=.21$ and a non-inferiority test (test for non-negative) returns $p=.01$ and the pre-specified alpha level is .05, then report that you could not reject an effect of 0.5 or smaller ($p = 0.21$) but could reject an effect of -0.2 or larger.

In a Neyman-Pearson framework, statistical tests and the corresponding error rates are chosen before data collection, and study design should carefully weigh the relative cost of Type I and Type II errors (Lakens et al., 2018a; Mudge et al., 2012). Researchers tend to blindly use defaults of a 5% Type I error rate and a 20% Type II error rate (or 80% power) despite long-standing cautions that the researcher should determine these as appropriate (e.g. Neyman & Pearson; 1933; Cohen, 1977). Researchers also must provide a sample size justification that incorporates these error rates. Such a sample size justification may involve practical considerations about limited sample sizes and feasibility constraints (Mudge et al., 2012; Allison et al., 1997). For example, one guidance regarding studies of the shelf-life of drugs recommends testing at the unusually high alpha level of .25 to compensate for expected limited sample sizes (ICH, 2003). In such cases, the researcher must still justify the study given the associated error rates and discuss the costs of each type of error. Lakens (2020a) provides some guidelines.

We recommend that researchers pre-register their study plan, including their alpha level and smallest effect size of interest before the data are analysed, by storing an online, date-stamped document in a repository such as aspredicted.org (Caldwell et al., 2020). This removes the temptation to move to higher alpha levels or lower bounds on meaningful effect sizes after looking at the data, which inflates Type I error rates (Lakens, 2020b).

Appendix 3 reviews standard procedures used to estimate Type I and II error rates, and the test levels and sample sizes required to achieve these for one-sided tests and for equivalence and two-sided minimal effects testing. These formulas are summarized in Table 6. Appendix 3 also gives formulas for the special cases of one-sided tests of harmful, non-beneficial and non-zero effect sizes that may be used in place of clinical MBI. Sample size calculations follow a form familiar to NHST calculations, and standard NHST calculators can be used to perform these calculations with simple modifications to the inputs. We gave worked

MBI AND ONE-SIDED TESTS

examples in section 3. Note that the summary formulas in Tables 6 and the calculations in Appendix 3 are only for a specific type of statistical comparison—the comparison of two means with equal sample sizes. These could, however, easily be modified for other comparisons or for unbalanced group sizes.

Table 6. Error rates and sample size estimates to control error rates. Normal distributions approximate t -distributions. Z_α is the 100 α % point and Φ is the cumulative density function of the standard normal distribution. Equivalence tests and minimal effects tests have symmetric smallest effect sizes of interest $\pm\Delta$. Formulas are for a two-group comparison of means; n is the number per group, σ^2 is the estimated group variance and ES is the anticipated effect size.

	Type I error rate given sample size n and test alpha level α	Type II error rate given sample size n and one-sided test alpha level α	Sample size needed to get Type I error rate α and Type II error rates β^{**}
One-sided test	α	$1 - \Phi\left(\frac{ \Delta - ES }{\sqrt{2\sigma^2 / n}} - Z_{1-\alpha}\right)$	$2\sigma^2 \left(\frac{Z_{1-\alpha} + Z_{1-\beta}}{\Delta - ES}\right)^2$
Equivalence test	α	Double the rate for one-sided test *	$2\sigma^2 \left(\frac{Z_{1-\alpha} + Z_{1-\beta/2}}{\Delta}\right)^2$
Minimal-effects test	2α *	Same as one-sided test	Same as one-sided test

* These are upper bounds on the true error rates. See Appendix 3.

**These sample size formulas are similar to those for NHST, and standard NHST calculators can be used to perform these calculations. For minimal effects tests, use an NHST sample size calculator but replace the effect size with the absolute difference between the effect size and the smallest effect size of interest. For equivalence testing, use an NHST sample size calculator but halve the target Type 2 error rate before inputting this value.

Note that there are MBI calculators that perform sample size calculations with an Excel spreadsheet (see sportsci.org). The sample size recommended in MBI documentation is the size required for a non-inferiority test to detect an effect on the opposite bound with power of 95% at an alpha level of 5% (mechanistic MBI) or power 75% at alpha level 0.5% (clinical MBI). This is a special case. The MBI sample size calculator is therefore not broadly applicable and we do not recommend its use.

Analysis and reporting stage

After the study is complete, the data may be analysed using standard statistical packages. TOST implementations for equivalence testing are readily available, for example, in proprietary packages such as SAS, in purchasable Excel add-ins, or in freeware such as JASP, or in R libraries or spreadsheets (Lakens, 2017). More generally, the p -values for one-sided tests as we have described in this manuscript can be obtained from statistical software packages through a standard one-sided test in the appropriate direction against the appropriate meaningful effect bound. Regardless of the choice of software, data analysis should always include checking, plotting, and understanding the data prior to performing formal hypothesis testing.

We recommend against use of verbal label such as ‘Very likely positive’ because, formally, hypothesis tests only allow statements about values that are rejected. Words such as ‘Likely’ have Bayesian interpretations that should be avoided in a frequentist analysis, and may obscure the test alpha level from the researcher and reader. We recommend users stick to reporting numerical statistics such as exact p -values (Wasserstein et al., 2019), effect sizes, and confidence intervals. Thus, instead of reporting ‘Very likely positive’ researchers should report non-positive values were rejected at $p = .12$, say, if that was the p -value.

6. Discussion

We have described MBI as a decision procedure for determining the size of an effect in relation to an equivalence range. The only formal description of the algorithms that MBI employ was through Excel spreadsheet formulae, and these differ between implementations. Although the developers have described MBI as “essentially Bayesian with a flat prior” (Batterham & Hopkins, 2006; p56) where “the usual frequentist confidence interval is interpreted in a Bayesian fashion” (Hopkins, 2019a), we set out to explain MBI in frequentist terms. To this end, we described MBI in relation to one-sided hypothesis tests that seek to reject effects that are either too small to matter (a minimum effect test), or test the opposite hypothesis that aims to reject the presence of effects large enough to matter (a non-inferiority test). These tests can be run using standard statistical software. We also gave procedures for controlling Type I and Type II error rates, and explained what information should be pre-specified to make the tests transparent.

When introducing a new statistical technique in the literature, its developers should provide an explicit formulation of the method in mathematical notation, explain how the method relates to existing statistical methods, and demonstrate the method’s empirical behavior through mathematical proofs and/or simulations. Although this accepted process will not completely prevent problems with how statistical methods are used, the failure of the MBI developers to follow it has led to additional problems.

Lohse et al. (2020) documented published analyses using MBI which were more closely related to data exploration than hypothesis testing, with large numbers of dependent variables being tested for effects at unspecified alpha levels. There is nothing wrong with exploratory studies, but researchers should be clear that this is what they are doing and follow conventions for reporting such studies (e.g., McIntosh, 2017). If researchers plan to draw

inferences from their findings while studying a large number of dependent variables, they need to either pre-specify a primary outcome or adjust the test alpha for multiple comparisons. In a Neyman-Pearson framework, each test requires clear *a priori* specification of the hypothesis of interest. The researcher must then clearly state how the hypothesis will be tested and the specific criteria that will determine whether the hypothesis is supported by the findings (Lakens et al., 2018a; Lakens & DeBruine, 2020). Without this, it is impossible to control error rates for incorrect decisions (Piaggio et al., 2006).

When testing is limited to a single alpha level, mechanistic MBI reduces mathematically to simultaneous minimum effects, non-inferiority and equivalence tests, as described in Table 1. We note that the theoretical basis for simultaneous testing against multiple bounds has long been established (Lehmann, 1957), and the practice has been adopted by some regulatory agencies (Committee for Proprietary Medicinal Products, 2001) and used in clinical trials (Goeman et al., 2010). Such testing of hypotheses concerning the same dependent variable does not require adjustment of alpha levels for multiple comparisons when the hypotheses are disjoint, i.e., when only one of them can be true at any one time, or when one implies the other. Type II error control is possible using the largest of the sample sizes required by the individual tests.

Researchers need to include sample size justifications as an essential component of their research design, whatever the analysis method. In the past, MBI proponents and critics have differed over claimed error rates because MBI defined errors in a non-standard way. Formalizing MBI as one-sided hypothesis tests in a Neyman-Pearson framework allows for standard definitions of Type I and Type II errors and makes it straightforward to design studies and perform tests with expected error rates.

The authors of this paper are divided over the value of using any labels to report the outcomes of statistical tests. Words are used to describe ranges of numbers in everyday life

and in almost any scientific field, where they aid communication to non-specialists (e.g. USGS Geologic Names Committee, 2010). This was the motivation for MBI’s verbal labels (Batterham & Hopkins, 2006). Labelling ranges of p -values is an entrenched custom even amongst statisticians (Neyman, 1977; Gelman, 2018); however, neither the terminology nor the ranges it describes have become standardised. Labels like MBI’s that suggest Bayesian interpretations are not applied to frequentist statistics. Also, their use may encourage an overemphasis on a narrow set of alpha levels, such as .05 or .25, and may obscure alpha levels from users. On balance, we recommend that researchers simply report numerical p -values, along with the hypotheses tested and the pre-specified alpha level.

The approaches discussed in this paper apply beyond simple independent t -tests, such as repeated-measures data or regression analyses. In principle, whenever a confidence interval around the effect of interest can be established, statistically comparing the observed results against an equivalence range is possible (see Wellek, 2000).

An important question that remains undiscussed so far is how to interpret previous research reports that have been analysed using MBI. MBI documentation advises that researchers report findings as 90% confidence intervals, and we note that the practice of reporting confidence intervals rather than p -values has wide support. In addition, Lohse et al. (2020) found almost half the studies using MBI analyses reported at least some NHST analyses, and where these are missing the analysis can be back-calculated from the confidence intervals. Each report needs, however, to be re-evaluated to see if its conclusions hold up. Researchers should be cautious of published claims of an effect that have been made on the basis of ‘Likely positive’ or ‘Possibly positive’ results. Some papers that have reported positive effects may in fact only support a conclusion of “not negative” (Lohse et al. 2020). Readers should also keep in mind that when a large number of dependent variables have been tested, the multiplicity of tests greatly inflates the Type I error rate. If a paper provides

insufficient information for reanalysis, the report authors would need to be asked to supply those items, as would be the case whatever the analysis method.

We have explained MBI as a multiple decision procedure, which supports the recent change of name to MBD (Hopkins, 2019a). We also identified how a formally correct approach from a frequentist perspective requires stripping MBI back to the underlying one-sided hypothesis tests to allow error control and sample size planning. However, it should be considered an achievement to get researchers to use any interval-based testing method, and move beyond NHST. MBI can also be credited with innovations to standard TOST approaches, including user-friendly presentation of results of simultaneous tests of disjoint hypotheses and the “lopsided” tests used in its clinical version. Such lopsided tests are justified when it is more important to detect effects of one polarity than of the other, which seems to be fitting in at least some sports science research. The recommendations in this manuscript aim to retain MBI’s focus on the magnitude of effects while providing transparent error control and ensuring the design and reporting of analyses accord with standard frequentist procedures.

REFERENCES

- Allison DB, Allison RL, Faith MS, Paultre F, Pi-Sunyer FX. Power and money: Designing statistically powerful studies while minimizing financial costs. *Psych Methods*. 1997; 2(1): 20–33. <https://doi.org/10.1037/1082-989X.2.1.20>
- Baker RJ, Schofield, MR. Inference about magnitudes of effects. *Int. J. Sports Physiol Perform*. 2008; 3: 547-557.

- 630 Batterham AM, Hopkins WG. Making meaningful inferences about magnitudes. *Int J*
 631 *Sports Physiol Perform.* 2006; 1:50–57. <https://doi.org/10.1123/ijsp.1.1.50>.
- 632 Batterham AM, Hopkins WG. The case for magnitude-based inference. *Med Sci Sports*
 633 *Exerc.* 2015; 47:885.
- 634 Berger RL, Hsu JC. Bioequivalence trials, intersection-union tests and equivalence
 635 confidence sets. *Stat. Sci.* 1996; 11 (4): 283--319. doi:10.1214/ss/1032280304.
- 636 Bernards JR, Sato K, Haff GG, Bazyler CD. Current research and statistical practices in
 637 sport science and a need for change. *Sports.* 2017; 5(4):87.
- 638 Bretz F, Hothorn T, Westfall PH. Multiple comparisons using R. 2011. New York:
 639 Chapman and Hall/CRC, <https://doi.org/10.1201/9781420010909>
- 640 Caldwell AR, Vigotsky AD, Tenan MS, Radel R, Mellor DT, Kreutzer A, Lahart IM,,
 641 Mills JP, Boisgontier MP. Moving sport and exercise science forward: A call for the adoption
 642 of more transparent research practices. *Sports Med.* 2020; 50(3); 449–459.
- 643 CPMP (Committee for Proprietary Medicinal Products). Points to consider on switching
 644 between superiority and non-inferiority. *British J. Clinical Pharmac.* 2001; 52:223-228.
- 645 Cohen J. Statistical power analysis for the behavioural sciences: Revised edition.
 646 Academic Press: NY 1977.
- 647 Cook, J.A., Julious, S.A., Sones, W. et al. Choosing the target difference (‘effect size’) for
 648 a randomised controlled trial - DELTA2 guidance protocol. *Trials.* 2017; 18: 271.
 649 <https://doi.org/10.1186/s13063-017-1969-5>.
- 650 Castelloe J, Watts D. Equivalence and noninferiority testing using SAS/STAT software.
 651 *Proc SAS Global Forum 2015 Conf*, Cary, NC: SAS Institute Inc.
- 652 Curran-Everett D. Magnitude-based Inference: Good idea but flawed approach. *Med Sci*
 653 *Sports Exerc.* 2018; 50:2164-2165.

- 654 Drinkwater E. Applications of confidence limits and effect sizes in sport research. Open
655 Sports Sci. 2008; 1:3–4.
- 656 Gigerenzer G. Statistical rituals: the replication delusion and how we got there. Adv
657 Methods & Practices Psych Sci, 2018;1(2): 198–218.
658 <https://doi.org/10.1177/2515245918771329>.
- 659 Goeman JJ, Solari A, Stijnen T. Three-sided hypothesis testing. Stat in Med. 2010;
660 29:2117-2125.
- 661 Hodges JL, Lehmann EL. Testing the approximate validity of statistical hypotheses. J.
662 Roy Statist Soc: Land Ser B. 1954; 16:261-268.
- 663 Hopkins WG. Estimating sample size for Magnitude-Based Inference. Sportscience. 2006;
664 10:63-70. <https://www.sportsci.org/resource/stats/xSampleSize.xlsx>. Accessed April 16,
665 2020.
- 666 Hopkins WG. A spreadsheet for deriving a CI, mechanistic inference and clinical
667 inference from a P value. Sportscience 2007: 11, 16-20. Updated Jan 2020.
668 <http://www.sportsci.org/2007/wghinf.htm>. Accessed April 16, 2020.
- 669 Hopkins WG. Important effect sizes for exercise and sport practitioners and scientists.
670 <https://www.sportsci.org/2010/ImportantEffectSizes.pptx>. Accessed from Linear models and
671 effect magnitudes for research, clinical and practical applications. Sportscience 2010; 14:49-
672 58. <https://www.sportsci.org/2010/wghlinmod.htm>. Accessed 15 March, 2020.
- 673 Hopkins WG. Magnitude-based decisions. Sportscience 23, i-iii. 2019a.
674 <http://sportsci.org/2019/inbrief.htm>. Accessed April 16, 2020.
- 675 Hopkins WG. A spreadsheet for Bayesian posterior compatibility intervals and
676 Magnitude-Based Decisions. Sportscience. 2019b: 23; 5-7.
677 <http://sportsci.org/2019/bayes.htm>. Accessed April 16, 2020.

Hopkins WG, Batterham AM. Error rates, decisive outcomes and publication bias with several inferential methods. *Sports Med.* 2016; 46:1563–1573.

<https://doi.org/10.1007/s40279-016-0517-x>.

Hopkins WG, Hawley JA, Burke LM. Design and analysis of research on sport performance enhancement. *Med & Sci Sports & Ex.* 1999; 31(3): 472-485

Hopkins WG, Marshall SW, Batterham AM, Hanin J. Progressive statistics for studies in sports medicine and exercise science. *Med. Sci. Sports & Exercise* 2009; 41:3-12.

ICH. Harmonised Tripartite Guideline: Evaluation for stability data Q1E. International conference on harmonisation of technical requirements for registration of pharmaceuticals for human use 2003. <https://database.ich.org/sites/default/files/Q1E%20Guideline.pdf>. Accessed June 30, 2020.

Julious SA, Campbell MJ. Tutorial in biostatistics: sample sizes for parallel group clinical trials with binary data. *Stat. Med.* 2012; 31 (24) 2904-36. ISSN 0277-6715.

Kruschke J. Rejecting or accepting parameter values in Bayesian estimation. *Adv Meth Pract Psych Sci.* 2018; 1(2); 270-280.

Lakens D. Equivalence Tests: A practical primer for t tests, correlations, and meta-analyses. *Soc Psychol Personal Sci.* 2017; 8(4):355-362. doi:

10.1177/1948550617697177. Lakens, D. (2020a). Feasibility sample size justification. <http://daniellakens.blogspot.com/2020/08/feasibility-sample-size-justification.html>

Lakens, D. (2020b). The value of preregistration for psychological science: A conceptual analysis. *Japanese Psychological Review*. <https://doi.org/10.31234/osf.io/jbh4w>

Lakens D, Scheel AM, Isager PM. Equivalence testing for psychological research: A tutorial. *Adv Meth Prac Psych Sci.* 2018a;1(2): 259–69

<https://doi.org/10.17605/OSF.IO/V3ZKT>.

- 702 Lakens D, Adolphi FG, Albers CJ et al. Justify your alpha. *Nat Hum Behav* 2018b; 2:168-
703 171.
- 704 Lakens, D, DeBruine LM. Improving transparency, falsifiability, and rigour by making
705 hypothesis tests machine readable. *PsyArXiv* 2020, <https://doi.org/10.31234/osf.io/5xcda>.
- 706 Lehmann EL. A theory of some multiple decision problems. Part II. *Ann Math Statist*
707 1957:28; 547-572.
- 708 Lohse KR, Sainani KL, Taylor JA, Butson ML, Knight EJ, Vickers AJ.. Systematic review
709 of the use of “magnitude-based inference” in sports science and medicine. *PLoS ONE* 2020;
710 15(6): e0235318. <https://doi.org/10.1371/journal.pone.0235318>
- 711 McIntosh RD. Exploratory reports: A new article type for *Cortex*. *Cortex*. 2017; 1
712 (96):A1-4.
- 713 Mudge JF, Baker LF, Edge CB, & Houlahan JE. Setting an optimal α that minimizes
714 errors in null hypothesis significance tests. *PLOS ONE* 2012; 7(2), e32734.
715 <https://doi.org/10.1371/journal.pone.0032734>.
- 716 Murphy KR, Myers B. Testing the hypothesis that treatments have negligible effects:
717 Minimum-effect tests in the general linear model. *J App Psych* 1999; 84(2): 234.
- 718 Neyman J. Frequentist probability and frequentist statistics. *Synthese* 1977: 36; 97–131.
719 <https://doi.org/10.1007/BF00485695>.
- 720 Neyman J, Pearson ES. On the problem of the most efficient tests of statistical hypotheses.
721 *Phil Trans Roy Soc A* 1933; 231: 289-337.
- 722 Nunnally J. The place of statistics in psychology. *Ed Psych Meas*. 1960; 20(4): 641-650.
723 <https://doi.org/10.1177/001316446002000401>.
- 724 Piaggio G1, Elbourne DR, Altman DG, Pocock SJ, Evans SJ, Altman DG; CONSORT
725 Group. Reporting of noninferiority and equivalence randomized trials: an extension of the
726 CONSORT statement. *JAMA*. 2012; 308(24):2594-2604.

MBI AND ONE-SIDED TESTS

Parkhurst DF. Statistical significance tests: equivalence and reverse tests should reduce misinterpretation. *Bioscience*. 2001; 51(12):1051–1057.

Rice WR, Gaines SD. ‘Heads I win, tails you lose’: Testing directional alternative hypotheses in ecological and evolutionary research. *Trends Ecol. & Evol.* 1994; 9:235-237.

Rogers JL, Howard KI, Vessey JT. Using significance tests to evaluate equivalence between two experimental groups. *Psychol Bull.* 1993; 113(3):553-65.

Sainani KL. The problem with “Magnitude-Based Inference”. *Med Sci Sports Exerc.* 2018; 50:2166-2176. <https://doi.org/10.1249/MSS.0000000000001645>.

Sainani KL, Lohse KR, Jones PR, Vickers A. Magnitude-Based Inference is not Bayesian and is not a valid method of inference. *Scand J Med Sci Sports* 2019; 29(9): 1428-1436.

Shieh G. Exact power and sample size calculations for the two one-sided tests of equivalence. *PLoS ONE*. 2016; 11(9): e0162093.

Speed HD, Andersen MB. What exercise and sport scientists don't understand. *Sci Med Sport*. 2000; 3 (1); 84-92.

USGS Geologic Names Committee, Divisions of geologic time—major chronostratigraphic and geochronologic units: U.S. Geological Survey Fact Sheet 2010. <https://pubs.usgs.gov/fs/2010/3059/pdf/FS10-3059.pdf>

Wasserstein RL, Schirm AL, & Lazar NA. Moving to a world beyond ‘ $p < 0.05$.’ *The American Statistician*. 2019; 73(sup1), 1–19. <https://doi.org/10.1080/00031305.2019.1583913>.

Wellek S. On a reasonable disaggregate criterion of population bioequivalence admitting of resampling-free testing procedures. *Stat Med*. 2000; 19: 2755–2767.

Welsh AH, Knight EJ. ‘Magnitude-based inference’: a statistical review. *Med Sci Sports Exerc.* 2015; 47:874.