What is your hypothesis? On the importance of knowing your hypothesis before conducting a hypothesis test

Cristian Mesquida[1], Joe Warne[2] and Daniël Lakens[1]

[1]Human-Technology Interaction Group, Eindhoven University of Technology, Eindhoven, The Netherlands
[2]School of Biological, Health and Sports Sciences, Technological University Dublin, Tallaght, Dublin, Ireland

ORCIDs
Cristian Mesquida – 0000-0002-1542-8355
Joe Warne – 0000-0002-4359-8132
Daniël Lakens – 0000-0002-0247-239X

Correspondence
Cristian Mesquida; Human-Technology Interaction Group, Eindhoven University of Technology, The Netherlands, c.mesquida.caldentey@tue.nl

This manuscript is a preprint
Please cite as: Mesquida, C., Warne, J. & Lakens, D. (2025). What is your hypothesis? On the importance of knowing your hypothesis before conducting a hypothesis test. *SportRXiv*.

Abstract

Null hypothesis significance testing (NHST) is a methodological procedure that allows sports and exercise scientists to make claims about the effect of interventions while controlling error rates. To be useful, NHST requires a clearly defined hypothesis and the use of an appropriate hypothesis test. However, we contend that these two conditions are often not met, and as a result, NHST is frequently applied without rigor. This can lead to situations in which stated hypotheses are actually not tested (Misalignments 1-3), or where type I and type II errors are inflated (Misalignments 4-5), ultimately resulting in misleading claims. In this paper, we deconstruct a series of hypotheses into basic statements to identify five types of commonly observed misalignments and offer recommendations for their proper testing. To address these issues, we recommend increased collaboration between sports scientists and applied statisticians, enhanced statistical training and the adoption of the PICO framework, which provides a structured approach for clearly defining the hypothesis that a study aims to test.

1. The logic of hypothesis tests

Sports scientists and exercise physiologists (henceforth referred to as sports scientists) often aim to establish reliable claims about physical performance–such as to determine the effects of training interventions on physical performance. One common approach for establishing claims is the hypothetico-deductive method (Hempel, 1966). In this approach, researchers derive a hypothesis from a theory or empirical observations, which is then tested and subsequently either corroborated or falsified. A hypothesis is a testable verbal statement about the presence, absence, or magnitude of an effect or relationship between one or more variables within a given population.

Importantly, the act of testing a hypothesis entails the commitment to evaluate whether it is corroborated or falsified. This is essentially a discrete probability space with only two answers (yes or no)– a dichotomous claim– (Tunç et al., 2023; but also see Frick, 1996; Nickerson, 2000). By making a dichotomous claim researchers want to communicate whether the hypothesis of interest has been corroborated or not by empirical data. Since data inherently contain random variability, researchers need a tool like hypothesis tests to distinguish between patterns that may be random noise and those that are unlikely to be mere random noise. Sports scientists interested in making reliable claims while controlling the maximum error rate usually rely on the Neyman-Pearson approach to null-hypothesis significance testing (NHST). According to the logic of NHST (Figure 1), a hypothesis (H) is first translated into a pair of statistical hypotheses: H_0 and H_1. H_0 represents the *negation* of H and it should always include an equality sign (i.e., =, ≥ and/or ≤). H_1 represents the *assertion* of H, should always include the opposite sign as H_0 (i.e., ≠, > and/or <), and it includes all other values than are not contained in H_0. Thus, H_0 and H_1 are mutually exclusive– they cannot be true at the same time– and collectively exhaustive– they encompass the entire range of possible outcomes–. H_0 is almost always specified as an effect of 0, but the null hypothesis can be set to any non-zero value, or even as a range of values around 0 (Mazzolari et al., 2022; K. R. Murphy & Myors, 1999).

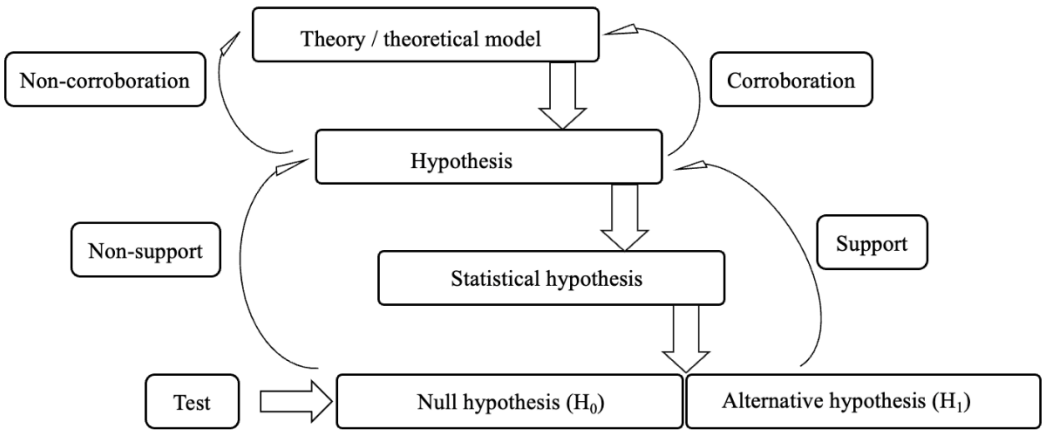

Figure 1. The logic of hypothesis testing according to the Neyman-Pearson approach to Null Hypothesis Significance Testing.

Once the researcher has assigned a specific value to H_0, H_1 will consist of all remaining possible values. A hypothesis test is performed where a test statistic is computed based on the data and an assumed statistical model, which is compared against the critical test statistic (t_c). The value of t_c defines the cutoff point for rejecting H_0 and it depends on the sample size, the statistical test, and the alpha level (α), which represents the maximum acceptable type I error rate (e.g., 5%) for the incorrect rejection of H_0. When the observed test statistic exceeds t_c, H_0 is

rejected. Conversely, if the observed test statistic is less than or equal to t_c, H_0 is not rejected. Alternatively, hypothesis tests can be framed in terms of *p*-values: the *p*-value will be smaller than α when the observed test statistic exceeds t_c, or equivalently, when the observed effect size exceeds the critical effect size, which corresponds to the minimal effect size that will reach statistical significance given a sample size, hypothesis test and α (Perugini et al., 2025). For example, suppose that a sports scientist recruits 30 participants per group to test the following hypothesis:

H: The new shoe model will improve running economy compared to the old shoe model during a 20-min time trial

This hypothesis leads to the formulation of two mutually exclusive statistical hypotheses:

H_0: Running_Economy_New_Show – Running_Economy_Old_Show ≤ 0 [the assertion of H]
H_1: Running_Economy_New_Show – Running_Economy_Old_Show > 0 [the negation of H]

In alignment with the directional hypothesis, the sports scientist conducts a one-sided *t*-test with an α of 5%. Given a sample size of 30 participants per group, the critical *t*-statistic for a one-sided *t*-test and α of 0.05 is 1.67, which corresponds to a critical effect size of Cohen's *d* of 0.43. Therefore, only effects sizes greater than or equal to $d = 0.43$ will result in a *p*-value less than or equal to α, leading to the rejection of H_0 in support of H. Observing an effect size $d \geq 0.43$ would allow the sports scientist to claim that "new shoes significantly improve running economy". Conversely, observing an effect size of $d < 0.43$, or $p > \alpha$, will lead to the non-rejection of H_0, and lead to the conclusion that "the data do not allow us to claim that new shoes significantly improve running economy".

One common criticism of null-hypothesis significance testing is that it is often performed poorly. For a scientific claim based on a hypothesis test to be valid, the tested hypothesis, the statistical test, and the scientific claim must be logically aligned. However, while coding the literature for meta-scientific projects (Mesquida et al., 2025), we noticed that this logical alignment is very often compromised in practice. When a misalignment between the stated hypothesis, the statistical test, and the scientific claim occurs, the scientific claim sport scientists make does not logically follow from the test they have performed. In previous research has highlighted related issues, including vague theoretical predictions (Frankenhuis et al., 2022), the difficulty of translating a theoretical prediction into a testable statistical hypothesis (Scheel et al., 2020), and the frequent absence of a clearly defined effect that researchers aim to establish or estimate (Kahan et al., 2024; Lundberg et al., 2021). In this paper we focus on a recurrent issue in sports and exercise science where researchers test hypotheses involving multiple interventions measured on multiple outcomes without clearly defining which test results would corroborate or falsify their prediction. This lack of clarity undermines the logic of hypothesis testing. As a consequence, claims in sports and exercise science are often not severely tested (Mayo & Spanos, 2006), in the sense that it is too easy for sport scientists to state that a claim is supported by data. For example, misalignments between hypotheses and statistical tests can lead to situations where multiple tests are used to test the same hypothesis, without correcting for multiple comparisons. The aim of this work is to start a scientific conversation among sports scientists about the importance of aligning the tested hypothesis, the statistical test, and the claim in scientific papers. Inspired by Hand's (1994)

seminal paper "Deconstructing Statistical Questions", we carefully analyze a series of hypotheses into basic statements to reveal common hypothesis misalignments in the published literature. Our examples come from published articles, but as our goal is to highlight general categories of misalignments, we do not cite the specific examples. Indeed, these examples are simply "one of many" and therefore do not deserve individual scrutiny. Throughout the article we will consistently use the following abbreviations: H refers to the hypothesis statement, H_0 and H_1 refer to the null and alternative hypothesis, respectively; EXP refers to the mean of the experimental condition or intervention group, while CON refers to the mean of the placebo or control group. The symbol α refers to the alpha level, typically set to 5%.

2. Misalignments

Misalignment 1: Testing a directional hypothesis using a two-sided statistical test

If sports scientists are testing a hypothesis that makes a directional prediction–such as caffeine will improve reaction time compared to a placebo–, then the statistical test must also reflect this directional nature. The first misalignment occurs when a directional hypothesis is tested using a two-sided statistical test, and after obtaining a $p < \alpha$, the scientist makes the directional claim that the experimental condition is superior to the control condition. H_0 is an effect of 0 (i.e., $\mu = 0$), while H_1 is any other effect (i.e., $\mu \neq 0$). Figure 2 illustrates a scenario where a two-sided independent *t*-test is performed with 50 participants per group, and a true effect size of Cohen's $d_s = 0.1$. We see that it is possible to observe a statistically significant effect in the negative direction (the blue area to the left of the vertical grey line at –1.99, which is the critical value in the negative direction). Logically, after rejecting a statistical test for any non-zero effect, it is not possible to claim an effect in a positive direction.

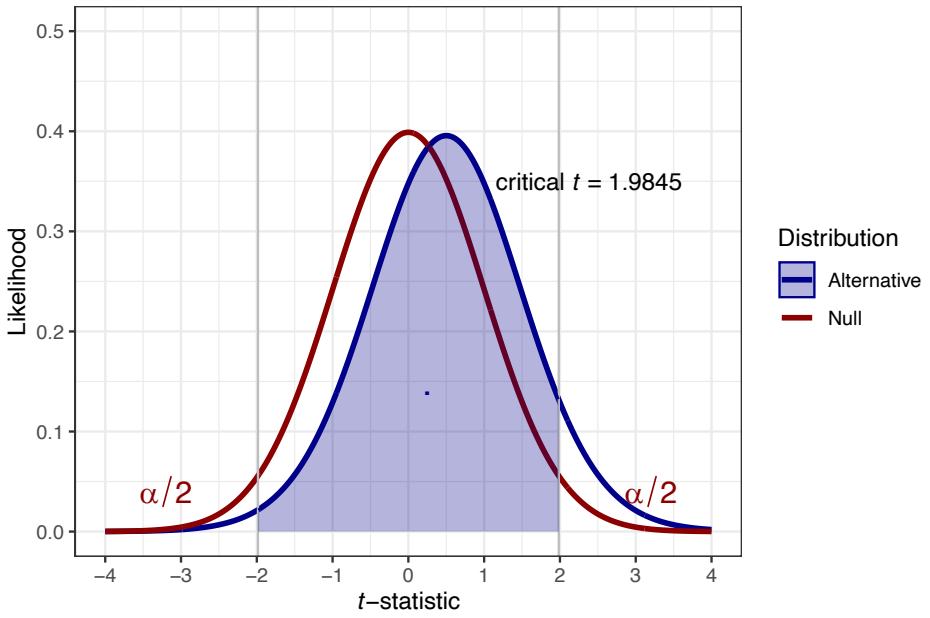

Figure 2. Illustration of the null (red curve) and alternative (blue curve) distributions of the *t*-statistic in a two-sided statistical test. The critical *t*-value (1.9845) indicates the threshold for statistical significance at the 5% significance level (α = 0.05). The areas labeled as *α*/2 under the null distribution correspond to the rejection regions, which jointly constitute the type I error rate (α = 0.05). The shaded blue area under the alternative

distribution curve represents the likelihood of observing data under the alternative hypothesis, corresponding to a true effect size of Cohen's $d_s = 0.1$.

It may seem obvious that making a directional claim after performing a two-sided statistical test is logically incoherent–after all, a directional hypothesis needs to be examined using a one-tailed statistical test–. This is precisely our point. Nevertheless, researchers commonly make directional claims based on two-sided tests, which can inadvertently lead them to accept their hypothesis but in the wrong direction. In a previous meta-scientific project where we selected the statistical result central to the tested hypothesis, we found that out of 350 studies, 6 (2%) studies claimed the existence of an effect in the opposite direction to what was expected, while still implicitly claiming support for the stated hypothesis.

The misalignment can be resolved in two ways. First, researchers can align their claim to the two-sided statistical test, and simply say that the null hypothesis is rejected, without making a scientific claim about the direction of the effect. This can still be followed by an estimate of the effect size. Second, researchers can align their test to the claim. There are two ways to do this. One approach is to perform a one-sided test at the α level if they are theoretically or practically only interested in an effect in one direction. A benefit of one-sided tests is that they have higher power to detect an effect size of interest compared to two-sided tests (Cho & Abe, 2013). For instance, given any effect size, and assuming a power of .8, a one-sided test requires 79% of the total sample size of a two-sided test. This is a benefit that should not be overlooked, especially in fields such as sports and exercise science where sample sizes are small, and studies might not have a high power to detect the effect of interest. The second approach is to perform two one-sided tests, each at α/2, one in the positive direction, and one in the negative direction (Kaiser, 1960). This test has the same type I error rate and the same power as a two-sided test, but it logically allows for a directional claim (unlike the two-sided test) without any additional cost. This solution might seem similar enough to the logically incoherent practice of making a directional claim after a two-sided test to wonder why one should worry about this misalignment. But if researchers are really interested in effects in a single direction, which they often seem to be, the optimal statistical test is a one-sided test, and awareness of this fact could increase the efficiency of statistical tests in sports and exercise science.

Misalignment 2: Testing a hypothesis of no difference with a classic null statistical test

Sports scientists often aim to test hypotheses that predict a difference between interventions, or that a new intervention is superior to a standard one. Less often, the hypothesis of interest predicts no difference (or in other words, equivalence) between two or more interventions. For example, a study might hypothesize that foam rolling and static stretching produce equivalent effects on range of motion. In a previous meta-scientific project where we selected the statistical result central to the tested hypothesis, we found that out of 350 studies, 36 (~10%) included a hypothesis of equivalence. This figure is likely an underestimation, as we only selected one hypothesis per study, whereas studies often test multiple hypotheses. Notably, all 36 studies (100%) used a classic statistical test to reject the null hypothesis of no difference. This leads to a misalignment in which scientists claim that two conditions are equivalent by failing to reject the null hypothesis of no difference (Aczel et al., 2018; S. L. Murphy et al., 2025). To illustrate the misalignment, consider the case of a sports scientist testing the following hypothesis:

H: there will be no difference in the decline of muscle force measured with dynamometry between cryotherapy (EXP) and napping (CON).

This H leads to the formulation of two statistical hypotheses:

H_0: EXP – CON = 0 [the assertion of H]
H_1: EXP – CON ≠ 0 [the negation of the assertion]

To test such hypothesis, 50 football players are randomly allocated to either the intervention group or to the control group after performing a high-intensity exercise bout. After collecting data on these 50 football players, a statistical test is conducted. When $p > \alpha$, the sports scientist should interpret this result as a failure to find sufficient evidence against the null hypothesis, rather than as evidence supporting the absence of an effect. In other words, they cannot confidently claim that there is no difference between the two interventions. Null hypothesis significance testing only allows us to reject the null hypothesis; if we fail to reject H_0, we cannot conclude that the null hypothesis is true or that it is supported–only that the evidence is insufficient to rule it out–. As H_1 encompasses all non-zero effects, such as effects of 0.01, it is practically impossible to detect all possible effects with low error rates. Therefore, we cannot argue there is no effect, because small effects remain possible, even after a non-significant test result.

To compound the issue, due to sampling error observed effects in a sample will almost always be non-zero even if the true population effect is 0, and there are no statistical tools that allow us to conclude an effect is exactly 0 (Frick, 1995). Even if a large number of participants is collected, the confidence interval around the observed effect size will converge to 0, but always contain a range of plausible non-zero effects (Figure 3) that cannot be statistically rejected.

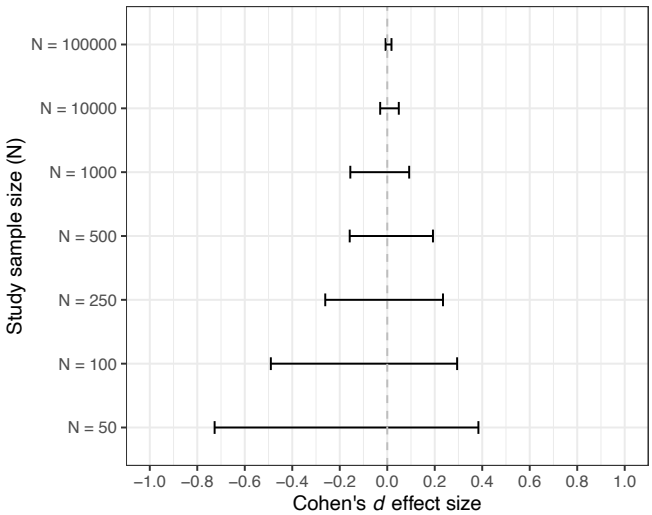

Figure 3. Illustration of how the width of the 95% confidence interval decreases as the sample size increases. The 95% confidence intervals were simulated assuming a true effect of 0. Although the 95% confidence interval becomes narrower with larger sample sizes, it continues to include both zero and small, non-zero effects. This

illustrates that, due to sampling error, it is statistically impossible to confirm that the true effect is zero, as the confidence interval always contains a range of plausible non-zero effects.

To resolve this misalignment, researchers can reformulate their research question in a way that it can be statistically answered. One solution is to test whether the null hypothesis of 'an effect' can be rejected in an equivalence test. To perform an equivalence test, researchers first need to specify which effects they consider large enough to be interesting. Researchers can specify a range of equivalence defined by the lower (Δ_L) and upper (Δ_U) bounds of the smallest effect size of interest (SESOI). Any observed difference falling within this range is deemed equivalent to a null effect, in the sense that even if the effect size is not numerically 0, it is too small to be considered meaningfully different from 0. The SESOI represents the smallest difference deemed practically relevant or theoretically important (Anvari & Lakens, 2021). An equivalence test reverses the pair of statistical hypotheses such that effects as large, or larger than, the smallest effect size of interest can be rejected:

H_{01}: EXP – CON $\leq \Delta_L$ and H_{02}: CON – EXP $\leq \Delta_U$
H_1: $\Delta_L <$ EXP – CON $< \Delta_U$

In other words, for H_0 to be rejected, differences should be more extreme than the upper and lower limits as determine by a SESOI, as opposed to simply different than 0. The structure of these statistical hypotheses is determined by the aim of the study–establish equivalence between two interventions by rejecting the presence of effects that are large enough to be deemed meaningful–. Thus, to support their hypothesis of equivalence, sports scientists must reject the null hypothesis of non-equivalence, which states that the respective difference between Cryotherapy and Napping is large enough to be meaningful. This null hypothesis is composed of two composite hypotheses: H_{01}: EXP – CON $\leq \Delta_L$ and H_{02}: EXP – CON $\leq \Delta_U$), which together define the boundaries of the equivalence region. To claim equivalence, the confidence interval of the observed difference must lie entirely within the pre-specified bounds between Δ_L and Δ_U, demonstrating that the observed difference is too small to be of practical or clinical importance. Determining these boundaries requires researchers define the SESOI. For example, when testing whether cryotherapy and napping produce equivalent effects on muscle force, the sport scientist might define the SESOI as ±30 Newtons. In this case, the equivalence region would be from –30 to +30, and statistical support for equivalence would require to reject H_{01}: EXP – CON $\leq$ –30 and H_{02}: EXP – CON $\leq$ +30. In other words, the entire confidence interval around the observed effect falls within the range of ±30 Newtons.

Specifying the SESOI is a complex process that depends on the research question and the intended aims of the study. One approach would be to use the anchor-based approach where one anchor functions as a reference to interpret the size of an effect (Anvari & Lakens, 2021). In this example, the change in muscle force could be anchored to an external indicator such as the rating-of-fatigue scale, assessing whether a change in muscle force corresponds to a meaningful reduction in perceived fatigue (Micklewright et al., 2017). Another approach would be a cost-benefit analysis, where the smallest effect size of interest is based on the minimal additional benefits cryotherapy would need to have compared to napping to make the investment of adopting cryotherapy worthwhile. Best practice would be prespecifying the SESOI in the preregistration or before data collection, and to design a test that has high power to reject the null hypothesis (e.g., the smallest effect size of interest). The narrower the

range of equivalence, or the smaller the effect size one tries to reject, the larger the sample size that is required. For tutorial papers on how to perform equivalence tests, readers are referred to Lakens (2017) and Mazzolari et al. (2022).

Misalignment 3: Failing to test whether the observed effect is of practical significance

Sports scientists sometimes claim that an effect is practically relevant based on the results of classic hypothesis test. For instance, consider a sports scientist testing the following hypothesis:

H: mean 20-min time-trial power output will be different after ingestion of ketones [EXP] compared to a control group [CON].

This H leads to the formulation of two statistical hypotheses:

H_0: EXP – CON = 0 [the negation of the assertion]
H_1: EXP – CON ≠ 0 [the assertion of H]

After performing a two-sided *t*-test, the sport scientist finds that the time-trial power output was 2.4% lower after EXP versus CON ingestion. The effect is statistically significant ($p < \alpha$). There are two possible misalignments. First, researchers might simply argue that the effect is practically significant, without having specified which effect sizes are large enough to matter (e.g., by specifying a smallest effect size of interest before data collection). In this case it is not possible to test the claim that the effect is practically relevant, as the comparison standard (which effect is of interest, and which effects are too small to be practically significant) is not specified. In essence, researchers are making an unsubstantiated claim. Second, the researcher might have specified a smallest effect size of interest (SESOI), but claim an effect is practically significant without testing against it. Let's assume a researcher has specified that an increase of 2% is practically meaningful, which is in absolute terms smaller than the observed 2.4% difference. However, the sports scientists cannot claim that the observed difference of 2.4% is of practical importance because this estimate comes with uncertainty (which can be quantified through a confidence interval around the effect size estimate), and a test is required to make the claim with a maximum error rate (for example by testing if the SESOI of 2% does not fall inside the 95% confidence interval of the mean difference). To make the claim that 2.4% is statistically different from 2%, the sport scientist should have explicitly tested against the null hypothesis of an effect as large or smaller than 2% in a minimum-effect test (K. R. Murphy & Myors, 1999). Following this example, the pair of statistical hypotheses should have included the SESOI (i.e., 2%), which could be stated as:

H_0: EXP – CON ≤ |2%| [the negation of the assertion]
H_1: EXP – CON > |2%| [the assertion of H]

After performing a minimum-effect test and observing a significant *p*-value (or the 95% confidence interval around the observed difference falls beyond the SESOI), the sport scientist can claim that the observed effect is

significantly different to the SESOI. For more information on minimal effect tests, see (Mazzolari et al., 2022; K. R. Murphy & Myors, 1999).

Misalignment 4: Omission of the test of an interaction

This misalignment occurs when sports scientists hypothesize an effect that requires the test of the interaction effect but do not perform or report the actual interaction test. This misalignment typically manifests when sports scientists (1) use a two-factorial design with pre-post repeated measures and (2) hypothesize a moderation effect.

Omission of the interaction effect in a two-factorial design with pre-post repeated measures

Sports scientists often hypothesize that one intervention improves performance more than another. To test this hypothesis, researchers design an experimental study where participants are assigned to two groups and the primary outcome is measured before and after the intervention for each group. The statistical hypothesis in this research design is to test whether there is a difference in the difference scores between groups, where H_0 states that the difference scores are the same, while H_1 is that the difference scores differ.

H_0: $(EXP_{POST} - EXP_{PRE}) - (CON_{POST} - CON_{PRE}) = 0$
H_1: $(EXP_{POST} - EXP_{PRE}) - (CON_{POST} - CON_{PRE}) \neq 0$

However, sports scientists often mistakenly claim that one intervention is superior to another intervention when the statistical test for the intervention's pre-post comparison is statistically significant (H_0: $EXP_{POST} - EXP_{PRE} = 0$), while the null hypothesis for the control's pre-post comparison is not rejected (H_0: $CON_{POST} - CON_{PRE} = 0$). As has been pointed out in the literature (Bland & Altman, 2011, 2015), the difference between significant and non-significant can itself be non-significant, meaning that this pattern of simple effects is not sufficient to claim the predicted interaction effect has been statistically supported. Consider the following hypothesis:

H: Omega-3 fatty acid supplementation will increase VO_{2peak} and improve running economy compared to a control group.

The study reported that there was no significant difference between groups in change in VO_{2peak} over the 12-wk intervention period ($p > 0.05$). However, a significant increase in VO_{2peak} from pre- to postintervention in intervention group was observed ($p < 0.05$) with no significant change in the control group $p > 0.05$). Based on these results, the study claimed that twelve weeks of omega-3 fatty acid supplementation during endurance training resulted in the improvement of running economy and increased VO_{2peak}.

Observing a significant difference between the pre- and post-measures for one intervention, but not for the other does not necessarily imply that these two interventions differ statistically from each other. For example, suppose the true difference between these two interventions is an effect size of 0, and the sports scientist conducts two paired *t*-tests for each intervention's pre-post comparison for which the critical effect size is $d = 0.32$. The effect size for the intervention's pre-post comparison is $d = 0.56$, $p = 0.002$, while the effect size for the control's pre-post interventions is $d = 0.22$, $p = 0.12$. Although one intervention shows a significant effect and the other does

not, the difference between these two effect sizes (0.56 and 0.22) is itself not statistically significant (p = 0.12; Figure 4).

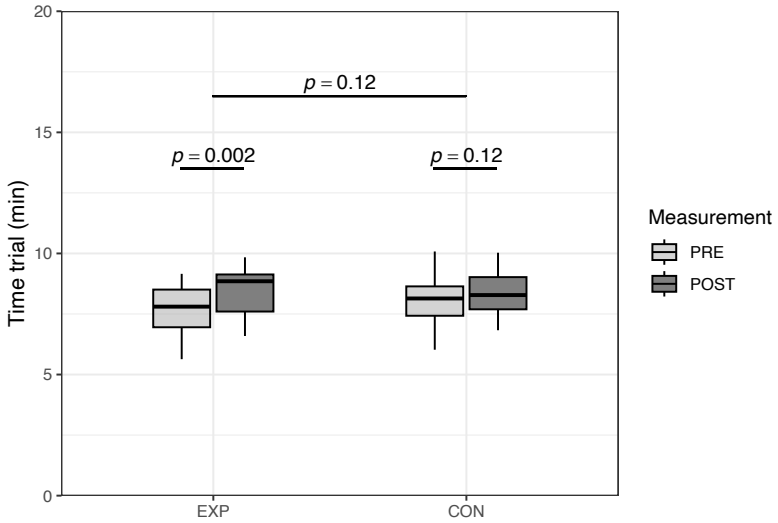

Figure 4. This example illustrates the case where a sports scientist designs a study with two between-subject groups with pre-post repeated measures and mistakenly claims that the new training intervention improved time trial performance. This error arises when the null hypothesis is rejected for the intervention's pre-post comparison (p = 0.002), but not for the control's pre-post comparison (p = 0.12), despite the difference between the two interventions being non-significant (p = 0.12).

To resolve this misalignment, the data provided by this design should be analyzed 1) with an analysis of covariance, 2) using the computed difference for each subject and then compare the groups in terms of score differences, or 3) testing the interaction effect using a two-way mixed analysis of variance (ANOVA) (Dimitrov & Rumrill, 2003; Huck & McLean, 1975). In other words, the aim is to reject the null hypothesis that the difference between the experimental and the control group is smaller or equal to zero. Although this null hypothesis should be rejected using the p-value from the interaction effect, sports scientists often do not report or overlook the interaction effect, and instead adopt two similar strategies involving the test of simple effects. The first is that some sports scientists perform two tests of simple effects where the goal is to reject the null hypothesis of no difference between the pre- and post-intervention scores for the experimental group (H_0: $EXP_{POST} - EXP_{PRE} = 0$) and fail to reject the null hypothesis of no difference between the pre- and post-intervention scores for the control group (H_0: $CON_{POST} - CON_{PRE} = 0$). Whenever H_0: $(EXP_{POST} - EXP_{PRE}) = 0$ is rejected, but not H_0: $(EXP_{POST} - EXP_{PRE}) = 0$, sports scientists implicitly claim that the intervention is different or superior to the control group. However, this claim is misleading, as no direct statistical comparison bas been made between the two groups.

A second strategy involves performing two tests of simple effects to reject the null hypothesis of no difference in post-intervention scores between the intervention and the control group (H_0: $EXP_{POST} - CON_{POST} = 0$) and fail to reject the null hypothesis of no difference in pre-intervention scores between the intervention and the control group (H_0: $EXP_{PRE} - CON_{PRE} = 0$). When H_0: $EXP_{POST} - CON_{POST} = 0$ is rejected, but H_0: $EXP_{PRE} - CON_{PRE} = 0$ is not rejected, sports scientists implicitly claim that interventions are statistically different because values were

different at post-test. However, when participants have not been randomly allocated to conditions this approach is also invalid. Without randomization and the inclusion of pre-treatment measurements in the statistical analysis, it is impossible to determine whether a significant difference between groups was caused by the treatment itself, or by pre-existing differences. Instead, sports scientists should directly compare the effect of the intervention against the control group by examining the interaction effect.

Omission of the interaction effect when testing for a moderation effect

Sports scientists are often interested in testing hypotheses that predict moderation effects. A moderation effect occurs when the (linear) effect of one factor on a primary outcome varies across the levels of a second factor (i.e., the moderator). For example, sports scientists may investigate whether the effect of an experimental condition on exercise performance is moderated by sex. In a 2 x 2 factorial design, this is tested statistically by evaluating the interaction between the two factors. H_0 states that the difference in the difference between experimental and control conditions is the same across sexes, while H_1 states that the difference scores vary by sex.

H_0: $(EXP_{FEMALE} - CON_{FEMALE}) - (EXP_{MALE} - CON_{MALE}) = 0$
H_1: $(EXP_{FEMALE} - CON_{FEMALE}) - (EXP_{MALE} - CON_{MALE}) \neq 0$

Although the interaction effect between sex and intervention is the effect of interest, it is often omitted or not reported (Garcia-Sifuentes & Maney, 2021). Consider the following hypothesis:

H: Beetroot juice supplementation will increase time to exhaustion more in males than in females.

The study reported that beetroot juice supplementation significantly increased time to exhaustion in males ($p < 0.05$) but not in females ($p > 0.05$). Based on these results, the study claimed that beetroot supplementation improved time to exhaustion in males but not in females.

Instead, researchers frequently rely on a significant main effect of sex or perform pairwise comparisons where the intervention and the control are compared within each sex. They then implicitly claim an effect exists when a significant and a nonsignificant *p*-value is observed for one sex but not the other. However, this approach does not test the hypothesis of interest. To resolve this misalignment, researchers should perform a test of the difference scores or interaction effect.

Misalignment 5: Incoherent multiple testing approaches

In Boolean logic, the connectors $\cap$ (interpreted as "and") and $\cup$ (interpreted as "or") determine whether a statement should be evaluated as a conjunction or disjunction set. For instance, consider a statement consisting of two elements (A and B), each of which is assigned a truth value (true or false). The truth value of the statement can be evaluated by each combination of the value of its elements with a truth table (Table 1). The $\cap$ connector determines that the statement ($A \cap B$) is tested as a conjunction set, hence the statement is evaluated as true (T) if and only if *both* A and B are true. On the other hand, the $\cup$ connector determines that the statement is tested as a disjunction set, hence the statement is evaluated as true when *either* A or B are true.

Table 1. Truth table illustrating the difference between a conjunction and disjunction set.

A	B	$A \cap B$	$A \cup B$
T	F	F	T
F	T	F	T
F	F	F	F
T	T	T	T

Boolean logic can be applied to hypothesis testing. A hypothesis can be regarded as a logical statement that researchers attempt to evaluate as corroborated (true) or falsified (false) by using a statistical test. When the effect of an intervention is tested on one primary outcome–what is known as an "individual testing" approach–, the type I error rate can be easily controlled at the prespecified α. However, a more common situation is when sports scientists test a hypothesis that will involve assessing multiple outcomes and/or compare multiple groups, often at several endpoints, resulting in a multiplicity of tests. In such cases, whether the type I error should be controlled at the prespecified α by adjusting for multiple comparisons depends on how the hypothesis is formulated, and specifically, whether it is formulated as conjunction or disjunction set (Dmitrienko & D'Agostino Sr, 2013; Rubin, 2021). A hypothesis formulated as a *conjunction* set implies that *all* hypothesis tests conducted to test the hypothesis must yield a $p < \alpha$ to claim that the hypothesis has been supported–what is known as an intersection-union testing approach–. This testing approach does not require researchers to adjust for multiple comparisons. Because all tests need to be statistically significant to claim the hypothesis is supported, the type I error rate is not inflated above the nominal α level.

In contrast, when a hypothesis is formulated as a *disjunction* set, it implies that multiple statistical tests are conducted, and the hypothesis is considered supported if *at least one* of these tests yields $p < \alpha$. This approach is known as union-intersection testing approach. However, in the context of a union-intersection testing approach, carrying out the multiple statistical tests without adjusting α level inflates the probability of incorrectly rejecting H_0, thereby increasing the risk of type I error. To address this problem, a correction of α for multiple comparisons is required (e.g., the Bonferroni correction).

Misalignment 5 arises when a hypothesis has been formulated as a conjunction or disjunction set, but the testing approach used does not align with the study's claim or the practical implications of the study (Cook & Farewell, 1996; Dmitrienko & D'Agostino Sr, 2013; Li et al., 2016; Molloy et al., 2022). Sports scientists should be aware of the distinction between formulating a hypothesis as a conjunction or disjunction set and consequently should adopt a testing approach that is aligned with their hypothesis.

Failing to adjust for multiple comparisons can inflate the type I error rate. Conversely, applying α adjustments can reduce statistical power, resulting in inflated type II errors (Nakagawa, 2004). In both cases, sports scientists risk making misleading claims about the effect of an intervention. Furthermore, if an unnecessary α adjustment is planned and incorporated into the sample size calculation, the study may require a larger sample size than necessary. Based on our experience, we are concerned that sports scientists may often be uncertain about which

testing approach to use, and in turn, when it is necessary to adjust α for multiple comparisons. As adjusting for multiple comparisons makes it less likely to provide statistical support for a hypothesis, this uncertainty can be used to opportunistically decide to not correct the alpha level, even when this should be done. When deciding about the testing approach, it can help to consider that within the Neyman-Pearson approach to NHST, (quasi)experimental studies are viewed as decision-making procedure where researchers make a claim about the effict of an intervention. Following this rationale, the key issue in determining which testing approach should be used is whether multiple hypothesis tests are performed to test the same hypothesis, and are therefore conceptually related (Cook & Farewell, 1996; Li et al., 2016; Molloy et al., 2022; Parker & Weir, 2020). Herein, we provide some examples of misalignment 5 and clarify when it is appropriate to use the union-intersection or the intersection-union testing approach.

When to apply the union-intersection approach

Sports scientists might often examine the effect of an intervention on related outcomes, without strong theoretical reasons, simply because it is easy to collect additional variables. Consider the following three hypotheses (H):

H: leg external pneumatic compression treatment compared to a static compression garment will improve performance following a muscle damaging protocol;

where performance is an umbrella outcome that was operationalized as isokinetic strength, countermovement jump, and squat jump.

H: cryotherapy will reduce force capacity, afferent feedback and neuromuscular propagation in comparison to a control group

H: a 30-minute nap after strenuous exercise will reduce the decline in muscle force;

where muscle force is measured at 5-, 60- and 120-minute post-nap, compared to a control group

What all these examples have in common is that they ignore that multiple statistical tests are used to a single hypothesis and make a claim based on any significant comparison and thus increase the risk of committing a type I error. Importantly, the increased risk of type I error depends on the performed number of statistical tests and the correlation between the primary outcomes or measurements (Stefan & Schönbrodt, 2023). That is, the higher the number of comparisons with lower correlations between primary outcomes/measurements, the more severe the inflation of the type I error rate.

To resolve this misalignment, sports scientists should adopt the union-intersection approach and perform adjustments. However, this approach can undermine researchers' ability to make a claim about the effect of an intervention when mixed results are obtained–where some outcomes are significant and others are not–(Cook & Farewell, 1996). For instance, consider the hypothesis that cryotherapy will reduce force capacity, afferent

feedback or neuromuscular propagation in comparison to a control group; would the sport scientist recommend the implementation of cryotherapy as a recovery method after only observing a significant difference in afferent feedback but not in force capacity and neuromuscular propagation? The best practice is to select one single primary outcome (at a predefined end point), based on theoretical reasons or established consensus about its importance, which accurately characterizes the effect of an intervention. In this case, there are no multiplicity issues because the sport scientist would be adopting an individual testing approach. Focusing on a single primary outcome also facilitates a more straightforward a priori power analysis. While additional variables may still be measured, they should be reported as secondary or exploratory outcomes (Ditroilo et al., 2025)

Another situation where sports scientists may overlook the issue of multiple comparisons is in studies involving multiple related interventions—for example, supplement studies with varying doses or intervention studies with different durations. In such studies, effectiveness is often claimed if any dose comparison reaches significance, but as this is a disjunction set, it requires α adjustment to avoid an inflated type I error. Alternatively, a closed testing approach can be used (Senn, 2008), where hypotheses are tested sequentially, and the order of testing is pre-specified before data analysis. For example, when comparing two doses and a placebo, it may be unreasonable to claim that a lower dose is effective unless a higher dose also shows superiority. In this approach, the higher dose is tested against the control first, and only if significant, is the lower dose compared to the placebo (Bachelez et al., 2015). This approach allows each comparison be tested while controlling the type I error at α. Alternatively, hypothesis of dose-response effect can be tested using a contrast test (Baguley, 2012; Stewart & Ruberg, 2000) or by modeling the dose-response effect (Senn, 2008), neither of which requires α adjustments.

When to apply the intersection-union approach

Sports scientists often frame their hypothesis as a conjunction set but overlook the fact that such hypothesis requires adopting the intersection-union testing approach. Consider the following hypothesis:

H: Wearing garments post-race will improve average power output, as measured by a 30-second Wingate test and 20-minute cycling time trial

Although the connector "and" denotes the adoption of the intersection-union approach, the study reports that the use of garments post-race significantly improved 30-second Wingate test but not 20-minute cycling time trial. Despite this, the study claims that the use of garments improves cycling performance. To resolve this misalignment, sports scientists should avoid making claims about the effect of an intervention based on a single significant comparison when adopting the intersection-union approach. In this framework, the hypothesis would only be supported if wearing garments post-race significantly improves both 30-second Wingate test and 20-minute cycling time trail. It cannot be one or the other, without controlling the type I error at α. Alternatively, they could adopt the union-intersection approach and perform α adjustments. However, in some cases, two or more outcomes might be crucial for evaluating the effectiveness of an intervention, and a significant effect on each outcome is required (European Medical Agency, 2017; Food Drug Administration, 2022). For example, in clinical investigations for the treatment of chronic obstructive pulmonary disease (COPD), lung function would be insufficient as a single primary outcome and needs to be accompanied by a symptom-based outcome or a patient-

related outcome (European Medical Agency, 2017; EMA/CHMP/44762/2017). In such cases, α adjustments are not necessary because the trial will only lead to the claim that the treatment should be recommended when *all* predefined outcomes reach statistical significance. Sports scientists should carefully consider whether multiple outcomes are required for evaluating the effectiveness of an intervention and choose their testing approach accordingly.

When to apply the individual testing approach

As we have mentioned, studies often include hypotheses that involve multiple unrelated interventions or outcomes. Consider the following additional hypotheses:

H: Beetroot juice supplementation will improve 20-min all-out power output *and* reduce perception of effort compared to a control group;
H: Creatine and beetroot juice (as separate interventions) will improve 3000-m running time trial performance in comparison to a control group;
H: Four weeks of high-altitude training will increase maximal oxygen uptake *and* enhance time trial performance at sea level in comparison to the control group.

Although these hypotheses include the connector "and", they contain multiple, distinct hypotheses that will lead to distinct claims, just as if they were evaluated in separate studies (Cook & Farewell, 1996; Molloy et al., 2022). In the first example, the two outcomes– power output and perception of effort– would support distinct claims. That is, the study may recommend beetroot juice for enhancing performance or for reducing perception of effort. Both claims are independently meaningful and could be evaluated and interpreted separately. In the second example, the focus is on determining whether each supplement is superior to a control. In such multi-intervention studies, the number of comparisons is not relevant because each intervention supports an independent claim (see Parker & Weir, 2020 for a thorough discussion). In the third example, investigating the physiological mechanism (i.e., changes in maximal oxygen uptake) is orthogonal to assessing the effect of altitude training on time trial performance. Even if high-altitude training enhances sea-level performance without improving VO_{2max}, the intervention would still be considered effective for improving endurance performance. Similarly, studies investigating several physiological mechanisms do not require α adjustments since each hypothesis test will lead to distinct claims. Therefore, studies that investigate unrelated interventions or assess unrelated primary outcomes should each result separately, as each may support a different claim (Cook & Farewell, 1996). Consequently, these should be presented as independent hypotheses rather than combining them into a single hypothesis statement. Importantly, researchers should explicitly report all individual tests they conduct regardless of significance, as selectively reporting (or even selectively highlighting only significant findings in an abstract) introduces bias, and will lead to an inflated type I error rate. Therefore, researchers should transparently communicate that they have reported all individual hypotheses, which can be achieved by preregistering their analysis plan before data collection (Lakens et al., 2024).

Improve conceptual clarity of the study

Evaluating whether the appropriate testing approach was used, and whether α adjustments were properly performed can be difficult because, in the published literature, hypotheses are often ambiguously stated (Mesquida

et al., 2023; Büttner et al., 2020). We define a hypothesis as ambiguous when the intervention effect—the specific comparison between the intervention and the control needed to test the hypothesis—is not defined and the primary outcome used to measure the intervention effect consists of an umbrella term or vague definition that is unmatched in the results section. Without a clearly specified statement, it becomes unclear what the hypothesis is truly predicting, making it easier for sports scientists to support their hypothesis. Consider the following example:

H: Napping will improve physical performance

Despite the simplicity of this hypothesis, the study design involves comparing four conditions: (1) normal sleep night and no nap, (2) sleep deprivation and no nap, (3) normal sleep and 20-min nap, and (4) sleep deprived and 90-min nap. The sport scientist can thus examine two effects: the effect of napping on physical performance after sleep deprivation or the effect of napping on physical performance after normal sleep. To further compound this ambiguity, the primary outcome (i.e., physical performance) is an umbrella outcome measured as reaction time, counter-movement jump height, and repeated sprint ability. Failing to clearly define the intervention effect and the primary outcome is particularly concerning in studies employing factorial designs or studies involving multiple interventions measured on several primary outcomes. When these two attributes are not clearly defined, researchers can conduct multiple hypotheses tests to test one single hypothesis, which may lead to inflated type I errors and misleading claims. Unfortunately, many studies published in sports and exercise fail to clearly define the intervention effect and the primary outcome (Mesquida et al., 2023; Büttner et al., 2020).

To address the issue of ambiguous hypotheses, the field could adopt the PICO framework (Schardt et al., 2007), which requires researchers to define four attributes of the intervention effect that the study aims to determine–namely target population, intervention, comparator and primary outcome, including at which end-point it will be assessed. More recently, the International Council for Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use proposed an extension of the PICO framework for clinical trials by including two additional attributes: intercurrent event, which defines strategies used to handle post-randomization occurrences such as treatment discontinuation, and the summary measure, which defines how outcomes are summarized (e.g., mean difference) (Kahan et al., 2024; Lawrance et al., 2020). Both the original and the extended frameworks aim to reduce ambiguity by requiring researchers to formulate the research question or hypothesis that can be directly linked to the intervention effect. The field of sport and exercise science would benefit from a similar structured framework that clearly defines the intervention effect by specifying the intervention and comparator, and the primary outcome. Best practice would be prespecifying the intervention effect with the four key attributes in the study preregistration. Sports scientists aiming to establish equivalence (Mazzolari et al., 2022) or whether two interventions differ by a certain margin (K. R. Murphy & Myors, 1999) would have to include the SESOI in the definition of their intervention effect.

The issue of ambiguous hypotheses is further compounded by ambiguous descriptions of how α adjustments are applied. For instance, a very common statement might read: *"A two-way ANOVA was used to investigate significant differences between groups and treatments. Significant main effects and interactions were further analyzed using the Bonferroni corrected post hoc test. All analyses used a significance level of $P < 0.05$."* Such

statements, coupled with an ambiguously formulated hypothesis, provide little insight into how α adjustments were applied since the number of hypothesis tests performed remains largely unknown.

To address this ambiguity, we propose—in addition to the adoption of the PICO framework—that researchers provide a preregistration of *all* statistical tests that lead to a scientific claim (as opposed to exploratory tests, that lead to a new hypothesis) where they clearly state how α adjustments are performed for all hypothesis tests. This information should also be reported in the article itself, and any deviations from the preregistration need to be highlighted and discussed. For example, the previously ambiguous hypothesis could be reformulated using the PICO framework as follows:

H: National-level cyclists (Population) taking a nap after sleep deprivation (Intervention) will show a significant greater mean increase in counter movement jump height or rating of perceived exertion during a 20-min time trial performed 30-min post-nap (Outcomes), compared to national-level cyclists who do not nap after sleep deprivation (Comparator).

The choice of the connector "and" or "or" in the hypothesis determines the adopted testing approach. Thus, sports scientists should ensure that the testing approach is aligned with its corresponding connector (i.e., the connector "and" should correspond to an intersection-union approach and "or" should correspond to the "union-intersection approach). Formulating the hypothesis following the PICO framework should be accompanied by a clear statement of the adopted testing approach and any performed α adjustments, if applicable. Such statement could read: *"α adjustments were applied based on a union-intersection testing approach, wherein the new intervention would be considered superior to the comparator group if it demonstrated a significant improvement in at least one of the two primary outcomes. To control the type I error, we applied a Bonferroni correction, dividing the* α level *of 0.05 by two. This resulted g in an adjusted α level of 0.025 for the main effect of each test".*

3. Discussion

The published literature in sports and exercise science is often characterized by claims based on a vaguely specified hypothesis, and statistical tests that are not aligned with scientific claims. Why do so many studies published in sports and exercise science seem to suffer from these two issues? We can identify three potential reasons. First, sport scientists often receive minimal training in statistics beyond basic techniques like *t*-tests and ANOVAs. As a result, hypothesis testing becomes more of a ritual than a carefully planned methodological process (Gigerenzer, 2004). It is therefore not surprising that many studies involve inappropriate statistical tests, leading to claims that are not logically aligned with the hypothesis test used. This issue could be mitigated through closer collaboration with applied statisticians, by strengthening statistical education and providing a more focused education pathway on the philosophy of science would greatly improve how sports scientists formulate and test their hypotheses.

Second, a growing concern in sports and exercise science is that many hypotheses in the published literature are inherently ambiguous (Mesquida et al., 2023; Büttner et al., 2020). The distinguishing feature of ambiguous hypotheses is that they can be tested in multiple ways, making the hypothesis easier to corroborate at the expense

of inflating type I error. Researchers can conduct multiple hypothesis tests as a strategy to obtain statistical significance. When the average power of studies in the field is 11% (Mesquida et al., 2025), one way to increase the probability of finding a significant effect is by conducting multiple hypothesis tests for a single hypothesis (Maxwell, 2004; Schmidt & Hunter, 2015). While the power of any specific test might be low, the probability of obtaining at least one significant effect increases with the number of hypothesis tests conducted. It is common practice to not define the intervention effect and use an umbrella outcome that can be measured or operationalized in several alternative allowing researchers to claim support for they tested hypothesis at the expense of inflated type I errors. This may explain why, despite an average power of 11% in sports and exercise science, nearly 70% of studies reported a significant result that supported the hypothesis tested (Mesquida et al., 2025). This is not a trivial issue. An inflated type I error rate in the published literature hinders the replicability of scientific findings. For example, the recent large-scale replication effort by the Sports Science Replication Centre (J. Murphy et al., 2025) reported a replication rate as low as 28%. Collectively, these concerns highlight the urgent need to improve research practices.

Researchers may also exploit the ambiguity of their hypotheses to engage in what Frankenhuis et al., (2022) term 'strategic ambiguity'. This ambiguity not only facilitates corroboration but also leave readers uncertain about what was actually hypothesized and what result(s) would corroborate or falsify the hypothesis. As a result, strategic ambiguity hinders scrutiny, making it difficult for peers to evaluate the intended intervention effect, assess the appropriateness of the study design and determine whether the study claim is justified. Scientific knowledge can progress when researchers make claims that can be scrutinized allowing peers to design experiments to replicate or falsify previous findings or test the same effect under different conditions. Thus, sports scientists should strive to state their hypothesis and the corresponding statistical tests with as much precision as possible (Lakens & DeBruine, 2021).

Third, in the early phases of research, researchers are not expected to have a strong understanding of the theoretical underpinnings of the phenomena under investigation. As a result, most hypotheses might be exploratory. The goal of exploratory research is to identify patterns, associations, and interactions between experimental conditions using statistical tests, without controlling error rates. In contrast, confirmatory research attempts to severely test pre-specified hypotheses using hypothesis tests and controlling error rates. In many fields, including sport and exercise science, exploratory research is not openly reported (Ditroilo et al., 2025), but represents a critical part of the research continuum (Scheel et al., 2020). Poorly defined intervention effects, especially when studies include multiple interventions and the measurement of several primary outcomes, imply many studies in which hypothesis tests are performed are more exploratory than confirmatory. Researchers might be able to control error rates be preregistering all tests they plan to perform, and lower the α level for each test (for example by using a Bonferroni correction). However, it might be more realistic to admit that it is too premature to test a hypothesis, and instead refrain from making any claims. In exploratory research, statistical tests function as a tool to generate hypotheses and not as a test of hypotheses.

4. Conclusion

NHST is a methodological procedure that allows sports scientists to make claims about the effect of interventions while controlling error rates. To be useful, NHST requires a clearly defined hypothesis and the use of an appropriate hypothesis test. However, NHST is often applied mindlessly, leading to situations in which stated hypotheses are not actually tested (Misalignments 1-3), or where type I and type II errors are inflated (Misalignments 4-5)—ultimately resulting in misleading claims. To address these issues, we recommend increased collaboration between sports scientists and applied statisticians and enhanced statistical training within the field. Additionally, we advocate for the adoption of preregistration and Registered Reports, and the adoption of the PICO framework, which provides a structured approach for defining of the intervention effect that a study aims to determine. By requiring researchers to specify the population, intervention, comparator and outcome, the PICO framework helps ensure clarity around the hypothesis being tested, enhancing transparency.

Competing interests

Authors declare no competing interests.

CRediT statement

Cristian Mesquida: conceptualization (lead); writing original draft (lead); Joe Warne: writing and editing (equal); Daniël Lakens: writing and editing (equal).

Data availability statement:

The data and analysis scripts related to this study are publicly available on the Open Science Framework and can be found at https://osf.io/axk5q/.

Funding

Cristian Mesquida was supported by the Ammodo Science Award 2023 for Social Sciences.

5. Reference

Aczel, B., Palfi, B., Szollosi, A., Kovacs, M., Szaszi, B., Szecsi, P., Zrubka, M., Gronau, Q. F., van den Bergh, D., & Wagenmakers, E.-J. (2018). Quantifying Support for the Null Hypothesis in Psychology: An Empirical Investigation. *Advances in Methods and Practices in Psychological Science*, *1*(3), 357–366. https://doi.org/10.1177/2515245918773742

Anvari, F., & Lakens, D. (2021). Using anchor-based methods to determine the smallest effect size of interest. *Journal of Experimental Social Psychology*, *96*, 104159. https://doi.org/10.1016/j.jesp.2021.104159

Bachelez, H., van de Kerkhof, P. C. M., Strohal, R., Kubanov, A., Valenzuela, F., Lee, J.-H., Yakusevich, V., Chimenti, S., Papacharalambous, J., Proulx, J., Gupta, P., Tan, H., Tawadrous, M., Valdez, H., Wolk, R., & OPT Compare Investigators. (2015). Tofacitinib versus etanercept or placebo in moderate-to-severe chronic plaque psoriasis: A phase 3 randomised non-inferiority trial. *Lancet (London, England)*, *386*(9993), 552–561. https://doi.org/10.1016/S0140-6736(14)62113-9

Baguley, T. S. (2012). *Serious stats: A guide to advanced statistics for the behavioral sciences* (pp. xxiii, 830). Palgrave Macmillan.

Bland, J. M., & Altman, D. G. (2011). Comparisons against baseline within randomised groups are often used and can be highly misleading. *Trials*, *12*, 264. https://doi.org/10.1186/1745-6215-12-264

Bland, J. M., & Altman, D. G. (2015). Best (but oft forgotten) practices: Testing for treatment effects in randomized trials by separate analyses of changes from baseline in each group is a misleading approach. *The American Journal of Clinical Nutrition*, *102*(5), 991–994. https://doi.org/10.3945/ajcn.115.119768

Büttner, F., Toomey, E., McClean, S., Roe, M., & Delahunt, E. (2020). Are questionable research practices facilitating new discoveries in sport and exercise medicine? The proportion of supported hypotheses is implausibly high. *British Journal of Sports Medicine*, *54*(22), 1365–1371. https://doi.org/10.1136/bjsports-2019-101863

Cho, H.-C., & Abe, S. (2013). Is two-tailed testing for directional research hypotheses tests legitimate? *Journal of Business Research*, *66*(9), 1261–1266. https://doi.org/10.1016/j.jbusres.2012.02.023

Cook, R. J., & Farewell, V. T. (1996). Multiplicity Considerations in the Design and Analysis of Clinical Trials. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, *159*(1), 93–110. https://doi.org/10.2307/2983471

Dimitrov, D. M., & Rumrill, P. D. (2003). Pretest-posttest designs and measurement of change. *Work (Reading, Mass.)*, *20*(2), 159–165.

Ditroilo, M., Mesquida, C., Abt, G., & Lakens, D. (2025). Exploratory research in sport and exercise science: Perceptions, challenges, and recommendations. *Journal of Sports Sciences*, 1–13. https://doi.org/10.1080/02640414.2025.2486871

Dmitrienko, A., & D'Agostino Sr, R. (2013). Traditional multiplicity adjustment methods in clinical trials. *Statistics in Medicine*, *32*(29), 5172–5218. https://doi.org/10.1002/sim.5990

European Medical Agency. (2017). *Guideline on multiplicity issues in clinical trials*. https://www.ema.europa.eu/en/documents/scientific-guideline/draft-guideline-multiplicity-issues-clinical-trials_en.pdf

Food Drug Administration. (2022). *Multiple Endpoints in Clinical Trials—Guidance for Industry*. https://www.fda.gov/media/162416/download

Frankenhuis, W., Panchanathan, K., & Smaldino, P. E. (2022). *Strategic ambiguity in the social sciences*. MetaArXiv. https://doi.org/10.31222/osf.io/kep5b

Frick, R. W. (1995). Accepting the null hypothesis. *Memory & Cognition*, *23*(1), 132–138. https://doi.org/10.3758/BF03210562

Frick, R. W. (1996). The appropriate use of null hypothesis testing. *Psychological Methods*, *1*(4), 379–390. https://doi.org/10.1037/1082-989X.1.4.379

Garcia-Sifuentes, Y., & Maney, D. L. (2021). Reporting and misreporting of sex differences in the biological sciences. *eLife*, *10*, e70817. https://doi.org/10.7554/eLife.70817

Gigerenzer, G. (2004). Mindless statistics. *The Journal of Socio-Economics*, *33*(5), 587–606. https://doi.org/10.1016/j.socec.2004.09.033

Hand, D. J. (1994). Deconstructing Statistical Questions. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, *157*(3), 317–356. https://doi.org/10.2307/2983526

Hempel, C. G. (1966). *Philosophy of natural Science* (pp. ix, 116). Prentice-Hall.

Huck, S. W., & McLean, R. A. (1975). Using a repeated measures ANOVA to analyze the data from a pretest-posttest design: A potentially confusing task. *Psychological Bulletin*, *82*(4), 511–518. https://doi.org/10.1037/h0076767

Kahan, B. C., Hindley, J., Edwards, M., Cro, S., & Morris, T. P. (2024). The estimands framework: A primer on the ICH E9(R1) addendum. *BMJ*, *384*, e076316. https://doi.org/10.1136/bmj-2023-076316

Kaiser, H. F. (1960). Directional statistical decisions. *Psychological Review*, *67*(3), 160–167. https://doi.org/10.1037/h0047595

Lakens, D. (2017). Equivalence Tests: A Practical Primer for t Tests, Correlations, and Meta-Analyses. *Social Psychological and Personality Science*. https://doi.org/10.1177/1948550617697177

Lakens, D., & DeBruine, L. M. (2021). Improving Transparency, Falsifiability, and Rigor by Making Hypothesis Tests Machine-Readable. *Advances in Methods and Practices in Psychological Science*, *4*(2), 2515245920970949. https://doi.org/10.1177/2515245920970949

Lakens, D., Mesquida, C., Rasti, S., & Ditroilo, M. (2024). The benefits of preregistration and Registered Reports. *Evidence-Based Toxicology*, *2*(1), 2376046. https://doi.org/10.1080/2833373X.2024.2376046

Lawrance, R., Degtyarev, E., Griffiths, P., Trask, P., Lau, H., D'Alessio, D., Griebsch, I., Wallenstein, G., Cocks, K., & Rufibach, K. (2020). What is an estimand & how does it relate to quantifying the effect of treatment on patient-reported quality of life outcomes in clinical trials? *Journal of Patient-Reported Outcomes*, *4*(1), 68. https://doi.org/10.1186/s41687-020-00218-5

Li, G., Taljaard, M., Van Den Heuvel, E. R., Levine, M. Ah., Cook, D. J., Wells, G. A., Devereaux, P. J., & Thabane, L. (2016). An introduction to multiplicity issues in clinical trials: The what, why, when and how. *International Journal of Epidemiology*, dyw320. https://doi.org/10.1093/ije/dyw320

Lundberg, I., Johnson, R., & Stewart, B. M. (2021). What Is Your Estimand? Defining the Target Quantity Connects Statistical Evidence to Theory. *American Sociological Review*, *86*(3), 532–565. https://doi.org/10.1177/00031224211004187

Maxwell, S. E. (2004). The Persistence of Underpowered Studies in Psychological Research: Causes, Consequences, and Remedies. *Psychological Methods*, *9*(2), 147–163. https://doi.org/10.1037/1082-989X.9.2.147

Mayo, D. G., & Spanos, A. (2006). Severe Testing as a Basic Concept in a Neyman-Pearson Philosophy of Induction. *The British Journal for the Philosophy of Science*, *57*(2), 323–357.

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*(3), 201–212. https://doi.org/10.1113/EP090171

Mesquida, C., Murphy, J., Warne, J., & Lakens, D. (2025). *On the replicability of sports and exercise science research: Assessing the prevalence of publication bias and studies with underpowered designs by a z-curve analysis*. SportRxiv. https://doi.org/10.51224/SRXIV.534

Micklewright, D., St Clair Gibson, A., Gladwell, V., & Al Salman, A. (2017). Development and Validity of the Rating-of-Fatigue Scale. *Sports Medicine*, *47*(11), 2375–2393. https://doi.org/10.1007/s40279-017-0711-5

Molloy, S. F., White, I. R., Nunn, A. J., Hayes, R., Wang, D., & Harrison, T. S. (2022). Multiplicity adjustments in parallel-group multi-arm trials sharing a control group: Clear guidance is needed. *Contemporary Clinical Trials*, *113*, 106656. https://doi.org/10.1016/j.cct.2021.106656

Murphy, J., Caldwell, A. R., Mesquida, C., Ladell, A. J. M., Encarnación-Martínez, A., Tual, A., Denys, A., Cameron, B., Van Hooren, B., Parr, B., DeLucia, B., Mason, B. R. J., Clark, B., Egan, B., Brown, C., Ade, C., Sforza, C., Taber, C. B., Kirk, C., … Warne, J. P. (2025). Estimating the Replicability of Sports and Exercise Science Research. *Sports Medicine*. https://doi.org/10.1007/s40279-025-02201-w

Murphy, K. R., & Myors, B. (1999). Testing the hypothesis that treatments have negligible effects: Minimum-effect tests in the general linear model. *Journal of Applied Psychology*, *84*(2), 234–248. https://doi.org/10.1037/0021-9010.84.2.234

Murphy, S. L., Merz, R., Reimann, L.-E., & Fernández, A. (2025). Nonsignificance misinterpreted as an effect's absence in psychology: Prevalence and temporal analyses. *Royal Society Open Science*, *12*(3), 242167. https://doi.org/10.1098/rsos.242167

Nakagawa, S. (2004). A farewell to Bonferroni: The problems of low statistical power and publication bias. *Behavioral Ecology*, *15*(6), 1044–1045. https://doi.org/10.1093/beheco/arh107

Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods*, *5*, 241–301. https://doi.org/10.1037/1082-989X.5.2.241

Parker, R. A., & Weir, C. J. (2020). Non-adjustment for multiple testing in multi-arm trials of distinct treatments: Rationale and justification. *Clinical Trials (London, England)*, *17*(5), 562–566. https://doi.org/10.1177/1740774520941419

Perugini, A., Toffalini, E., Gambarota, F., Lakens, D., Pastore, M., Finos, L., & Altoè, G. (2025). *The benefits of reporting critical effect size values*. OSF. https://doi.org/10.31234/osf.io/7qe92

Rubin, M. (2021). When to adjust alpha during multiple testing: A consideration of disjunction, conjunction, and individual testing. *Synthese*, *199*(3), 10969–11000.

Schardt, C., Adams, M. B., Owens, T., Keitz, S., & Fontelo, P. (2007). Utilization of the PICO framework to improve searching PubMed for clinical questions. *BMC Medical Informatics and Decision Making*, *7*(1), 16. https://doi.org/10.1186/1472-6947-7-16

Scheel, A. M., Tiokhin, L., Isager, P. M., & Lakens, D. (2020). Why Hypothesis Testers Should Spend Less Time Testing Hypotheses. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, 1745691620966795. https://doi.org/10.1177/1745691620966795

Schmidt, F. L., & Hunter, J. E. (2015). *Methods of Meta-Analysis: Correcting Error and Bias in Research Findings*. SAGE Publications, Ltd. https://doi.org/10.4135/9781483398105

Senn, S. S. (2008). *Statistical Issues in Drug Development*. John Wiley & Sons.

Stefan, A. M., & Schönbrodt, F. D. (2023). Big little lies: A compendium and simulation of p-hacking strategies. In *Royal Society Open Science* (Vol. 10, Issue 2, p. 220346). Royal Society. https://doi.org/10.1098/rsos.220346

Stewart, W. H., & Ruberg, S. J. (2000). Detecting dose response with contrasts. *Statistics in Medicine*, *19*(7), 913–921. https://doi.org/10.1002/(sici)1097-0258(20000415)19:7<913::aid-sim397>3.0.co;2-2

Tunç, D., Tunç, M. N., & Lakens, D. (2023). The epistemic and pragmatic function of dichotomous claims based on statistical hypothesis tests. In *Theory & Psychology* (Vol. 33, Issue 3, pp. 403–423). SAGE Publications Ltd. https://doi.org/10.1177/09593543231160112