Myths and Methodologies: the use of equivalence and non-inferiority tests for interventional studies in exercise physiology and sport science

Supplementary materials:
https://osf.io/ndqhe/
For correspondence:
rmazzolari001@ikasle.ehu.eus

Raffaele Mazzolari[1,2], Simone Porcelli[2,3], David J Bishop[4], and Daniël Lakens[5]

[1]Department of Physical Education and Sport, University of the Basque Country (UPV/EHU), [2]Department of Molecular Medicine, University of Pavia, Pavia, Italy, [3]Institute for Biomedical Technologies, National Research Council, Segrate, Italy, [4]Institute for Health and Sport (iHeS), Victoria University, Melbourne, Australia, [5]Human Technology Interaction Group, Eindhoven University of Technology (TU/e), Eindhoven, the Netherlands

This is the pre-peer reviewed version of the following article: 'Mazzolari, R., Porcelli, S. Bishop, D. J., & Lakens, D. (2022). Myths and methodologies: The use of equivalence and non-inferiority tests for interventional studies in exercise physiology and sport science. *Experimental Physiology,* 10.1113/EP090171. Advance online publication. https://doi.org/10.1113/EP090171', which has been published in final form at https://physoc.onlinelibrary.wiley.com/doi/epdf/10.1113/EP090171. This article may be used for non-commercial purposes in accordance with Wiley Terms and Conditions for Use of Self-Archived Versions.

All authors have read and approved this version of the manuscript. This article was last modified on February 23, 2022.

Authors Raffaele Mazzolari *@RaffaeleM28*, Simone Porcelli *@PorcelliSimone*, David J Bishop *@BlueSpotScience*, and Daniël Lakens *@lakens* can be reached on Twitter.

ABSTRACT

Exercise physiology and sport science have traditionally made use of the null hypothesis of no difference to make decisions about experimental interventions. This article aims to review current statistical approaches typically used by exercise physiologists and sport scientists for the design and analysis of experimental interventions and to highlight the importance of including equivalence and non-inferiority studies, which address different research questions than deciding whether two interventions work differently. Firstly, we briefly describe the most common approaches, along with their rationale, to investigate the effects of different interventions. We then discuss the main steps involved in the design and analysis of equivalence and non-inferiority studies, commonly performed in other research fields, with worked examples from exercise physiology and sport science scenarios. Finally, we provide recommendations to exercise physiologists and sport scientists who would like to apply the different approaches in future research. We hope this work will promote the correct use of equivalence and non-inferiority designs in exercise physiology and sport science whenever the research context, conditions, applications, researchers' interests, or reasonable beliefs, justify these approaches.

INTRODUCTION

An often-overlooked aspect when designing and analysing interventional studies in exercise physiology and sport science concerns the type and direction of the research hypothesis(es) (Caldwell & Cheuvront, 2019). Most studies use the null hypothesis of no effect when making decisions about experimental interventions. That is, researchers usually examine whether there is a statistical difference between the experimental and the control group on one or more primary outcomes. However, other hypothesis tests may be more appropriate when researchers are interested in whether two interventions are similar in efficacy but substantially differ with respect to factors such as cost56 effectiveness, invasiveness, or administrative procedures (Hecksteden et al., 2018). The correct approach to designing and analysing interventional studies in exercise physiology and sport science continues to be extensively discussed in the literature (Caldwell & Cheuvront, 2019; Hecksteden et al., 2018; Hopkins et al., 1999; Mansournia & Altman, 2018). Recently, several researchers have recommended complementing the traditional null hypothesis with tests of equivalence and non-inferiority, which evaluate whether two interventions or conditions are similar or do not differ by more than a given amount (Aisbett et al., 2020; Caldwell & Cheuvront, 2019, Dixon et

al., 2018). In this article, we will review and expand the statistical toolset that can be used by sport and exercise scientists when designing and analysing interventional studies. We will refer to the best practices as developed in biomedical, social, and behavioural research since we recognise sufficient similarities with exercise physiology and sport science regarding the design of interventional studies. To increase understanding by exercise physiologists and sport scientists, we will also provide two worked examples from exercise physiology and sport science research that highlight how typical research designs and analyses conducted using traditional null hypothesis tests could be re-imagined using equivalence or non-inferiority tests. Moreover, we will provide theoretical and practical recommendations to exercise physiologists and sport scientists who would like to apply the different hypothesis tests in future research.

INVESTIGATING STATISTICAL DIFFERENCES (SUPERIORITY)

Unless otherwise specified, most interventional studies in exercise physiology and sport science have the implicit aim of determining if the efficacy of a given intervention is superior, or possibly inferior, to a control or reference intervention. In the most common study design, researchers randomise participants to either an experimental or a control group. The observed difference in group means after the intervention period (i.e., the effect size) is used to perform a hypothesis test examining a difference in population means. Following traditional null hypothesis testing, a difference between interventions can be concluded, while controlling the Type I error rate, whenever the *p*-value calculated from a particular test statistic indicates the observed or more extreme data are surprising (i.e., the *p*-value is less than or equal to the significance level, or α), assuming there is no difference between the interventions and all other modelling assumptions are met. Alternatively, researchers can choose a confidence interval (CI) approach. These two approaches lead to identical decisions in a hypothesis test, as *p* is less than or equal to .05 when a 95% CI excludes the value that is tested against (i.e., zero) (Figure 1a – upper example).

Regardless of the inferential approach employed, investigating differences between interventions without taking into consideration any meaningful value does not permit informed decisions regarding the *practical significance* of the outcome(s). From an exercise physiology and sport science perspective, testing the superiority of the experimental intervention against an effect size that is exactly zero may increase the risk of endorsing interventions, such as exercise training protocols or nutritional strategies, that are expensive, demanding, or time-consuming, but have no practical benefit – that is, they do not provide a noticeable advantage

over an existing benchmark. A more informative criterion for assessing superiority consists of determining whether the mean difference, after having considered its uncertainty, is larger than the smallest effect size of interest (SESOI), which should be defined *a priori* and justified on sound grounds (Lakens, 2021) (Figure 1a – middle example). This approach leads to the same conclusions as testing the shifted (non-zero) null hypothesis (Victor, 1987) or a '*minimum*-effect test', whose null hypothesis assumes that the mean difference between the interventions falls within a range of practically irrelevant values (Murphy et al., 2014).

Although the definition of SESOI is self-explanatory, exercise physiologists and sport scientists should be aware that several different methods exist to determine this value, depending on data and applications (Cook et al., 2018; Lakens, 2021). The 'anchor-based' method, which uses the researcher's judgment or participant's experience to define the SESOI, provides a common approach to interpret study outcomes in clinical research. The expert panel approach, also known as the Delphi method, is an alternative (although not necessarily straightforward) way to define the SESOI based on expert consensus. Previous studies may give an indication of the expected effect sizes. However, researchers should be aware that due to publication bias published effect sizes often overestimate the true effect of interventions, and that the distribution of effect sizes observed in literature does not necessarily inform about the SESOI, whose determination needs careful consideration and justification. Cohen's classical benchmarks (Cohen, 1988), developed for the social and behavioural sciences, are not recommended as guidance on identifying the SESOI in exercise physiology and sport science since an effect size of interest is context-dependent and should be decided based on a substantive research question (Caldwell & Vigotsky, 2020). Although some authors (Hopkins et al., 1999; Rhea, 2004) have developed scales for assessing the magnitude of effect sizes in some specific areas of exercise physiology and sport science, researchers should be aware that determining the SESOI is not a straightforward process, and it may be challenging in many sporting and physiological contexts.

Interpreting inconclusive evidence for superiority, or interpreting failure to reject the null hypothesis, as evidence for the equality of two interventions, is a common misconception (Altman & Bland, 1995). A statistically non-significant result (e.g., $p > .05$) cannot be interpreted as the absence of an effect. To be able to conclude an effect is absent, one needs to specify the alternative hypothesis explicitly, and perform a test that statistically rejects the alternative hypothesis. The traditional null hypothesis testing only rejects the null hypothesis, and, especially in small studies, a statistically non-significant result is not informative about whether the alternative hypothesis can be rejected. Exercise physiologists and sport scientists must

keep in mind that no correct conclusions other than superiority or inferiority can be drawn using traditional hypothesis tests. Because a well-designed study is informative about both the presence and absence of an effect of interest, researchers should consider complementing traditional null hypothesis tests with equivalence and non-inferiority tests.

INVESTIGATING EQUIVALENCE AND NON-INFERIORITY

Proving that two interventions or conditions are perfectly equal in efficacy is impossible from a statistical standpoint. What is possible in a statistical test is to reject the presence of a difference that is large enough to be practically relevant, defined by the upper (Δ_U) and lower (Δ_L) equivalence margins (Hodges & Lehmann, 1954; Lakens, 2017). Although various approaches exist to perform an equivalence test (Meyners, 2012), equivalence is typically investigated via the 'two one-sided tests' (TOST) procedure, which is a simple variation of a traditional hypothesis test (Schuirmann, 1987). In this procedure, the null and alternative hypotheses within each set are reversed and data are tested against Δ_U and Δ_L in two one-sided tests, each carried out at the α level (conventionally set to .05). Equivalence can be concluded at the α level only if both tests statistically reject the presence of effects equal to or larger than the equivalence margins. It is common to report only the greater *p*-value of the two one-sided tests when testing for equivalence since this *p*-value is also the one for the overall equivalence test (Berger & Hsu, 1996). The TOST procedure is operationally identical to concluding equivalence whenever the two-sided 100(1 − 2α)% CI for the mean difference between the interventions lies entirely within the equivalence margins (Schuirmann, 1987; Westlake, 1981) (Figure 1b – middle example).

Equivalence studies are very common in clinical research, in which new drug formulations or generic versions of the product are often compared to brand-name pharmaceuticals to prove bioequivalence (Senn, 2007). Moreover, this design has attracted growing interest in the social and behavioural sciences for its utility in evaluating replication results and corroborating risky predictions (Lakens, 2017; Lakens et al., 2018a). The latter application of equivalence hypotheses may also make them valuable for exercise physiology and sport science, which suffers from a shortage of replication experiments (Halperin et al., 2018). Nevertheless, until recently, investigating equivalence did not appear to be a common practice among exercise physiologists and sport scientists, who have so far restricted the use of equivalence tests mostly to measurement agreement research as an alternative or complementary approach to the Bland–Altman method (Dixon et al., 2018).

If there is an interest, along with a solid rationale, to investigate whether a given intervention is not unacceptably worse than a standard one with no restriction for its maximal efficacy, researchers can opt for a non-inferiority study. This is usually the case when the new intervention has better cost-effectiveness, is safer, is easier to implement, or is less demanding than the standard intervention. Non-inferiority studies can also be useful to evaluate modifications to well-established interventions and extend applicability to special populations. These research questions may also apply to exercise physiology and sport science. In non-inferiority testing, the non-zero null hypothesis is shifted towards the negative side of the *nil* (zero) effect, favouring the standard. It follows that, when applying the CI approach, non-inferiority is conventionally concluded when the lower margin of the two-sided 95% CI for the mean difference between the interventions lies above the non-inferiority margin (Δ_{NI}) (Senn, 2007) (Figure 1c – middle and lower examples).

**** Figure 1 near here ****

Compared with classical parallel-group studies, the design and analysis of non-inferiority studies face several additional methodological challenges, which include the suitability of the reference intervention, the determination of the Δ_{NI}, and sample size estimation. We will briefly review and discuss the main aspects of each of these challenges in the following sections. Since some of these issues also apply to equivalence studies, we will expand those parts where relevant.

Suitability of the reference intervention

From a clinical perspective, the non-inferiority of an experimental intervention can be firmly concluded only when compared to a reference intervention of well-established efficacy (Committee for Medicinal Products for Human Use, 2005; Committee for Proprietary Medicinal Products, 2000; International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use, 1998, 2001). The design characteristics of the reference intervention (population selection, intervention protocol, primary outcome measures, etc.) should be replicated as closely as possible to reduce the risk of violating the 'constancy assumption', which requires consistency between the effect of the reference group in the new study and the historical effect estimated from the literature. Violating this assumption may increase the chances of incorrectly concluding non-inferiority for inefficacious or even harmful interventions. When considering the extreme paucity of replication

experiments (Halperin et al., 2018), along with the small sample sizes characterising exercise physiology and sport science research (Speed & Andersen, 2000), it becomes self-evident that satisfying the prerequisite for the choice of the comparator arm represents the first critical issue to be addressed by exercise physiologists and sport scientists interested in conducting non-inferiority studies. Even when a discrete amount of evidence is available, the large sampling variability related to studies with small sample sizes (e.g., 8−16 participants per group) makes it difficult to identify an intervention whose efficacy had been consistently demonstrated across the literature. Moreover, questionable practices such as publication bias and *p*-hacking (i.e., the manipulation of data collection and analysis to obtain statistically significant results) tend to overestimate the intervention effect in meta-analyses and thus impact the 'assay sensitivity' of the new investigation, which is the ability of a study to distinguish between an efficacious and less efficacious intervention. Several graphical and statistical approaches seeking to quantify or adjust for publication bias in meta-analyses have been developed (Carter et al., 2019; Simonsohn et al., 2014). However, most of these methods lack large-scale empirical validation, do not work well when there are few studies or large heterogeneity in effect sizes, and their performance and efficiency are often highly sensitive to deviations from the model assumptions. Note that the problem of publication bias and *p*-hacking would be dramatically reduced if pre-registration or Registered Reports Protocols became common practice in exercise physiology and sport science (Caldwell et al., 2020; Lakens & Evers, 2014). These aspects highlight the importance of gaining reliable knowledge about effect sizes reported in the literature before deciding whether to adopt a non-inferiority design. This also emphasises the need for more collaborations across exercise physiology and sport science departments to design and conduct studies with high accuracy, and the need for more transparent research practices, as stressed by several scientists in a recent call (Caldwell et al., 2020).

Determination of non-inferiority and equivalence margin(s)

Once the reference intervention has been chosen, the next step in designing non-inferiority studies concerns the choice for the margin. An appropriate Δ_{NI} should be based on a combination of statistical reasoning and domain expertise (Committee for Medicinal Products for Human Use, 2005; Committee for Proprietary Medicinal Products, 2000; International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use, 1998, 2001). The general principle states that the Δ_{NI} should not be larger than

the smallest effect the reference intervention would be reliably expected to have compared with a placebo. Despite more sophisticated approaches being proposed (Snapinn & Jiang, 2008a; Yu et al., 2019), the 'point-estimate method' and the 'fixed-margin method' are the most widely used for specifying the margin in clinical research (Althunian et al., 2017). In the point-estimate method, the Δ_{NI} is based upon the pooled effect estimate of the active comparator from a meta-analysis without considering the uncertainty in the estimate (Δ_{NI-P}). In the fixed margin method, the two-sided 95% CI of the meta-analytic effect size estimate that is closest to the null effect is used to determine the non-inferiority Δ (Δ_{NI-C}) (Figure 2). This makes the latter approach more conservative than the former, especially when − as is often the case in exercise physiology and sport science − the precision of the individual study estimates is generally low, and the total number of studies is small. A third common approach to analyse non-inferiority trials applies the same criteria as the fixed-margin method to determine Δ_{NI} but also adjusts the CI derived from the non-inferiority trial to account for the sampling variability in the effect of the active comparator against placebo (Althunian et al., 2017; Holmgren, 1999). This 'synthesis method' is slightly more efficient than the fixed-margin method but it is also more sensitive to a violation in the assumptions of assay sensitivity and constancy (Schumi & Wittes, 2011).

Regardless of the method used to determine the Δ_{NI}, several factors such as the importance of the outcome measure, clinical or practical considerations in terms of cost-effectiveness of the active comparator, model misspecification, or violation of the constancy assumption can make putative superiority over placebo alone an insufficient criterion to establish non-inferiority and additional assurance may be needed. In this respect, pre-specifying a percentage of the historical effect of the reference intervention that must be retained by the new one (usually 50%), the so-called 'preserved fraction' (λ), has become common practice in non-inferiority clinical trials (Figure 2) (Snapinn, 2004; Snapinn & Jiang, 2008b). Despite its widespread use in clinical research, it is important to note that there is no consensus as to whether setting the Δ_{NI} by including a preserved fraction represents an effective discounting approach (Snapinn, 2004; Snapinn & Jiang, 2008b).

**** Figure 2 near here ****

Whether or not the stringency in the criteria to determine non-inferiority should be further adjusted according to the degree of magnitude of the historical effect of the comparator is a matter of debate among clinical researchers (Schumi & Wittes, 2011). Although

the choice of the preserved fraction would have negligible implications on the study conclusions for small to moderate effects, considerable discrepancies may take place for largely efficacious standard interventions. In these cases, determining the fraction without any adjustment for the historical effect of the comparator may rule out a large part of the effect, eventually leading to the paradoxical situation in which non-inferiority is established although the experimental intervention is inferior compared with the standard (Althunian et al., 2018; Schumi & Wittes, 2011). A maximum margin criterion that prevents clinically important differences between the standard and the new intervention may be applied in these situations (Schumi & Wittes, 2011).

Whereas (bio)equivalence margins in clinical trials are often set by regulatory authorities (Committee for Medicinal Products for Human Use, 2010), several approaches to justify the equivalence range have been proposed in the social and behavioural sciences (Lakens, 2017, 2021; Lakens et al., 2018a). Among them, it is worth mentioning a method based on the maximum sample size researchers are willing to collect given the available resources. This approach may be used for those situations, also common in exercise physiology and sport science, in which there are time, money, or population size constraints that limit the effect size that can be properly investigated, especially in novel lines of research. Under such conditions, determining Δ_U and Δ_L based on feasibility may be justified, and may represent a starting point for future studies aiming for a more precise assessment, if researchers see no way to specify the SESOI based on theoretical predictions or practical concerns.

Sample size planning for non-inferiority and equivalence studies

In superiority studies, sample size estimation conventionally aims to achieve the desired level of statistical power (typically 80% or 90%) against an alternative hypothesis, expressed in terms of a target difference between interventions in the primary outcome(s), at a given α (Cook et al., 2018). Since superiority and non-inferiority are logically opposite tests, sample size estimation for non-inferiority studies follows the same principles as for superiority studies. However, because the Δ_{NI} is usually smaller than the superiority difference, a larger sample size is often needed. Due to the nature of the TOST procedure, in which each one-sided test must statistically reject effects as small as the equivalence margins to prove efficacy, the power of an equivalence test equals the power to detect the smallest margin. In the light of the above, researchers should be aware that the adequate sample size for equivalence and non-inferiority tests may be prohibitively large for very small effects. For this reason, researchers should carefully consider the target or expected effect size, along with the margin(s), when

planning equivalence and non-inferiority studies. Whenever there is substantial uncertainty about the mean difference between the interventions, or when it is plausible that the true effect is larger or smaller than the margin the test was powered to detect, researchers may opt for sequential analysis (Lakens et al., 2021). This efficient approach allows terminating data collection while controlling the Type I error rate as soon as there is convincing evidence to decide on the presence, or absence, of an effect.

Julious (2004) provided detailed overviews for superiority, equivalence, and non-inferiority designs. Moreover, there are several spreadsheets (Lakens, 2017), statistical packages (Castelloe & Watts, 2015; Lakens, 2017), and web-based applications (Magnusson, 2016) that exercise physiologists and sport scientists can use to estimate sample sizes for equivalence and non-inferiority tests.

RE-IMAGING INTERVENTIONAL STUDIES USING EQUIVALENCE AND NON-INFERIORITY TESTS

We provide two worked examples from exercise physiology and sport science research comparing sprint interval training (SIT) against moderate-intensity continuous training (MICT) to show how the statistical approaches discussed above can be applied to real-world data. We have included all the formulas used in these examples in an accompanying spreadsheet (openly available − along with the SAS and R code used for validation − at https://osf.io/ndqhe/), which can also be used to perform calculations based on summary statistics or complete datasets.

Example 1 - use of equivalence hypothesis: In a comprehensive study investigating the effects of four weeks of SIT (60 min per week) or MICT (300 min per week) on cardiorespiratory, musculoskeletal, and metabolic characteristics in obese men, Cocks et al. (2016) concluded that SIT and MICT have equal benefits on aerobic capacity, as no statistical difference was observed between the two groups with respect to the changes in maximal oxygen uptake (VO_{2max}). As previously stated, the absence of an effect cannot be concluded based on $p > .05$ from the traditional null-hypothesis test. However, we wanted to determine whether the authors' conclusions concerning the absence of an effect between the groups can indeed be inferred from the observed data. Unfortunately, the authors did not report the nominal *p*-value for the time · group interaction in the 2 · 2 mixed analysis of variance (ANOVA) model, or any other necessary information about the differences in the changes in VO_{2max} between the groups. Since the authors did not make the raw data available along with the manuscript, we

cannot perform a proper covariate-adjusted analysis; nonetheless, we can still appraise the between-group differences by extracting summary data from the paper. Specifically, we can estimate the standard deviation (SD) of the change score within each group by imputing different plausible correlation coefficients (*r*) between pre- and post-training scores, construct the two-sided 90% CI for the mean difference between the groups using the different SD estimates, and then perform a sensitivity analysis on the results (Higgins et al., 2019). For $r = .5$, the SIT – MICT 90% CI around the observed mean difference of −2.3 mL · kg^{-1} · min^{-1} ranges from −7.1 to 2.5 mL · kg^{-1} · min^{-1}. The SDs of the change scores decrease at greater values of *r*, and the 90% CI narrows by ~ 17% (ranging from −6.3 to 1.7 mL · kg^{-1} · min^{-1}) when $r = .7$. However, even in the optimistic scenario in which $r = .9$, the 90% CI for the between-group difference ranges from −5.2 to 0.6 mL · kg^{-1} · min^{-1}, which indicates a large imprecision of the parameter estimate. Since a difference in VO_{2max} as small as 3.5 mL · kg^{-1} · min^{-1} has been associated with a 10–25% risk reduction in mortality (Ross et al., 2016), the mean difference between SIT and MICT that was observed by Cocks and colleagues of −2.3 mL · kg^{-1}· min^{-1} is hardly trivial after having considered its uncertainty.

If we wish, we can also formally test for equivalence against symmetric margins Δ_U and Δ_L of 3.5 mL · kg^{-1} · min^{-1} by using the TOST procedure, which is very similar to the Student's *t*-test when assuming equal population variances. This equivalence test examines the question of whether we can reject the presence of an effect as large, or larger than 3.5 mL · kg^{-1}· min^{-1}, which we know is large enough to have practical benefits.

For Δ_U

$$t_U = \frac{M_1 - M_2 - \Delta_U}{SD_P \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

where t_U is the test statistic for the one-sided *t*-test on Δ_U, M_1, and M_2 are the means of the SIT and MICT group respectively, n_1 and n_2 are the sample size in each group, and SD_P is the pooled SD:

$$SD_P = \sqrt{\frac{(n_1 - 1)\, SD_1^2 + (n_2 - 1)\, SD_2^2}{n_1 + n_2 - 2}}$$

where SD_1 and SD_2 are the SD of the SIT and MICT group, respectively.

In this example,

$$SD_P = \sqrt{\frac{(8-1)\ 2.1^2 + (8-1)\ 4.2^2}{8+\ 8\ -2}} = 3.3$$

Therefore

$$t_U = \frac{2.4 - 4.7 - 3.5}{3.3\sqrt{\frac{1}{8}+\frac{1}{8}}} = -3.5$$

which correspond to a *p*-value lower than 0.1 from the *t*-distribution with 14 degrees of freedom (*df*) for a left-sided test.

For Δ_L

$$t_L = \frac{M_1 - M_2 - \Delta_L}{SD_P\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$$

being t_L the test statistic for the one-sided *t*-test on Δ_L.

In this example,

$$t_L = \frac{2.4 - 4.7 - (-3.5)}{3.3\sqrt{\frac{1}{8}+\frac{1}{8}}} = 0.7$$

which corresponds to a *p*-value of .24 from the *t*-distribution with 14 *df* for a right-sided test.

Since the one-sided test with the greater *p*-value is not statistically significant [t(14) = 0.7, p = .24] based on an α = .05, we cannot reject differences larger than 3.5 mL · kg^{-1} · min^{-1}. Therefore, we cannot conclude that the difference between the two interventions is too small to matter (given a SESOI of 3.5 mL · kg^{-1} · min^{-1}) with respect to the changes in VO_{2max}.

It is important to note that, unlike in traditional hypothesis tests where effects that are substantially greater than expected can compensate small sample sizes, underpowered tests inevitably increase the risk of inconclusive results in equivalence studies. If we want to estimate how many individuals Cocks and colleagues should have recruited and tested to reach an adequate level of power (e.g., 80%) for the TOST procedure at the desired α level (e.g., .05), the most informative approach is to perform an *a priori* power analysis. For the sake of simplicity in calculations, we can define equivalence margins that are symmetric around a zero difference in population means ($\mu_1 - \mu_2$). Moreover, we assume that the estimated pooled SD represents the true SD for the two populations (σ). We can then rely on the normal approximation of the

power equation for equivalence tests and estimate the sample size (*n*) required in each group to achieve the desired power against Δ_U and Δ_L as (Julious, 2004):

$$n_U = \frac{(r+1)\,\sigma^2\,(z_\alpha + z_{\beta/2})^2}{r\,|\Delta_U|^2}$$

and

$$n_L = \frac{(r+1)\,\sigma^2\,(z_\alpha + z_{\beta/2})^2}{r\,|\Delta_L|^2}$$

where *r* is the allocation ratio (n_1 / n_2), and z_α and $z_{\beta/2}$ are the standardized normal deviates corresponding to the levels of α and $\beta / 2$ respectively (with 1 – β that represents the desired power). With an equal allocation (1:1 ratio), the equations 5 and 6 are reduced to:

$$n_U = n_L = \frac{2\,\sigma^2\,(z_\alpha + z_{\beta/2})^2}{|\Delta_U = \Delta_L|^2}$$

In this example,

$$n = \frac{2 \times 3.3^2\,(1.6 + 1.3)^2}{3.5^2} = 16$$

which indicates that the *minimum* sample size that Cocks and colleagues should have recruited to have a properly powered test for equivalence was double the *n* = 8 per group that was collected in that study. Note that this also represents an optimistic estimation: any situation in which some inequality between interventions can be expected (i.e., the expected difference is not 0), would increase the required sample size, all else being equal.

Example 2 - use of non-inferiority hypothesis: Gillen et al. (2016) investigated whether 30 min per week of SIT was a time-efficient exercise strategy to improve indices of cardiometabolic health in healthy men to the same extent as 150 min per week of MICT. Although the time · group interaction in the 3 · 2 mixed ANOVA model was significant for VO_{2max}, the authors were unable to reject a *nil* effect and conclude statistical differences between the groups after 12 weeks of training intervention. The exact *p*-value and the 95% CI for the between-group comparison were not reported; however, since the authors reported the 95% CI for the change scores of the two groups, as well as their sample sizes, we can obtain the information we need using statistical first principles (Higgins et al., 2019). The calculations reveal a *p*-value of .94 and a 95% CI ranging from −2.9 to 2.7 mL · kg^{-1} · min^{-1} constructed around a mean difference between the interventions of −0.1 mL · kg^{-1} · min^{-1}.

From a superiority standpoint, the study is inconclusive for what concerns the ability of SIT to improve the VO_{2max} compared with MICT. Given the rationale supporting the study, a more informative research question might be whether the improvements in the VO_{2max} induced by SIT are not substantially lower than those induced by a standard MICT program. To answer such a question, first, we must define the Δ_{NI} that we will use to test our hypothesis. The net effect of MICT against no-exercise control on VO_{2max} has been estimated to be 4.9 mL · kg^{-1} · min^{-1} with a 95% CI ranging from 3.5 to 6.3 mL · kg^{-1} · min^{-1} (Milanović et al., 2015). If we assume the MICT protocol prescribed by Gillen and colleagues is sufficiently representative of the 'typical' MICT from which the average intervention effect has been estimated and we prefer a conservative approach to the margin determination without further need for a preserved fraction, we can rely on the fixed-margin method and test the SIT – MICT difference against a Δ_{NI} of −3.5 mL · kg^{-1} · min^{-1}. The calculation of the t-statistic for the non-inferiority test is identical to those for the one-sided test against the ΔL in the TOST procedure.

$$t_{NI} = \frac{M_1 - M_2 - \Delta_{NI}}{SD_P \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

being t_{NI} the test statistic for the non-inferiority test.

In this example,

$$t_{NI} = \frac{5.9 - 6 - (-3.5)}{2.9 \sqrt{\frac{1}{9} + \frac{1}{10}}} = 2.6$$

which corresponds to a *p*-value of .02 from the *t*-distribution with 17 *df* for a two-sided test. If all the assumptions underlying the statistical model are correct, the non-inferiority test is significant [$t(17) = 2.6$, $p = .02$] for an $\alpha = .05$. We can then reject a loss in the efficacy of SIT compared with MICT larger than 3.5 mL · kg^{-1} · min^{-1} and conclude that SIT is non-inferior to MICT for what concerns increase in VO_{2max}. Unsurprisingly, given the close relationship between *p*-values and CIs, the CI approach leads to the same conclusion as the formal non-inferiority test since the lower 95% confidence limit of the SIT – MICT difference (i.e., −2.9 mL · kg^{-1} · min^{-1}) is larger than the Δ_{NI} (i.e., −3.5 mL · kg^{-1} · min^{-1}), which indicates that the entire set of plausible values for the population parameter contained in the 95% CI is consistent with the non-inferiority of SIT against MICT.

SWITCHING BETWEEN HYPOTHESES

Switching the objective of a clinical trial from non-inferiority to superiority or *vice versa* may be possible at the analysis stage of the study; however, the change is not always straightforward, and several points need to be considered (Committee for Proprietary Medicinal Products, 2000; Schumi & Wittes, 2011). From a statistical perspective, testing first for non-inferiority and then for superiority, does not require a statistical penalty for multiple testing, since the closed testing procedure properly controls the overall Type I error rate of the two tests. When the Δ_{NI} has been prespecified, and the trial design and conduct have been strict, it is also possible to test for non-inferiority after a superiority test that does not show any statistical benefit. Despite being statistically appropriate, researchers should be warned that this testing order could result in paradoxical outcomes (i.e., a new intervention that is both non-inferior and inferior to the standard), especially for largely efficacious standard interventions. As stated previously, considering the SESOI as a criterion for the largest acceptable Δ_{NI} may help to minimise this risk.

Departing from the initial aim of establishing equivalence does not appear to be a common practice in clinical research (Senn, 2007). Moreover, the greater value of α usually adopted in such investigations would lead to an inflated Type I error rate if the researcher attempted to draw straightforward conclusions on superiority or non-inferiority. Nonetheless, various comprehensive methods to investigate equivalence along with superiority have been recently presented in the social and behavioural sciences literature (Lakens, 2017; Lakens et al., 2018a) (Figure 3). Exercise physiologists and exercise scientists interested in conducting equivalence and non-inferiority studies may benefit from exploring these approaches.

It is also worth mentioning the possibility to test against both the nil effect and the SESOI in all those situations in which the researcher, after having concluded that the effect is non-zero, is interested in rejecting effects too small to be relevant.

**** Figure 3 near here ****

LIMITATIONS AND ADDITIONAL CONSIDERATIONS

In the present review, we have detailed how to expand the statistical toolset used to design and analyse interventional studies in exercise physiology and sport science. To achieve clarity and brevity, we focused on parallel-group studies with means and variances determined from pairs of independent random samples of normally distributed observations. Readers must be aware that the analytical approach to other research designs or variables with different probability distributions may slightly differ from the one presented herein.

When discussing the acceptable standard of evidence, we maintained consistency with the defaults commonly used in biomedical, social, and behavioural research. Nonetheless, the optimal error rates should be decided based on a cost-benefit analysis, depending on the context, goals, and resources (Lakens et al., 2018b).

It is worth keeping in mind that frequentist estimation (i.e., CI) and hypothesis testing do not represent the only way to draw inferences from data. Among the alternative or complementary methods, Bayesian statistics or Likelihood approaches can also be used to answer the questions that might be of interest to researchers (Lakens et al., 2020; van Ravenzwaaij et al., 2019; Wang & Blume, 2011). These approaches have the main advantage of allowing researchers to make probabilistic statements about the (random) parameter of interest. Whenever prior data are available from other studies, Bayesian statistics also allows incorporating such information in the analysis to update the (posterior) probability of the parameter and provide the relative weight of evidence for the alternative hypothesis compared with the null. Although presenting such methods to design and analyse superiority, equivalence, and non-inferiority studies were beyond the scope of this manuscript, exercise physiologists and sport scientists should consider their use within the context of statistical inference when deciding which method(s) is the most appropriate for their research purpose(s).

CONCLUSIONS AND RECOMMENDATIONS FOR FUTURE RESEARCH

Exercise physiology and sport science have largely relied on the traditional null hypothesis test to make informed decisions in interventional studies. This approach, combined with underpowered tests, has often led to the misinterpretation of a non-significant test result as support for the equivalence between interventions. While it should be clear at this point that this is a statistical misconception, exercise physiologists and exercise scientists should also understand that research should not be limited to investigating whether one intervention is superior or inferior to another. Equivalence and non-inferiority designs may be adopted whenever the research context, conditions, applications, researchers' interests, or reasonable beliefs justify them. Although these research hypotheses require additional methodological considerations than superiority hypotheses to be properly investigated, they may also better answer the empirical question researchers are interested in. Equivalence and non-inferiority studies may help exercise physiologists and exercise scientists to answer questions that the traditional null hypothesis cannot address. Figure 4 provides a flowchart to facilitate the decision-making process about the most informative study design.

**** Figure 4 near here ****

Figures

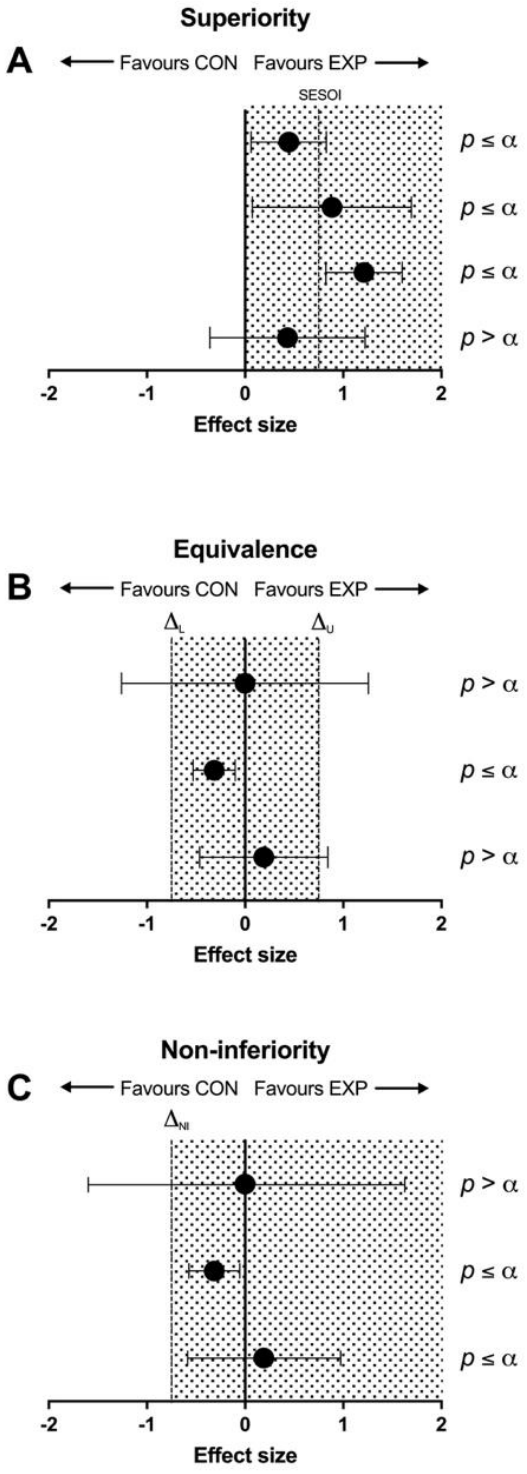
Superiority
A
Favours CON
Favours EXP
SESOI
p ≤ α
p ≤ α
p ≤ α
p > α
-2
-1
0
1
2
Effect size
Equivalence
B
Favours CON
Favours EXP
Δ_L
Δ_U
p > α
p ≤ α
p > α
-2
-1
0
1
2
Effect size
Non-inferiority
C
Favours CON
Favours EXP
Δ_NI
p > α
p ≤ α
p ≤ α
-2
-1
0
1
2
Effect size

Figure 1 Testing for superiority, equivalence, and non-inferiority within a typical parallel-group design. The error bars indicate the 95% confidence interval (CI) in relation to the traditional null-hypothesis test (Figure 1a) and non-inferiority test (Figure 1c) and the 90% CI in relation to the two one-sided test procedure (Figure 1b). The shaded areas indicate the rejection region for each hypothesis test. Figure 1a The superiority of the experimental group (EXP) compared with the control (CON) can be concluded in both the upper and middle scenarios. However, it is possible to reject effects that are smaller than the smallest effect of interest (SESOI) only in the middle scenario. Superiority cannot be concluded in the lower scenario, since the 95% CI extends beyond zero. Figure 1b It is possible to conclude equivalence between the interventions only in the middle example since in the upper and lower example the 90% CI spans beyond the lower (Δ_L) or the upper (Δ_U) equivalence margin. Figure 1c The observed data are identical to Figure 1b. Despite the wider CI, the absence of an upper margin allows concluding non-inferiority in both the middle and lower scenario.

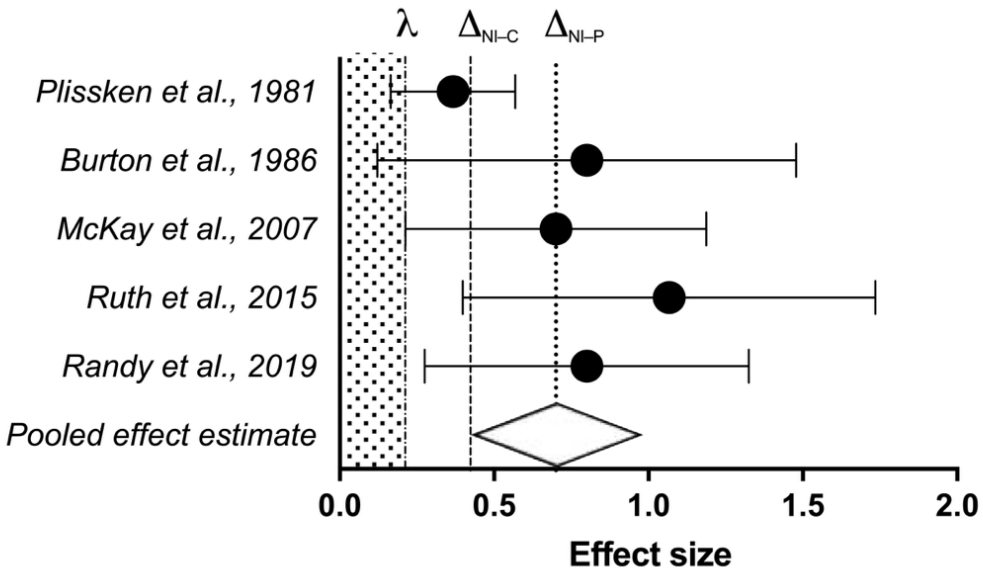

Figure 2 The two-step process commonly employed to determine the non-inferiority margin (Δ_{NI}) in clinical research. A pooled effect estimate is calculated from a meta-analysis of hypothetical studies and the margin is determined using either the point estimate (point-estimate method; Δ_{NI-P}) or the lower 95% confidence limit (fixed-margin method; Δ_{NI-C}) of the effect size. The chosen margin (Δ_{NI-C} in the example) is then multiplied by a pre-specified factor (λ; usually 50%) to preserve a fraction of the active-control effect (shaded area).

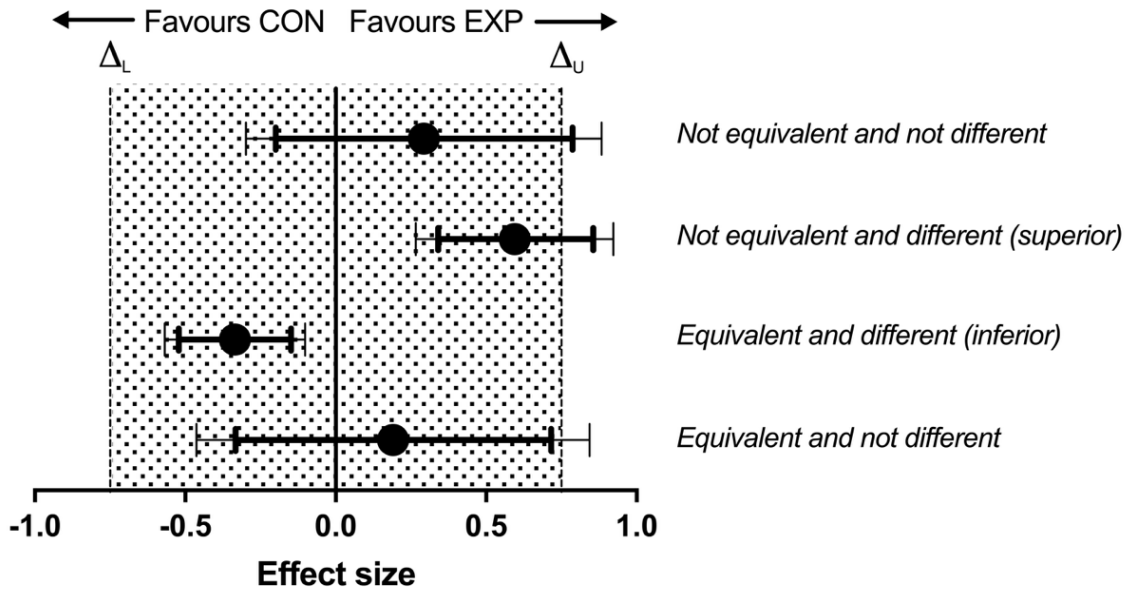

Figure 3 Testing for both equivalence and superiority. The thin error bars indicate the 95% confidence interval (CI) in relation to the traditional null-hypothesis test, whereas the thick error bars indicate the 90% CI in relation to the two one-sided tests procedure. The solid vertical lines indicate the traditional null hypothesis, whereas the shaded area indicates the equivalence region. Conclusions for hypothesis tests are reported next to each example.

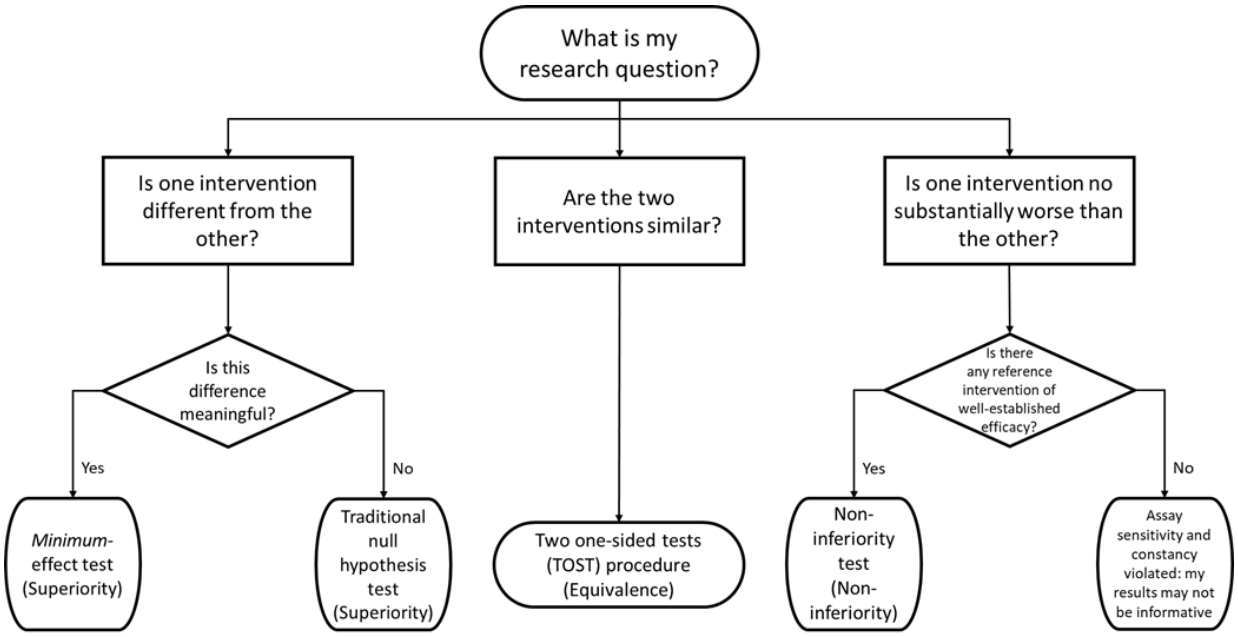

Figure 4 Processes of decision making for selecting the different hypothesis tests based on the research question that is being asked.

Contributions

Contributed to conception and design: RM, SP, DJB, DL
Contributed to acquisition of data: N/A
Contributed to analysis and interpretation of data: RM, SP, DJB, DL
Drafted and/or revised the article: RM, SP, DJB, DL
Approved the submitted version for publication: RM, SP, DJB, DL

Funding information

This work was funded by VIDI Grant 452-17-013 from the Netherlands Organisation for Scientific Research.

Data and Supplementary Material Accessibility

The workbook and the code used to perform all the calculations reported in this review are openly available at: https://osf.io/ndqhe/

REFERENCES

[1] J. Aisbett, D. Lakens, and K. L. Sainani, K. L. “Magnitude based inference in relation to one-sided hypotheses testing procedures”. In: SportRxiv (2020). https://doi.org/10.31236/osf.io/pn9s3
[2] T. A. Althunian, A. de Boer, R. H. H. Groenwold, and O. H. Klungel. “Defining the noninferiority margin and analysing noninferiority: an overview”. In: British Journal of Clinical Pharmacology 83 (2017), pp. 1636–1642. https://doi.org/10.1111/bcp.13280
[3] T. A. Althunian, A. de Boer, R. H. H. Groenwold, and O. H. Klungel. “Using a single noninferiority margin or preserved fraction for an entire pharmacological class was found to be inappropriate”. In: Journal of Clinical Epidemiology 104 (2018), pp. 15–23. https://doi.org/10.1016/j.jclinepi.2018.07.004

[4] D. G. Altman and J. M. Bland. “Absence of evidence is not evidence of absence”. In: British Medical Journal 311 (1995), pp. 485. https://doi.org/10.1136/bmj.311.7003.485
[5] R. L. Berger and J. C. Hsu. “Bioequivalence trials, intersection-union tests and equivalence confidence sets”. In: Statistical Science 11 (1996), pp. 283–319. https://doi.org/10.1214/ss/1032280304
[6] A. R. Caldwell and S. N. Cheuvront. “Basic statistical considerations for physiology: the journal Temperature toolbox”. In: Temperature (Austin) 6 (2019), pp. 181−210. https://doi.org/10.1080/23328940.2019.1624131
[7] A. Caldwell and A. D. Vigotsky. “A case against default effect sizes in sport and exercise science”. In: PeerJ 8 (2020), pp. e10314. https://doi.org/10.7717/peerj.10314
[8] A. R. Caldwell, A. D. Vigotsky, M. S. Tenan, R. Radel, D. T. Mellor, A. Kreutzer, I. M. Lahart, J. P. Mills, M. P. Boisgontier, and Consortium for Transparency in Exercise Science (COTES) Collaborators. “Moving sport and exercise science forward: a call for the adoption of more transparent research practices”. In: Sports Medicine 50 (2020), pp. 449–459. https://doi.org/10.1007/s40279-019-01227-1
[9] E. C. Carter, F. D. Schönbrodt, W. M. Gervais, and J. Hilgard. “Correcting for bias in psychology: a comparison of meta-analytic methods”. In: Advances in Methods and Practices in Psychological Science 2 (2019), pp. 115−144. https://doi.org/10.1177/2515245919847196
[10] J. Castelloe and D. Watts. “Equivalence and Noninferiority Testing Using SAS/STAT® Software”. In: Proceedings of the SAS Global Forum 2015 Conference. Cary, NC: SAS Institute Inc. https://support.sas.com/resources/papers/proceedings15/SAS1911-2015.pdf
[11] M. Cocks, C. S. Shaw, S. O. Shepherd, J. P. Fisher, A. Ranasinghe, T. A. Barker, and A. J. Wagenmakers. “Sprint interval and moderate-intensity continuous training have equal benefits on aerobic capacity, insulin sensitivity, muscle capillarisation and endothelial eNOS/NAD(P)Hoxidase protein ratio in obese men”. In: Journal of Physiology 594 (2016), pp. 2307−2321. https://doi.org/10.1113/jphysiol.2014.285254
[12] J. Cohen, J. “Statistical power analysis for the behavioral sciences (2nd ed.)”. (1988). Hillsdale, NJ: Lawrence Earlbaum Associates. https://doi.org/10.4324/9780203771587
[13] Committee for Medicinal Products for Human Use. (2005). “Guideline on the choice of the non-inferiority margin”. European Medicines Agency. https://www.ema.europa.eu/en/documents/scientific-guideline/guideline-choice-non-inferiority-margin_en.pdf
[14] Committee for Medicinal Products for Human Use. (2010). “Guideline on the investigation of bioequivalence”. European Medicines Agency.

https://www.ema.europa.eu/en/documents/scientific-guideline/guideline-investigation-bioequivalence-rev1_en.pdf
[15] Committee for Proprietary Medicinal Products. (2000). “Points to consider on switching between superiority and non-inferiority”. European Medicines Agency. https://www.ema.europa.eu/en/documents/scientific-guideline/points-consider-switching-between-superiority-non-inferiority_en.pdf
[16] J. A. Cook, S. A. Julious, W. Sones, L. V. Hampson, C. Hewitt, J. A. Berlin, D. Ashby, R. Emsley, D. A. Fergusson, S. J. Walters, E. C. F. Wilson, G. Maclennan, N. Stallard, J. C. Rothwell, M. Bland, L. Brown, C. R. Ramsay, A. Cook, D. Armstrong, ... L. D. Vale. “DELTA^2 guidance on choosing the target difference and undertaking and reporting the sample size calculation for a randomised controlled trial”. In: British Medical Journal 363 (2018), pp. k3750. https://doi.org/10.1186/s13063-018-2884-0
[17] P. M. Dixon, P. F. Saint-Maurice, Y. Kim, P. Hibbing, Y. Bai, and G. J. Welk. “A primer on the use of equivalence testing for evaluating measurement agreement”. In: Medicine & Science in Sports & Exercise 50 (2018), pp. 837−845. https://doi.org/10.1249/MSS.0000000000001481
[18] J. B. Gillen, B. J. Martin, M. J. MacInnis, L. E. Skelly, M. A. Tarnopolsky, and M. J. Gibala. “Twelve weeks of sprint interval training improves indices of cardiometabolic health similar to traditional endurance training despite a five-fold lower exercise volume and time commitment”. In: PLoS One 11 (2016), pp. e0154075. http://doi.org/10.1371/journal.pone.0154075
[19] I. Halperin, A. D. Vigotsky, C. Foster, and D. B. Pyne. “Strengthening the practice of exercise and sport-science research”. In: International Journal of Sports Physiology and Performance 13 (2018), pp. 127−134. https://doi.org/10.1123/ijspp.2017-0322
[20] A. Hecksteden, O. Faude, T. Meyer, and L. Donath. “How to construct, conduct and analyze an exercise training study?” in: Frontiers in Physiology 9 (2018), pp. 1007. https://doi.org/10.3389/fphys.2018.01007
[21] J. P. T. Higgins, J. Thomas, J. Chandler, M. Cumpston, T. Li, M. J. Page, and V. A. Welch. “Cochrane handbook for systematic reviews of interventions (2nd ed.)”. (2019). Chichester: Wiley. https://doi.org/10.1002/9781119536604
[22] J. L. Hodges and E. L. Lehmann. “Testing the approximate validity of statistical hypotheses.” In: Journal of the Royal Statistical Society Series B (Statistical Methodology) 16 (1954), pp. 261–268. https://doi.org/10.1111/j.2517-6161.1954.tb00169.x

[23] E. B. Holmgren. “Establishing equivalence by showing that a specified percentage of the effect of the active control over placebo is maintained”. In: Journal of Biopharmaceutical Statistics 9 (1999), pp. 651–659. https://doi.org/10.1081/bip-100101201
[24] W. G. Hopkins, J. A. Hawley, and L. M. Burke. Design and analysis of research on sport performance enhancement. In: Medicine & Science in Sports & Exercise 31 (1999), pp. 472−485. https://doi.org/10.1097/00005768-199903000-00018
[25] International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use (1998). “ICH E9: statistical principles for clinical trials.” European Medicines Agency. https://www.ema.europa.eu/en/documents/scientific-guideline/ich-e-9-statistical-principles-clinical-trials-step-5_en.pdf
[26] International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use (2001). “ICH E10: Choice of control group in clinical trials.” European Medicines Agency. https://www.ema.europa.eu/en/documents/scientific-guideline/ich-e-10-choice-control-group-clinical-trials-step-5_en.pdf
[27] S. A. Julious. “Sample sizes for clinical trials with normal data”. In: Statistics in Medicine 23 (2004), pp. 1921–1986. https://doi.org/10.1002/sim.1783
[28] D. Lakens. “Equivalence tests: a practical primer for t tests, correlations, and meta-analyses”. Social Psychological and Personality Science 8 (2017), pp. 355−362. https://doi.org/10.1177/1948550617697177
[29] D. Lakens. “Sample size justification”. In: PsyArXiv (2019). https://doi.org/10.31234/osf.io/9d3yf
[30] D. Lakens, F. G. Adolfi, C. J. Albers, F. Anvari, M. A. J. Apps, S. E. Argamon, T. Baguley, R. B. Becker, S. D. Benning, D. E. Bradford, E. M. Buchanan, A. R. Caldwell, B. Van Calster, R. Carlsson, S−C. Chen, B. Chung, L. J. Colling, G. S. Collins, Z. Crook, … Zwaan, R. A. (2018). “Justify your alpha”. In. Nature Human Behaviour 2 (2018), pp. 168–171. https://doi.org/10.1038/s41562-018-0311-x
[31] D. Lakens and E. R. K. Evers. “Sailing from the seas of chaos into the corridor of stability: practical recommendations to increase the informational value of studies”. In: Perspectives on Psychological Science 9 (2014), pp. 278–292. https://doi.org/10.1177/1745691614528520
[32] D. Lakens, N. McLatchie, P. M. Isager, A. M. Scheel, and Z. Dienes. “Improving inferences about null effects with Bayes factors and equivalence tests”. In: J The Journals of Gerontology, Series B: Psychological Sciences and Social Sciences 75 (2020), pp. 45−57. https://doi.org/10.1093/geronb/gby065

[33] D. Lakens, F. Pahlke, and G. Wassmer. “Group sequential designs: a tutorial”. In: PsyArXiv (2021). https://doi.org/10.31234/osf.io/x4azm
[34] D. Lakens, A. M. Scheel, and P. M. Isager. “Equivalence testing for psychological research: a tutorial”. In: Advances in Methods and Practices in Psychological Science 1 (2018), pp. 259−269. https://doi.org/10.1177/2515245918770963
[35] K. Magnusson (2021, October 24). “Equivalence, non-inferiority and superiority testing − an interactive visualization”. In: R Psychologist. https://rpsychologist.com/d3/equivalence/
[36] M. A. Mansournia and D. G. Altman. “Invited commentary: methodological issues in the design and analysis of randomised trials”. In: British Journal of Sports Medicine 52 (2018), pp. 553−555. https://doi.org/10.1136/bjsports-2017-09824515
[37] M. Meyners. “Equivalence tests – a review”. In: Food Quality and Preference 26 (2012), pp. 231–245. https://doi.org/10.1016/j.foodqual.2012.05.003
[38] Z. Milanović, G. Sporiš, and M. Weston. “Effectiveness of high-intensity interval training (HIT) and continuous endurance training for VO_{2max} improvements: a systematic review and meta-analysis of controlled trials”. In: Sports Medicine 45 (2015), pp. 1469−1481. http://doi.org/10.1007/s40279-015-0365-0
[39] K. R. Murphy, B. Myors, and A. Wolach. “Statistical power analysis: A simple and general model for traditional and modern hypothesis tests (4^{th} ed.)”. (2014). New York, NY: Routledge. https://doi.org/10.4324/9781315773155
[40] M. R. Rhea. “Determining the magnitude of treatment effects in strength training research through the use of the effect size”. In: The Journal of Strength & Conditioning Research 18 (2004), pp. 918−920. https://doi.org/10.1519/14403.1
[41] R. Ross, S. N. Blair, R. Arena, T. S. Church, J. P. Després, B. A. Franklin, W. L. Haskell, L. A. Kaminsky, B. D. Levine, C. J. Lavie, J. Myers, J. Niebauer, R. Sallis, S. S. Sawada, X. Sui, U. Wisløff, American Heart Association Physical Activity Committee of the Council on Lifestyle and Cardiometabolic Health, Council on Clinical Cardiology, Council on Epidemiology and Prevention, Council on Cardiovascular and Stroke Nursing, Council on Functional Genomics and Translational Biology, and Stroke Council. “Importance of assessing cardiorespiratory fitness in clinical practice: a case for fitness as a clinical vital sign: a scientific statement from the American Heart Association”. In: Circulation 134 (2016), pp. e653−e699. https://doi.org/10.1161/CIR.0000000000000461
[42] D. J. Schuirmann. “A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability”. In: Journal of

Pharmacokinetics and Pharmacodynamics 15 (1987), pp. 657–680. https://doi.org/10.1007/BF01068419
[43] J. Schumi and J. T. Wittes. "Through the looking glass: understanding non-inferiority". In: Trials 12 (2011), pp. 106. https://doi.org/10.1186/1745-6215-12-106
[44] S. Senn. "Statistical issues in drug development (2nd ed.)." (2007). Hoboken, NJ: Wiley. https://doi.org/10.1002/9780470723586
[45] U. Simonsohn, L. D. Nelson, and J. P. Simmons. "*p*-curve and effect size: correcting for publication bias using only significant results". In: Perspectives on Psychological Science 9 (2014), pp. 666–681. https://doi.org/10.1177/1745691614553988
[46] S. M. Snapinn. "Alternatives for discounting in the analysis of noninferiority trials". In: Journal of Biopharmaceutical Statistics 14 (2004), pp. 263–273. https://doi.org/10.1081/BIP-120037178
[47] S. Snapinn and Q. Jiang. "Controlling the type 1 error rate in non-inferiority trials". In: Statistics in Medicine 27 (2008), pp. 371–381. https://doi.org/10.1002/sim.3072
[48] S. Snapinn and Q. Jiang. "Preservation of effect and the regulatory approval of new treatments on the basis of non-inferiority trials". In: Statistics in Medicine 27 (2008), pp. 382–391. https://doi.org/10.1002/sim.3073
[49] H. D. Speed and M. B. Andersen. "What exercise and sport scientists don't understand". In: Journal of Science and Medicine in Sport 3 (2000), pp. 84–92. https://doi.org/10.1016/s1440-2440(00)80051-1
[50] D. Van Ravenzwaaij, R. Monden, J. N. Tendeiro, and J. P. A. Ioannidis. "Bayes factors for superiority, non-inferiority, and equivalence designs". In: BMC Medical Research Methodology 19 (2019), pp. 71. https://doi.org/10.1186/s12874-019-0699-7
[51] N. Victor. "On clinically relevant differences and shifted null hypotheses". In: Methods of Information in Medicine 26 (1987), pp. 109–116. https://doi.org/10.1055/s-0038-1635499
[52] S. J. Wang and J. D. Blume. "An evidential approach to non-inferiority clinical trials." In: Pharmaceutical Statistics 10 (2011), pp. 440–447. https://doi.org/10.1002/pst.513
[53] W. J. Westlake. "Response to T.B.L. Kirkwood: Bioequivalence testing – a need to rethink". In: Biometrics 37 (1981), pp. 589–594. https://doi.org/10.2307/2530573
[54] B. Yu, H. Yang, and B. Sabin. "A note on the determination of non-inferiority margins with application in oncology clinical trials". In: Contemporary Clinical Trials Communications 16 (2019), pp. 100454. https://doi.org/10.1016/j.conctc.2019.100454