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

Received: 31 October 2021 | Accepted: 10 January 2022

DOI: 10.1113/EP090171

Experimental Physiology WILEY

MYTHS AND METHODOLOGIES

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

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

[1] Department of Physical Education and Sport, University of the Basque Country (UPV/EHU), Vitoria-Gasteiz, Spain

[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, Victoria, Australia

[5] Human Technology Interaction Group, Eindhoven University of Technology (TU/e), Eindhoven, The Netherlands

Correspondence
Raffaele Mazzolari, Department of Physical Education and Sport, University of the Basque Country, Portal de Lasarte 71, Vitoria-Gasteiz 01007, Spain.
Email: rmazzolari001@ikasle.ehu.eus

Linked articles: This article is highlighted in a Viewpoint article by Batterham. To read this paper, visit https://doi.org/10.1113/EP090321.

Funding information
Nederlandse Organisatie voor Wetenschappelijk Onderzoek, Grant/Award Number: VIDI Grant 452-17-013

Edited by: Jeremy Ward

Abstract

Exercise physiology and sport science have traditionally made use of the null hypothesis of no difference to make decisions about experimental interventions. In this article, we aim 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 from deciding whether an effect is present. Initially, 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.

KEYWORDS
intervention efficacy, methodology, statistical review

1 | 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 group and the control group on one or more primary outcomes. However, other hypothesis tests might be more appropriate when researchers are interested in whether two interventions are similar in efficacy but differ substantially with respect to factors such as cost-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 discussed extensively 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 review and expand the statistical toolset that can be used by exercise physiologists and sport scientists when designing and analysing interventional studies. We refer to the best practices as developed in biomedical, social and behavioural research, because we recognize sufficient similarities with exercise physiology and sport science regarding the design of interventional studies. To increase understanding by exercise physiologists and sport scientists, we 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 provide theoretical and practical recommendations to exercise physiologists and sport scientists who would like to apply the different hypothesis tests in future research.

2 | INVESTIGATING STATISTICAL DIFFERENCES (SUPERIORITY)

Unless otherwise specified, most interventional studies in exercise physiology and sport science have the implicit aim of determining whether the efficacy of a given intervention is superior, or possibly inferior, to a placebo, sham or reference intervention. In the most common study design, researchers randomize participants to either an experimental group 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 that 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 that there is no difference between the interventions and that all other modelling assumptions are met. Alternatively, researchers can choose a confidence interval (CI) approach. Confidence intervals can be used to inspect and interpret the point estimate and the lower and upper limits of the interval in relationship to effects of practical importance. Thus, a properly derived CI can also be used to evaluate superiority or any other family of null hypotheses (Bauer & Kieser, 1996). The two approaches lead to identical decisions in a hypothesis test, because *P* is ≤ 0.05 when a 95% CI excludes the value that is tested against (i.e., zero for the traditional null hypothesis) (Figure 1a, first example).

Regardless of the inferential approach adopted, 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 might 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. For adequately powered tests (i.e., 80–90% power), testing data against the nil (zero) effect using the smallest effect size of interest (SESOI), which should be defined a priori and justified on sound grounds, as a target mean difference might lead to concluding efficacy for effects as low as the 60–70% of SESOI, the so-called 'decision value' (Chuang-Stein et al., 2011; Roychoudhury et al., 2018). Chuang-Stein et al. (2011) recommended this approach as a reasonable compromise between desirability and feasibility, stressing how this approach also acknowledges the impact of sampling variation in reducing the observed intervention effect. However, although an effect as low as 60–70% of SESOI might be observed when the true effect equals the SESOI, the opposite might not necessarily be true. By rearranging the equation used to determine the decision value, it is possible to obtain an adequate sample size that leads to rejecting the null hypothesis when the decision value equals at least the SESOI. In this way, statistical significance is ensured whenever practical relevance is observed (Figure 1a, second example). For a deeper insight into the statistical aspects of this approach (named 'dual-criterion designs'), we refer the reader to Roychoudhury et al. (2018).

An even more conservative criterion for assessing superiority consists of determining whether the mean difference, after having considered its uncertainty, is larger than the SESOI (Lakens, 2021) (Figure 1a, third 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). However, raising the standard of evidence to claim superiority comes at a cost. Testing data against the SESOI can require sample sizes that are prohibitively large when the 'true' effect size is close to the SESOI unless prespecifying unrealistically large effect sizes with an attendant risk of type II error (Gelman & Carlin, 2014). Therefore, researchers should decide very carefully what standard of evidence they want to achieve for intervention efficacy when designing their studies, taking into account the implications of their findings and their resources.

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, participant's experience or clinical endpoint(s) to define the SESOI, provides a common approach to interpret study outcomes in clinical research. In this field, the SESOI (also known as the minimal clinically important difference) is often determined by examining the association between a certain change in an outcome variable and a meaningful change in a (hard) clinical outcome from prospective epidemiological data or randomized controlled trials.

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 can give an indication of the expected effect sizes. However, researchers should

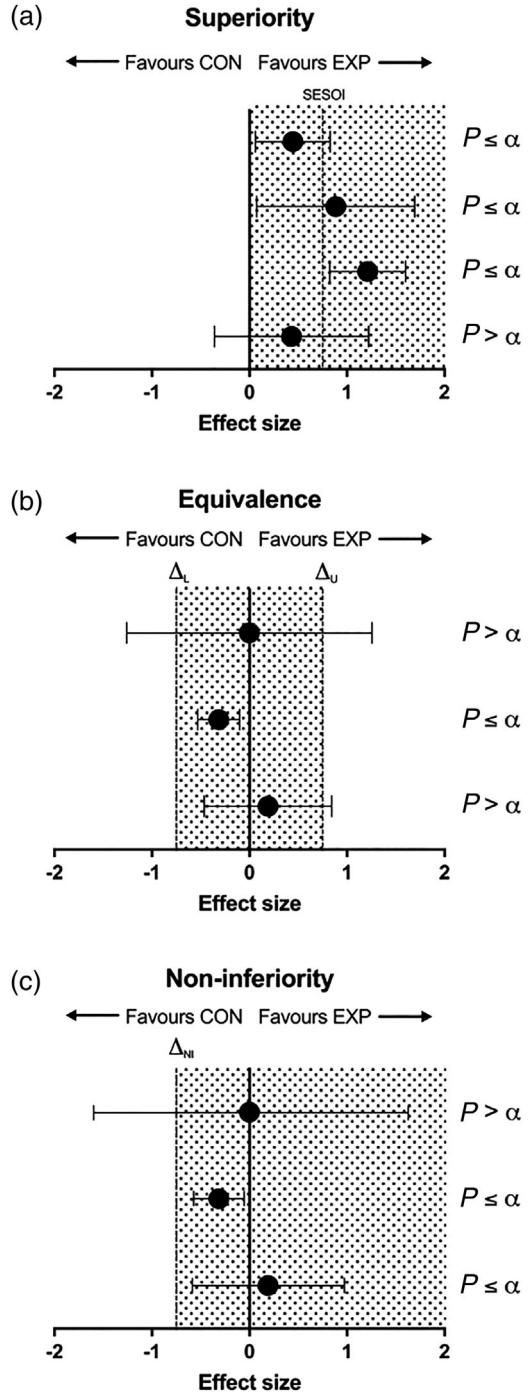

FIGURE 1 Testing for superiority, equivalence and non-inferiority within a typical parallel-group design. The error bars indicate the 95% confidence interval (CI) in relationship to the traditional null-hypothesis test (a) and non-inferiority test (c) and the 90% CI in relationship to the two one-sided test procedure (b). The stippled areas indicate the rejection region for each hypothesis test. (a) From a traditional perspective (i.e., deciding on the presence of an effect), the superiority of the experimental group (EXP) compared with the control (CON) can be concluded in the first three scenarios. However, the standard of evidence to claim superiority differs between the scenarios. In the first scenario, it is only possible to reject effects that are smaller than zero. In the second scenario, it is also possible to claim practical importance besides statistical significance. In the third scenario, it is possible to reject any effect that is not practically important [i.e., an effect that is smaller than the smallest effect of interest (SESOI)]. Superiority cannot be concluded in the fourth scenario, because the 95% CI extends beyond zero, which reflects in a *P*-value > α. (b) It is possible to conclude equivalence between the interventions only in the middle example, because in the upper and lower examples the 90% CI spans beyond the lower (Δ_L) or the upper (Δ_U) equivalence margin. (c) The observed data are identical to those in (b). Despite the wider CI, the absence of an upper margin allows conclusion of non-inferiority in both the middle and lower scenarios

be aware that owing to publication bias, published effect sizes often overestimate the true effect of interventions, and that the distribution of effect sizes observed in the 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, because 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 can 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., $P > 0.05$) cannot be interpreted as the absence of an effect (Figure 1a, fourth example). To be able to conclude that an effect is absent, one needs to specify the alternative hypothesis explicitly and perform a test that rejects the alternative hypothesis statistically. The traditional null hypothesis testing rejects only 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. Given that a well-designed study is informative about both the presence and the absence of an effect of interest, researchers should consider complementing traditional null hypothesis tests with equivalence and non-inferiority tests.

3 | 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 relevant practically, 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 0.05 or even to 0.025 in some regulated settings). 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, because 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 with brand-name pharmaceuticals to prove bioequivalence (Senn, 2021). 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, Scheel et al., 2018). The latter application of equivalence hypotheses might also make them valuable for exercise physiology and sport science, which suffer 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, in investigating 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 can 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 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, 2021) (Figure 1c, middle and lower examples).

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 briefly review and discuss the main aspects of each of these challenges in the following sections. Given that some of these issues also apply to equivalence studies, we expand those parts where relevant.

3.1 | Suitability of the reference intervention

From a clinical perspective, the non-inferiority of an experimental intervention can be concluded firmly only when compared with 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 can 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 characterizing 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 crucial 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 demonstrated consistently 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 reduced dramatically if pre-registration or Registered Reports Protocols became common practice in exercise physiology and sport science (Caldwell et al., 2020; Lakens & Evers, 2014). In this regard, the recent initiative of *Experimental Physiology* to publish Registered Reports Protocols (Stewart, 2021) deserves credit.

The aforementioned 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

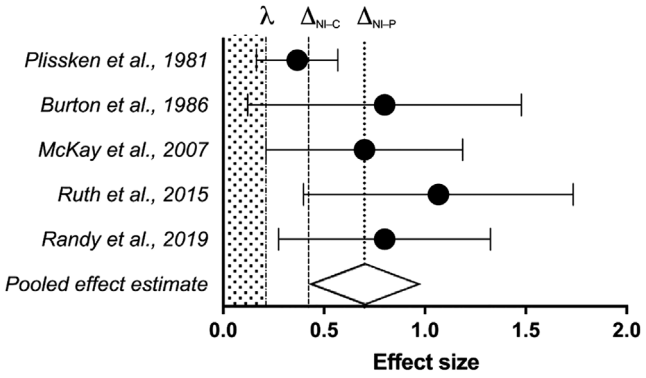

FIGURE 2 The two-step process commonly used to determine the non-inferiority margin (Δ_{NI}) in clinical research. A pooled effect estimate is calculated from a meta-analysis of hypothetical studies, and the margin is determined using either the point estimate (point-estimate method; Δ_{NI-P}) or the 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 prespecified factor (λ; usually 50%) to preserve a fraction of the active-control effect (stippled area)

emphasizes 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).

3.2 | 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, 2018a; 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-analytical 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, prespecifying 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, 2018b). Despite its widespread use in clinical research, it is important to note that there is no consensus on whether setting the Δ_{NI} by including a preserved fraction represents an effective discounting approach (Snapinn, 2004; Snapinn & Jiang, 2018b).

Whether or not the stringency in the criteria to determine non-inferiority should be adjusted further 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 might take place for largely efficacious standard interventions. In these cases, determining the fraction without any adjustment for the historical effect of the comparator might 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 can be applied in these situations (Schumi & Wittes, 2011).

(Bio)equivalence margins in clinical trials are often set by regulatory authorities (Committee for Medicinal Products for Human Use, 2010), whereas several approaches to justify the equivalence range have been proposed in the social and behavioural sciences (Lakens, 2017, 2021; Lakens, Scheel et al., 2018). 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 can be taken 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 investigated properly, especially in new lines of research. In such conditions, determining Δ_U and Δ_L based on feasibility can be justified and can 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.

3.3 | Sample size planning for non-inferiority and equivalence studies

As we discussed previously, sample size estimation in superiority studies 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 value of α (Cook et al., 2018).

Given that 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. Owing 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 might 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 can opt for sequential analysis (Lakens et al., 2021). This efficient approach allows terminatiion of 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 and approximations to calculate power and sample size in superiority, equivalence and non-inferiority studies. Researchers who wish to exact solutions for power and sample size for equivalence designs might look at the paper by Shieh (2016). Moreover, there are several spreadsheets (Lakens, 2017), statistical packages (Castelloe & Watts, 2015; Lakens, 2017) and Web-based applications (Kovacs et al., 2021; Magnusson, 2016) that exercise physiologists and sport scientists can use to estimate sample sizes for equivalence and non-inferiority tests.

4 | RE-IMAGINING 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 workbook (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 data sets.

4.1 | Example 1: Use of equivalence hypothesis

In a comprehensive study investigating the effects of 4 weeks of SIT (60 min per week) or MICT (300 min per week) on cardio-respiratory, musculoskeletal and metabolic characteristics in obese men, Cocks et al. (2016) concluded that SIT and MICT have equal benefits on aerobic capacity, because no statistical difference was observed between the two groups with respect to the changes in maximal oxygen uptake ($\dot{V}_{O_2\,max}$). As already stated, the absence of an effect cannot be concluded based on $P > 0.05$ 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 $\dot{V}_{O_2\,max}$ between the groups. Given that 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 $r = 0.5$, the SIT–MICT 90% CI around the observed mean difference of −2.3 ml/kg/min ranges from −7.1 to 2.5 ml/kg/min. 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/min) when $r = 0.7$. However, even in the optimistic scenario in which $r = 0.9$, the 90% CI for the between-group difference ranges from −5.2 to 0.6 ml/kg/min, which indicates a large imprecision of the parameter estimate. Given that a difference in $\dot{V}_{O_2\,max}$ as small as 1 ml/kg/min has been associated with a 9% instantaneous relative risk reduction for all-cause mortality (hazard ratio 0.91) (Laukkanen et al., 2016), the mean difference between SIT and MICT that was observed by Cocks et al. (2016) of −2.3 ml/kg/min is hardly trivial, let alone after having considered its uncertainty.

If we wish, we can also test formally for equivalence against symmetric margins Δ_U and Δ_L of 1 ml/kg/min by using the TOST procedure, which is very similar to Student's unpaired *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 1 ml/kg/min, 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 sizes 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:

$$t_U = \frac{2.4 \; - \; 4.7 \; - \; 1}{3.3\sqrt{\frac{1}{8} + \frac{1}{8}}} = -2$$

which corresponds to a *P*-value of 0.03 from the *t*-distribution with 14 degrees of freedom (d.f.) for a left-sided test.

For Δ_L:

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

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

In this example,

$$t_L = \frac{2.4 - 4.7 - (-1)}{3.3\sqrt{\frac{1}{8} + \frac{1}{8}}} = -0.8,$$

which corresponds to a *P*-value of 0.78 from the *t*-distribution with 14 d.f. for a right-sided test.

Given that the one-sided test with the greater *P*-value is not statistically significant [$t(14) = -0.8$, $P = 0.78$] based on an $\alpha = 0.05$, we cannot reject differences larger than 1 ml/kg/min. Therefore, we cannot conclude that the difference between the two interventions is too small to matter (given a SESOI of 1 ml/kg/min) with respect to the changes in $\dot{V}_{O_2\,max}$.

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 et al. (2016) 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., 0.05), 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 rely on the normal approximation of the power equation for equivalence tests (Julious, 2004) and estimate the sample size (n) required in each group to achieve the desired power against Δ_U and Δ_L as:

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

and

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

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 $1 - \beta$ that represents the desired power). With an equal allocation (1:1 ratio), the two equations above are reduced to:

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

In this example,

$$n = \frac{2 \times 3.3^2(1.6 + 1.3)^2}{1^2} = 192,$$

which indicates that the minimum sample size that Cocks et al. (2016) should have recruited to have a properly powered test for equivalence was 24 times larger than the $n = 8$ per group that was collected in that study. Even using a much more liberal SESOI of 3.5 ml/kg/min, associated with ≤25% risk reduction in mortality (Ross et al., 2016), the minimum sample size should have been double the one collected. Note that these also represent optimistic estimations; any situation in which some inequality between interventions can be expected (i.e., the expected difference is not zero) would increase the required sample size, all else being equal.

4.2 | 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 $\dot{V}_{O_2\,max}$, 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, given that the authors reported the 95% CI for the change scores of the two groups, in addition to their sample sizes, we can obtain the information we need from statistical first principles (Higgins et al., 2019). The calculations reveal a *P*-value of 0.94 and a 95% CI ranging from −2.9 to 2.7 ml/kg/min constructed around a mean difference between the interventions of −0.1 ml/kg/min. From a superiority standpoint, the study is inconclusive regarding the ability of SIT to improve the $\dot{V}_{O_2\,max}$ compared with MICT. Given the rationale supporting the study, a more informative research question might be whether the improvements in the $\dot{V}_{O_2\,max}$ induced by SIT are not substantially lower than those induced by a standard MICT programme. To answer such a question, we must initially define the Δ_{NI} that we will use to test our hypothesis. The net effect of MICT against no-exercise control on $\dot{V}_{O_2\,max}$ has been estimated to be 4.9 ml/kg/min, with a 95% CI ranging from 3.5 to 6.3 ml/kg/min (Milanović et al., 2015). If we assume that the MICT protocol prescribed by Gillen et al. (2016) is sufficiently representative

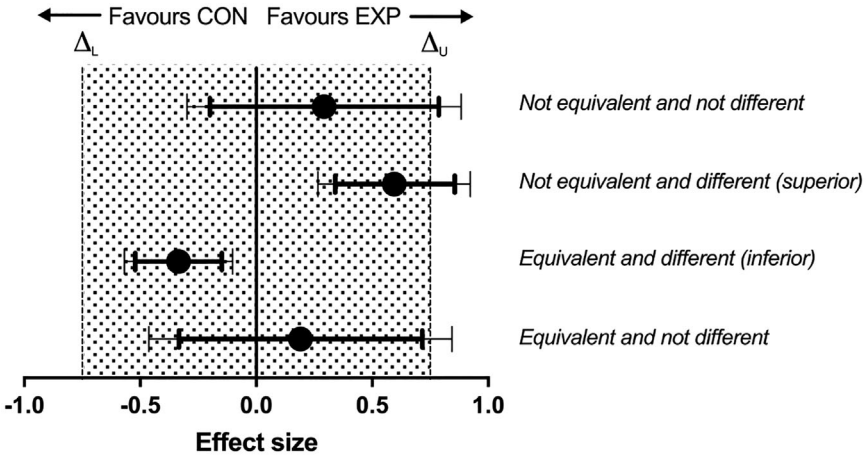

FIGURE 3 Testing for both equivalence and superiority. The thin error bars indicate the 95% confidence interval (CI) in relationship to the traditional null-hypothesis test, whereas the thick error bars indicate the 90% CI in relationship to the two one-sided tests procedure. The continuous vertical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

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/min. The calculation of the *t*-statistic for the non-inferiority test is identical to those for the one-sided test against the Δ_L in the TOST procedure:

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

With t_{NI} being 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 0.02 from the *t*-distribution with 17 d.f. for a two-sided test. If all the assumptions underlying the statistical model are correct, the non-inferiority test is significant [$t(17) = 2.6$, $P = 0.02$] for an $\alpha = 0.05$. We can then reject a loss in the efficacy of SIT compared with MICT larger than 3.5 ml/kg/min and conclude that SIT is non-inferior to MICT regarding the increase in $\dot{V}_{O_2\,max}$. Unsurprisingly, given the close relationship between *P*-values and CIs, the CI approach leads to the same conclusion as the formal non-inferiority test, because the lower 95% confidence limit of the SIT–MICT difference (i.e., −2.9 ml/kg/min) is larger than the Δ_{NI} (i.e., −3.5 ml/kg/min), 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.

5 | SWITCHING BETWEEN HYPOTHESES

Switching the objective of a clinical trial from non-inferiority to superiority or vice versa can 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, because 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 order of testing 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} might help to minimize this risk.

Departing from the initial aim of establishing equivalence does not appear to be a common practice in clinical research (Senn, 2021). 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 presented recently in the social and behavioural sciences literature (Lakens, 2017; Lakens, Scheel et al., 2018) (Figure 3). Exercise physiologists and sport scientists interested in conducting equivalence and non-inferiority studies might benefit from exploring these approaches.

It is also worth mentioning the possibility of testing 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.

6 | 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

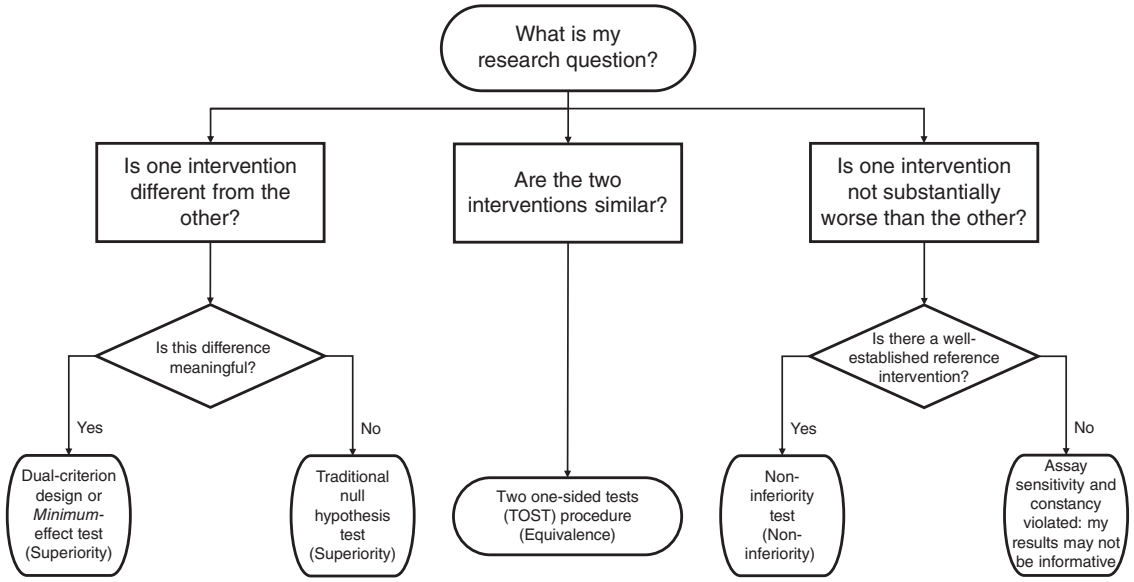

FIGURE 4 Processes of decision making for selecting the different hypothesis tests based on the research question that is being asked

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 might differ slightly 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, Adolfi et al., 2018).

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. Wald's statistical decision theory provides a coherent frequentist framework to use sample data to make decisions on interventions (Manski, 2019). Compared with hypothesis testing, the Wald framework has the advantage of taking into account the magnitudes of the losses that type I and II errors (whose probabilities are considered symmetrically) yield as an integral part of the framework. Among the alternative or complementary methods to frequentist statistics, 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 allow the incorporation of 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 was beyond the scope of the present paper, 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).

7 | 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. Although it should be clear at this point that this is a statistical misconception, exercise physiologists and sport 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 can be adopted whenever the research context, conditions, applications, researchers' interests or reasonable beliefs justify them. Although these research hypotheses require methodological considerations additional to superiority hypotheses to be investigated properly, they might also provide better answers to the empirical question in which researchers are interested. Equivalence and non-inferiority studies might help exercise physiologists and sport 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.

ACKNOWLEDGEMENTS

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

COMPETING INTERESTS

None declared.

AUTHOR CONTRIBUTIONS

R.M. conceived the initial idea; all authors contributed to its refinement. All authors contributed to the intellectual content and drafting of the manuscript. All authors approved the final version of the manuscript and agree to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. All persons designated as authors qualify for authorship, and all those who qualify for authorship are listed.

DATA AVAILABILITY STATEMENT

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

ORCID

Raffaele Mazzolari https://orcid.org/0000-0002-9923-6018
Simone Porcelli https://orcid.org/0000-0002-9494-0858
David J. Bishop https://orcid.org/0000-0002-6956-9188
Daniël Lakens https://orcid.org/0000-0002-0247-239X

REFERENCES

Aisbett, J., Lakens, D., & Sainani, K. L. (2020). Magnitude based inference in relation to one-sided hypotheses testing procedures. *SportRxiv*. https://doi.org/10.31236/osf.io/pn9s3

Althunian, T. A., de Boer, A., Groenwold, R. H. H., & Klungel, O. H. (2017). Defining the noninferiority margin and analysing noninferiority: An overview. *British Journal of Clinical Pharmacology*, *83*, 1636–1642. https://doi.org/10.1111/bcp.13280

Althunian, T. A., de Boer, A., Groenwold, R. H. H., & Klungel, O. H. (2018). Using a single noninferiority margin or preserved fraction for an entire pharmacological class was found to be inappropriate. *Journal of Clinical Epidemiology*, *104*, 15–23. https://doi.org/10.1016/j.jclinepi.2018.07.004

Altman, D. G., & Bland, J. M. (1995). Absence of evidence is not evidence of absence. *British Medical Journal*, *311*, 485. https://doi.org/10.1136/bmj.311.7003.485

Bauer, P., & Kieser, M. (1996). A unifying approach for confidence intervals and testing of equivalence and difference. *Biometrika*, *83*, 934–937. https://doi.org/10.1093/biomet/83.4.934

Berger, R. L., & Hsu, J. C. (1996). Bioequivalence trials, intersection–union tests and equivalence confidence sets. *Statistical Science*, *11*(4), 283–319. https://doi.org/10.1214/ss/1032280304

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

Caldwell, A., & Vigotsky, A. D. (2020). A case against default effect sizes in sport and exercise science. *PeerJ*, *8*, e10314. https://doi.org/10.7717/peerj.10314

Caldwell, A. R., Vigotsky, A. D., Tenan, M. S., Radel, R., Mellor, D. T., Kreutzer, A., Lahart, I. M., Mills, J. P., Boisgontier, M. P., & Consortium for Transparency in Exercise Science (COTES) Collaborators. (2020). Moving sport and exercise science forward: A call for the adoption of more transparent research practices. *Sports Medicine*, *50*, 449–459. https://doi.org/10.1007/s40279-019-01227-1

Carter, E. C., Schönbrodt, F. D., Gervais, W. M., & Hilgard, J. (2019). Correcting for bias in psychology: A comparison of meta-analytic methods. *Advances in Methods and Practices in Psychological Science*, *2*, 115–144. https://doi.org/10.1177/2515245919847196

Castelloe, J., & Watts, D. (2015). Equivalence and noninferiority testing using SAS/STAT® software (paper SAS1911-2015). *Proceedings of the SAS Global Forum 2015 Conference*. SAS Institute. https://support.sas.com/resources/papers/proceedings15/SAS1911-2015.pdf

Chuang-Stein, C., Kirby, S., Hirsch, I., & Atkinson, G. (2011). The role of the minimum clinically important difference and its impact on designing a trial. *Pharmaceutical Statistics*, *10*, 250–256. https://doi.org/10.1002/pst.459

Cocks, M., Shaw, C. S., Shepherd, S. O., Fisher, J. P., Ranasinghe, A., Barker, T. A., & Wagenmakers, A. J. (2016). 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. *The Journal of Physiology*, *594*, 2307–2321. https://doi.org/10.1113/jphysiol.2014.285254

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Earlbaum Associates. https://doi.org/10.4324/9780203771587

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

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

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

Cook, J. A., Julious, S. A., Sones, W., Hampson, L. V., Hewitt, C., Berlin, J. A., Ashby, D., Emsley, R., Fergusson, D. A., Walters, S. J., Wilson, E. C. F., Maclennan, G., Stallard, N., Rothwell, J. C., Bland, M., Brown, L., Ramsay, C. R., Cook, A., Armstrong, D., … Vale, L. D. (2018). DELTA^2 guidance on choosing the target difference and undertaking and reporting the sample size calculation for a randomised controlled trial. *British Medical Journal*, *19*, k3750. https://doi.org/10.1186/s13063-018-2884-0

Dixon, P. M., Saint-Maurice, P. F., Kim, Y., Hibbing, P., Bai, Y., & Welk, G. J. (2018). A primer on the use of equivalence testing for evaluating measurement agreement. *Medicine & Science in Sports & Exercise*, *50*, 837–845. https://doi.org/10.1249/MSS.0000000000001481

Gelman, A., & Carlin, J. (2014). Beyond power calculations: Assessing type S (sign) and type M (magnitude) errors. *Perspectives on Psychological Science: a Journal of the Association for Psychological Science*, *9*(6), 641–651. https://doi.org/10.1177/1745691614551642

Gillen, J. B., Martin, B. J., MacInnis, M. J., Skelly, L. E., Tarnopolsky, M. A., & Gibala, M. J. (2016). 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. *PLoS One*, *11*, e0154075. https://doi.org/10.1371/journal.pone.0154075

Halperin, I., Vigotsky, A. D., Foster, C., & Pyne, D. B. (2018). Strengthening the practice of exercise and sport-science research. *International Journal of Sports Physiology and Performance*, *13*, 127–134. https://doi.org/10.1123/ijspp.2017-0322

Hecksteden, A., Faude, O., Meyer, T., & Donath, L. (2018). How to construct, conduct and analyze an exercise training study? *Frontiers in Physiology*, *9*, 1007. https://doi.org/10.3389/fphys.2018.01007

Higgins, J. P. T., Thomas, J., Chandler, J., Cumpston, M., Li, T., Page, M. J., & Welch, V. A. (2019). *Cochrane handbook for systematic reviews of interventions* (2nd ed.). Wiley. https://doi.org/10.1002/9781119536604

Hodges, J. L., & Lehmann, E. L. (1954). Testing the approximate validity of statistical hypotheses. *Journal of the Royal Statistical Society Series B (Statistical Methodology)*, *16*, 261–268. https://doi.org/10.1111/j.2517-6161.1954.tb00169.x

Holmgren, E. B. (1999). Establishing equivalence by showing that a specified percentage of the effect of the active control over placebo is maintained.

Journal of Biopharmaceutical Statistics, *9*, 651–659. https://doi.org/10.1081/bip-100101201

Hopkins, W. G., Hawley, J. A., & Burke, L. M. (1999). Design and analysis of research on sport performance enhancement. *Medicine & Science in Sports & Exercise*, *31*, 472–485. https://doi.org/10.1097/00005768-199903000-00018

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

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

Julious, S. A. (2004). Sample sizes for clinical trials with normal data. *Statistics in Medicine*, *23*, 1921–1986. https://doi.org/10.1002/sim.1783

Kovacs, M., van Ravenzwaaij, D., Hoekstra, R., & Aczel, B. (2021). SampleSizePlanner: A tool to estimate and justify sample size for two-group studies. *MetaArXiv*. https://doi.org/10.31222/osf.io/rm9dn

Lakens, D., & Evers, E. R. K. (2014). Sailing from the seas of chaos into the corridor of stability: Practical recommendations to increase the informational value of studies. *Perspectives on Psychological Science*, *9*, 278–292. https://doi.org/10.1177/1745691614528520

Lakens, D. (2017). Equivalence tests: A practical primer for *t* tests, correlations, and meta-analyses. *Social Psychological and Personality Science*, *8*, 355–362. https://doi.org/10.1177/1948550617697177

Lakens, D., Adolfi, F. G., Albers, C. J., Anvari, F., Apps, M. A. J., Argamon, S. E., Baguley, T., Becker, R. B., Benning, S. D., Bradford, D. E., Buchanan, E. M., Caldwell, A. R., Van Calster, B., Carlsson, R., Chen, S.-C., Chung, B., Colling, L. J., Collins, G. S., Crook, Z., … Zwaan, R. A. (2018). Justify your alpha. *Nature Human Behaviour*, *2*, 168–171. https://doi.org/10.1038/s41562-018-0311-x

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

Lakens, D., McLatchie, N., Isager, P. M., Scheel, A. M., & Dienes, Z. (2020). Improving inferences about null effects with Bayes factors and equivalence tests. *The Journals of Gerontology, Series B: Psychological Sciences and Social Sciences*, *75*, 45–57. https://doi.org/10.1093/geronb/gby065

Lakens, D. (2021). Sample size justification. *PsyArXiv*. https://doi.org/10.31234/osf.io/9d3yf

Lakens, D., Pahlke, F., & Wassmer, G. (2021). *Group sequential designs: A tutorial*. PsyArXiv. https://doi.org/10.31234/osf.io/x4azm

Laukkanen, J. A., Zaccardi, F., Khan, H., Kurl, S., Jae, S. Y., & Rauramaa, R. (2016). Long-term change in cardiorespiratory fitness and all-cause mortality: A population-based follow-up study. *Mayo Clinic Proceedings*, *91*, 1183–1188. https://doi.org/10.1016/j.mayocp.2016.05.014

Magnusson, K. (2016) *Equivalence, non-inferiority and superiority testing – An interactive visualization*. R Psychologist. https://rpsychologist.com/d3/equivalence/

Manski, C. F. (2019). Treatment choice with trial data: Statistical decision theory should supplant hypothesis testing. *The American Statistician*, *75*, 265–275. https://doi.org/10.1080/00031305.2020.1717621

Mansournia, M. A., & Altman, D. G. (2018). Invited commentary: Methodological issues in the design and analysis of randomised trials. *British Journal of Sports Medicine*, *52*, 553–555. https://doi.org/10.1136/bjsports-2017-09824515

Meyners, M. (2012). Equivalence tests – A review. *Food Quality and Preference*, *26*, 231–245. https://doi.org/10.1016/j.foodqual.2012.05.003

Milanović, Z., Sporiš, G., & Weston, M. (2015). Effectiveness of high-intensity interval training (HIT) and continuous endurance training for VO_{2max} improvements: A systematic review and meta-analysis of controlled trials. *Sports Medicine*, *45*, 1469–1481. https://doi.org/10.1007/s40279-015-0365-0

Murphy, K. R., Myors, B., & Wolach, A. (2014). *Statistical power analysis: A simple and general model for traditional and modern hypothesis tests* (4th ed.). Routledge. https://doi.org/10.4324/9781315773155

Rhea, M. R. (2004). Determining the magnitude of treatment effects in strength training research through the use of the effect size. *The Journal of Strength & Conditioning Research*, *18*, 918–920. https://doi.org/10.1519/14403.1

Ross, R., Blair, S. N., Arena, R., Church, T. S., Després, J. P., Franklin, B. A., Haskell, W. L., Kaminsky, L. A., Levine, B. D., Lavie, C. J., Myers, J., Niebauer, J., Sallis, R., Sawada, S. S., Sui, X., & Wisløff, U., American Heart Association Physical Activity Committee of the Council on Lifestyle and Cardiometabolic Health, Council on Clinical Cardiology, Council on Epidemiology and Prevention, … Stroke Council. (2016). 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. *Circulation*, *134*, e653–e699. https://doi.org/10.1161/CIR.0000000000000461

Roychoudhury, S., Scheuer, N., & Neuenschwander, B. (2018). Beyond *p*-values: A phase II dual-criterion design with statistical significance and clinical relevance. *Clinical Trials*, *15*, 452–461. https://doi.org/10.1177/1740774518770661

Schuirmann, D. J. (1987). A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Pharmacodynamics*, *15*, 657–680. https://doi.org/10.1007/BF01068419

Schumi, J., & Wittes, J. T. (2011). Through the looking glass: Understanding non-inferiority. *Trials*, *12*, 106. https://doi.org/10.1186/1745-6215-12-106

Senn, S. (2021). *Statistical issues in drug development* (3rd ed.), Wiley.

Shieh, G. (2016). Exact power and sample size calculations for the two one-sided tests of equivalence. *PLoS One*, *11*, e0162093. https://doi.org/10.1371/journal.pone.0162093

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). *p*-Curve and effect size: Correcting for publication bias using only significant results. *Perspectives on Psychological Science*, *9*, 666–681. https://doi.org/10.1177/1745691614553988

Snapinn, S. M. (2004). Alternatives for discounting in the analysis of non-inferiority trials. *Journal of Biopharmaceutical Statistics*, *14*, 263–273. https://doi.org/10.1081/BIP-120037178

Snapinn, S., & Jiang, Q. (2018a). Controlling the type 1 error rate in non-inferiority trials. *Statistics in Medicine*, *27*, 371–381. https://doi.org/10.1002/sim.3072

Snapinn, S., & Jiang, Q. (2018b). Preservation of effect and the regulatory approval of new treatments on the basis of non-inferiority trials. *Statistics in Medicine*, *27*, 382–391. https://doi.org/10.1002/sim.3073

Speed, H. D., & Andersen, M. B. (2000). What exercise and sport scientists don't understand. *Journal of Science and Medicine in Sport*, *3*, 84–92. https://doi.org/10.1016/s1440-2440(00)80051-1

Stewart, A. (2021). Experimental physiology publishes one of the first registered reports in physiology. The Physiological Society. https://www.physoc.org/blog/experimental-physiology-publishes-one-of-the-first-registered-reports-in-physiology/

Van Ravenzwaaij, D., Monden, R., Tendeiro, J. N., & Ioannidis, J. P. A. (2019). Bayes factors for superiority, non-inferiority, and equivalence designs. *BMC Medical Research Methodology*, *19*, 71. https://doi.org/10.1186/s12874-019-0699-7

Victor, N. (1987). On clinically relevant differences and shifted null hypotheses. *Methods of Information in Medicine*, *26*, 109–116. https://doi.org/10.1055/s-0038-1635499

Wang, S. J., & Blume, J. D. (2011). An evidential approach to non-inferiority clinical trials. *Pharmaceutical Statistics*, *10*, 440–447. https://doi.org/10.1002/pst.513

Westlake, W. J. (1981). Response to T.B.L. Kirkwood: Bioequivalence testing—a need to rethink. *Biometrics*, *37*, 589–594. https://doi.org/10.2307/2530573

Yu, B., Yang, H., & Sabin, B. (2019). A note on the determination of non-inferiority margins with application in oncology clinical trials. *Contemporary Clinical Trials Communications*, *16*, 100454. https://doi.org/10.1016/j.conctc.2019.100454

How to cite this 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*, *107*, 201–212. https://doi.org/10.1113/EP090171