Sports Medicine (2023) 53:1865–1890
https://doi.org/10.1007/s40279-023-01880-7

REVIEW ARTICLE

Extraordinary Claims in the Literature on High-Intensity Interval Training (HIIT): I. Bonafide Scientific Revolution or a Looming Crisis of Replication and Credibility?

Panteleimon Ekkekakis[1] · Paul Swinton[2] · Nicholas B. Tiller[3]

Accepted: 15 June 2023 / Published online: 10 August 2023

Abstract
The literature on high-intensity interval training (HIIT) contains claims that, if true, could revolutionize the science and practice of exercise. This critical analysis examines two varieties of claims: (i) HIIT is effective in improving various indices of fitness and health, and (ii) HIIT is as effective as more time-consuming moderate-intensity continuous exercise. Using data from two recent systematic reviews as working examples, we show that studies in both categories exhibit considerable weaknesses when judged through the prism of fundamental statistical principles. Predominantly, small-to-medium effects are investigated in severely underpowered studies, thus greatly increasing the risk of both type I and type II errors of statistical inference. Studies in the first category combine the volatility of estimates associated with small samples with numerous dependent variables analyzed without consideration of the inflation of the type I error rate. Studies in the second category inappropriately use the $p > 0.05$ criterion from small studies to support claims of 'similar' or 'comparable' effects. It is concluded that the situation in the HIIT literature is reminiscent of the research climate that led to the replication crisis in psychology. As in psychology, this could be an opportunity to reform statistical practices in exercise science.

1 Introduction

In the mid-1990s, exercise science underwent what can be characterized as the most consequential paradigmatic shift in its history, expanding its focus from exercise training for fitness enhancement to lifestyle physical activity for the promotion of public health [1, 2]. This new perspective resulted in a series of physical activity recommendations from organizations in the United States, including the Centers for Disease Control and Prevention [3], the Surgeon General [4], and the National Institutes of Health [5, 6], followed by similar initiatives in other countries. These recommendations converged on a common, easy-to-remember message: adults should accumulate (in short bouts, dispersed throughout the day) at least 30 min of physical activity, performed at least at a moderate intensity, on most, but preferably all, days of the week.

At the time, several aspects of these recommendations were criticized for their lack of specificity (e.g., what is 'moderate' intensity?) or for relying on a weak empirical basis (e.g., scant evidence on 'accumulated' physical activity). Furthermore, while the recommendations implied that additional health benefits could be obtained with activities of higher-than-moderate intensity, the emphasis was clearly placed on activity options that involve moderate intensity, such as brisk walking, based on the assumption that such options are realistic and non-intimidating for a largely hypoactive adult population [7]. This rationale was supported by a meta-analysis showing that interventions attempting to implement activity of higher intensity were associated with lower participation [8].

Despite good intentions, the guidelines had no measurable effect on public participation in physical activity. Accelerometry data from the 2003–2004 National Health and Nutritional Examination Survey (NHANES), a nationally representative study in the United States (with 6329 individuals providing at least one day of data), showed that only 3.5% of individuals 20–59 years of age and 2.4% of those

✉ Panteleimon Ekkekakis
ekkekaki@msu.edu

[1] Department of Kinesiology, Michigan State University, 308 W Circle Dr #134, East Lansing, MI 48824, USA

[2] School of Health Sciences, Robert Gordon University, Aberdeen, Scotland, UK

[3] The Lundquist Institute for Biomedical Innovation at Harbor-UCLA Medical Center, Torrance, CA, USA

aged 60 years or older registered at least 30 min of moderate-intensity physical activity per day on at least 5 days per week [9]. Less than 1% of adults registered 20 min of vigorous-intensity activity on at least 3 days per week [10]. In the 2005–2006 NHANES, the situation was unchanged, with only 3.2% of adults achieving the recommended dose of moderate-intensity activity [11]. The absence of positive results from population surveys encouraged calls for renewed emphasis on higher intensity activity [12–14]. Indeed, reformulated physical activity guidelines explicitly offered a choice between moderate intensity (for at least 30 min on 5 days per week, or 150 min per week), vigorous intensity (for at least 20–25 min on 3 days per week, or 75 min per week), or an equivalent combination [15, 16].

In 2005, in the midst of the debate preceding the reformulation of the guidelines and the renewed emphasis on vigorous-intensity activities, researchers published results from a doctoral dissertation [17] in the *Journal of Applied Physiology*. The article reported a remarkable finding, namely that a group of two women and six men doubled their cycling endurance performance (time to fatigue while pedaling at 80% $\dot{V}O_{2peak}$) after a total of only about 15 min of high-intensity interval training (HIIT) over 2 weeks, without changing their maximal aerobic capacity. An accompanying editorial [18] underscored the "effectiveness and remarkable time efficiency" of high-intensity training but noted that the 'price' participants have to pay is a need for "a high level of motivation" and "a feeling of severe fatigue lasting for at least 10–20 min" (p. 1983) [18]. Over the next several years, fueled by extensive media coverage in which HIIT was portrayed as a solution for individuals with limited available discretionary time, HIIT became a top trend in the fitness industry worldwide [19]. Moreover, since 2005, HIIT has been the subject of approximately 4000 articles, with more than 700 new articles being added to the literature each year, 10% of them being meta-analyses (see Fig. 1).

The data on the fitness and health benefits of HIIT have been characterized as "clear and convincing" (p. 1231) [20]. Nevertheless, as claims about HIIT are now influencing policy on a national and global scale (e.g., through exercise prescription guidelines and physical activity recommendations), it would be prudent to assess whether these claims can withstand statistical scrutiny. Steen [21] has argued that "error and fraud are the main sources of scientific misinformation" but "error is more prevalent than fraud" (p. 501). He insisted that "bias can also result from earnest error, statistical naiveté, or other innocent causes; not all bias is fraud" (p. 502). However, it has already been established that some of the extraordinary claims surrounding HIIT cannot be attributed solely to earnest human error. For example, on 14 February 2019, the *British Journal of Sports Medicine* issued a press release promoting the publication of a meta-analysis entitled "Is interval training the magic bullet for fat loss?"

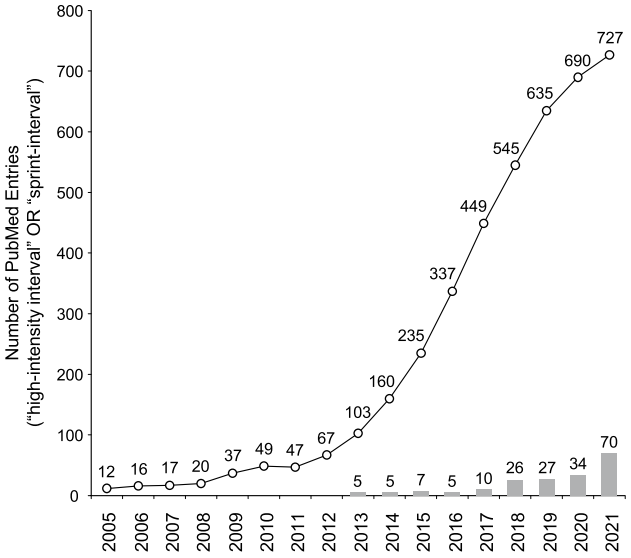

Fig. 1 The number of entries per year in PubMed that include the strings 'high intensity interval' or 'sprint interval' are shown in the line chart. The number of meta-analyses (subsample) is shown in bars

[22], which purportedly showed that, indeed, HIIT results in significantly larger reduction in total absolute fat mass than moderate-intensity continuous exercise (−2.28 kg, 95% CI −4.00 to −0.56, $p = 0.0094$). The press release issued by the journal appeared under the title "Interval training may shed more pounds than continuous moderate intensity workout," and attracted the attention of major news outlets, including the global news agency *Reuters* and influential magazines like *Runner's World*.[1] However, the meta-analysis was later retracted because the authors could not explain how they obtained their data (e.g., a larger reduction of body fat by −13.44 kg in HIIT than moderate-intensity continuous exercise, associated with a 12-week study that reported no relevant data).

Drawing lists of studies from two recently published systematic reviews, the present critical analysis focuses on statistical concerns emanating from the rapidly expanding literature on HIIT. This analysis highlights alarming parallels between prevalent practices in the HIIT literature and the emergence of a replication crisis in other scientific fields. The narrative culminates in a call for a return to fundamental principles of statistics. Unlike some of the more complicated scenarios outlined by Sainani et al. [23], the points raised in the following sections refer to elementary

[1] See: (1) https://bjsm.bmj.com/content/bjsports/suppl/2019/02/19/bjsports-2018-099928.DC1/bjsports-2018-099928.pdf; (2) https://www.reuters.com/article/us-health-exercise-training/interval-training-burns-off-more-pounds-than-jogging-or-cycling-idUSKCN1Q71TT; (3) https://www.runnersworld.com/news/a26339798/interval-training-for-weight-loss-study/

statistical principles, such as the mechanisms that raise the risk of type I and type II errors of statistical inference. The analysis culminates in a call not for the implementation of novel, obscure, or advanced statistical methods but rather for a *return* to fundamental statistical principles, along with the *readoption* of the critical outlook that should, in principle, characterize all manner of scientific inquiry.

2 Statistical Preliminaries: (Mis-) Understanding Null-Hypothesis Significance Testing

Studies evaluating the effectiveness of HIIT reach their conclusions following the statistical methodology known as null-hypothesis significance testing (NHST). Despite strong concerns [24, 25] and the presence of alternatives (i.e., Bayesian inference and fiducial inference) [26], NHST has been established as the standard method for evaluating statistical tests in most domains of human-science research, including the exercise sciences. Despite its popularity, however, the NHST is frequently misunderstood, misapplied, and misinterpreted [24, 25, 27].

NHST represents the amalgamation of the testing methodologies proposed during the period 1915–1933 by Ronald Aylmer Fisher (1890–1962) and the duo of Jerzy Neyman (1894–1981) and Egon Sharpe Pearson (1895–1980). Fisher on the one hand, and Neyman and Pearson on the other, contributed different pieces of what evolved into the NHST methodology, but it is important to emphasize that, as applied today, the NHST is "essentially an anonymous hybrid" and "a marriage of convenience that neither party would have condoned" (p. 171) [28].

Fisher, who emphasized the importance of inductive reasoning (i.e., analyzing samples to draw inferences about the population), is credited with the concept of the null hypothesis (i.e., data demonstrating random variance) and the use of exact *p* values as a quantitative measure of the 'extremeness' of the data given the null hypothesis. By extension, he considered *p* values as an indication of the plausibility or implausibility of the null hypothesis. However, although he famously wrote that "we shall not often be astray if we draw a conventional line at 0.05" (p. 82) [29], for Fisher, a low *p* value, such as $p < 0.05$, represented merely a sign that a finding may be worthy of further study, starting with an attempt at replication.

In the central point of contention with Fisher, Neyman and Pearson espoused a deductive approach, in which the null hypothesis is either rejected in favor of an alternative or retained for further study (which is not the same as accepting that the null hypothesis is true). Unlike Fisher, who believed that a specific hypothesis can be tested using data from a single study, Neyman and Pearson were not interested in developing a method for drawing inductive inferences about a single hypothesis based on the 'statistical significance' of data from a single study. Instead, their goal was to use a deductive approach and probability theory to develop 'rules of behavior' (i.e., rejection vs non-rejection of a hypothesis) to ensure that the frequency of errors (i.e., the erroneous rejection or non-rejection) would be kept below an acceptably low limit over a series of many studies:

> *But we may look at the purpose of tests from another view-point. Without hoping to know whether each separate hypothesis is true or false, we may search for rules to govern our behaviour with regard to them, in following which we insure that, in the long run of experience, we shall not be too often wrong. Here, for example, would be such a "rule of behaviour": to decide whether a hypothesis, H, of a given type be rejected or not, calculate a specified character, x, of the observed facts; if* $x > x_0$ *reject H, if* $x \leq x_0$ *accept H. Such a rule tells us nothing as to whether in a particular case H is true when* $x \leq x_0$ *or false when* $x > x_0$*. But it may often be proved that if we behave according to such a rule, then in the long run we shall reject H when it is true not more, say, than once in a hundred times, and in addition we may have evidence that we shall reject H sufficiently often when it is false* (p. 291) [30].

The Neyman-Pearson approach, therefore, implied two types of errors, called type I and type II, with the rate of those errors symbolized by the Greek letters α and β, respectively, as well as the concept of statistical power, symbolized as 1-β [31]. A type I error (α) occurs when "if we reject H_0, we may reject it when it is true," whereas a type II error (β) occurs when "if we accept H_0, we may be accepting it when it is false, that is to say, when really some alternative is true" (p. 296) [30]. Statistical power (1-β) is defined as "the probability of rejecting the hypothesis tested, H_0, when the true hypothesis is H_i" (p. 498) [32].

Fisher [33] concurred with the notion of type I errors and was keenly aware of the risk of raising the rate of such errors as a result of performing a multitude of tests. For example, he argued that a comparison between two extreme values "picked out from the results, will often appear to be significant, even from undifferentiated material" (p. 66). His proposed remedy was analogous to alpha-splitting, namely making the criterion for evaluating the *p* value more stringent: "We might, therefore, require the probability of the observed difference to be as small as 1 in 900, instead of 1 in 20, before attaching statistical significance to the contrast" (p. 66). On the other hand, arguing from an inductive standpoint, Fisher rejected the notion of type II errors because he believed that scientific research is a process of "learning by experience" and, in such a process, a priori knowledge is "almost always absent or negligible" (p.

73) [34]. Thus, although he considered the rate of type I error "calculable, and therefore controllable," he insisted that type II error is "incalculable both in frequency and in magnitude" (p. 73).

Interestingly, while Fisher rejected the notion of type II error, he was aware of the importance of statistical power (although he used the term 'sensitivity' or 'sensitiveness') and the role of sample size and a higher number of repetitions in increasing statistical power: "By increasing the size of the experiment, we can render it more sensitive, meaning by this that it will allow of the detection of a lower degree of sensory discrimination, or, in other words, of a quantitatively smaller departure from the null hypothesis" (p. 25) [33]. Commentators have noted that "Fisher's 'sensitivity' and Neyman-Pearson's 'power' refer to the same concept" (p. 173) [28], but Fisher "denied the possibility of assessing it quantitatively" (p. 1245) [35].

The main misinterpretations surrounding the NHST emerged following the merger of the Fisher and Neyman-Pearson approaches by anonymous researchers [35, 36], a merger "that neither party would have condoned," to repeat the phrase of Hubbard and Bayarri (p. 171) [28]. This anonymous and unsanctioned merger has resulted in several persistent misuses and misinterpretations that have plagued research for decades [24, 37, 38]. Of these, the following problems are arguably most relevant to research on HIIT.

2.1 The *p* Value as an Indication of the Plausibility of the Null Hypothesis

First, there is a widespread but mistaken belief that a p value of 0.05 means that there is only 5% probability of the null hypothesis being true (or, conversely, for 1-p, that there is 95% probability that the null hypothesis is false). This belief is mistaken because p values are calculated from the data under the assumption that the null hypothesis is true [39]. A p value merely indicates the probability (assuming that the null hypothesis is true) of observing a test statistic (e.g., a t value) as extreme or more extreme than the value observed in the present sample. This can be expressed as $\Pr(\text{data}|H_0)$ in probability notation. This statement is not equivalent to the interpretation that a p value of 0.05 means that there is only 5% probability of the null hypothesis being true, namely $\Pr(H_0|\text{data})$. While the p value does provide some indication of the plausibility or implausibility of the null hypothesis, a p near 0.05 "greatly overstates the evidence against the null hypothesis" (p. 139) [37]. Berger and Sellke [40] calculated that the lower bound of $\Pr(H_0|\text{data})$ can be estimated as:

$$\Pr(H_0|\text{data}) = \left(1 + (1+n)^{-1/2}\exp\left\{t^2/\left[2(1+1/n)\right]\right\}\right)^{-1}$$

Using a t value that yields $p = 0.05$ ($t = 1.96$) and a sample size of $n = 50$ per group results in $\Pr(H_0|\text{data}) = 0.52$, which surpasses $p = 0.05$ by more than an order of magnitude [40, 41].

2.2 The *p* Value as an Index of the Risk of Type I Errors

Second, related to the previous point, there is pervasive confusion between a p value, namely the probability of obtaining a test statistic at least as extreme as that obtained from a given study under the assumption that the null hypothesis is true, and α, namely the rate of type I errors [28]. In actuality, a single number (i.e., a p value) cannot simultaneously serve the dual function of providing an indication of the 'extremeness' of the data from any given study and, at the same time, an indication of the 'long-run' frequency of improperly rejecting the null hypothesis when it is true [39]. Nevertheless, statisticians [40–42] have estimated that, at least for the range $p < 1/e$, where e is Euler's constant (2.71828), namely $p < 0.36787$, the lower bound of α (i.e., the minimum risk of a type I error when rejecting the null hypothesis) can be estimated by:

$$\alpha(p) = \left(1 + \left[-e\,p\log(p)\right]^{-1}\right)^{-1}$$

where $\log(p)$ is the natural logarithm of the p value. Substituting $p = 0.05$ yields $\alpha = 0.289$. This means that there is at least 28.9% probability of a type I error when rejecting the null hypothesis on the basis of a p value close to 0.05. In other words, at least 28.9% of p values near 0.05 can be expected to come from studies in which the null hypothesis is true.

2.3 The *p* Value as an Index of Replicability

Third, researchers often mistakenly assume that a low p value (e.g., $p < 0.05$) entails that, if the same test were performed on a different sample randomly drawn from the same population (e.g., same sample sizes, same treatments), there would be high probability (e.g., > 95%) that the new p value would be similarly low (e.g., $p < 0.05$) [43]. In fact, except in studies with levels of statistical power over 90%, p values are characterized by extraordinary uncertainty [44, 45]. Thus, for a comparison between two means resulting in $p < 0.05$, the probability of finding $p < 0.05$ in a (theoretical) 'identical' replication (with the difference between the means being in the same direction) has been estimated as only 50% [46–49].

2.4 A Non-Significant *p* Value As a Basis for Accepting the Null Hypothesis

Fourth, a widely prevalent and persistent misunderstanding is that obtaining a nonsignificant test result (e.g., $p > 0.05$) can be interpreted as an indication that the null hypothesis (e.g., $\mu_1 - \mu_2 = 0$) is true or as indication of the absence of an effect [24, 37, 38, 50–52]. Fisher [33] famously asserted that "the null hypothesis is never proved or established, but is possibly disproved, in the course of experimentation" (p. 19). Accordingly, one of the oft-quoted admonitions of statisticians is that "the absence of evidence is not the same as evidence of absence" [53, 54]. A non-significant *p* value cannot provide a basis for accepting the null hypothesis as true or for the rejection of alternatives. It only suggests that a null effect is statistically consistent (or not inconsistent) with the data, along with the range of other effects encompassed within the confidence interval. However, $p > 0.05$ provides no indication that the null effect, specifically, is the most likely among these. Moreover, using non-significant *p* values as an indication in support of the null hypothesis is especially precarious in scientific fields, such as the exercise sciences [55], that are characterized by a preponderance of underpowered studies. Authors have warned that "null results are surprisingly easy to obtain by mere statistical artefacts; simply using a small sample or a noisy measure can suffice to produce a false negative" (p. 97) [56].

Collectively, the aforementioned misinterpretations suggest that NHST is a potentially useful, but delicate, test methodology. As such, it should be approached cautiously, recognizing and respecting its considerable limitations. The wide prevalence of the misinterpretations and misuses of the NHST across many domains of scientific research cannot be deemed a valid excuse for their ubiquity within the field of exercise science in general and research on HIIT in particular. Likewise, the fact that prestigious journals within the field of exercise science have permitted such practices does not render them any less egregious or harmful.

While there is ongoing debate about the causes and potential remedies of these misinterpretations and misuses of the NHST [57], many statistical experts see these misinterpretations and misuses as contributors to the phenomenon of non-replicable research [58–61]. Whether implemented deliberately or inadvertently, questionable statistical practices can result in intriguing, albeit fanciful, findings, with a high probability of attracting the attention of other researchers and the public. Serra-Garcia and Gneezy [62] speculated that, when evaluating manuscripts, journal editors and peer reviewers probably weigh two considerations against each other, namely the likely robustness or reliability of the result on one hand and its interest or curiosity on the other: "when the paper is more interesting, the review team may apply lower standards regarding its reproducibility" (p. 4).

3 Misuses of Null-Hypothesis Significance Testing in Research on HIIT

The following two sections present critical commentaries on two major variants of claims pertaining to HIIT, namely (i) that HIIT is effective in improving a variety of fitness and health outcomes, and (ii) that HIIT is as effective as more time-consuming moderate-intensity continuous exercise. We examine studies contained in two recent systematic reviews to demonstrate that deviating from elementary statistical principles can result in data that can be portrayed as supporting both of these conclusions, but with a high probability that such conclusions reflect errors of statistical inference. It is important to reiterate that the problems to be discussed are certainly not unique to the HIIT literature but have long plagued the broader exercise-science literature [63].

3.1 The 'Is Effective' Problem

As evidenced in meta-analyses [64, 65], a striking feature of the research literature on HIIT is an abundance of implausibly large effect sizes (e.g., standardized mean differences over 2.0 or 2.5 standard deviations) reportedly demonstrating the extraordinary effectiveness of HIIT compared with control conditions or even compared with active interventions consisting of moderate-intensity continuous exercise training. Some of these can be dismissed as mistakes, such as standardized mean differences (Hedges' g) of 11, 16, or 29 standard deviations [64], which can be readily attributed to computational errors (e.g., mistaking standard errors of the mean as standard deviations). Other cases, however, may be more complicated. For example, a remarkable standardized mean difference in maximal oxygen consumption of 4.59 standard deviations [65] from a 12-week comparison between HIIT and moderate-intensity continuous exercise [66] could be due to a host of well-established but frequently overlooked sources of methodological bias. These include, but are not limited to, the inadequate concealment of the randomization sequence, the absence of intention-to-treat analyses, and the use of unblinded outcome assessors. In addition, exercise researchers are aware of the biasing effect of several exercise-specific factors, such as the lack of control for verbal encouragement during tests of maximal performance [67–69]. When exercise testing is conducted by researchers who are ardent proponents of HIIT (e.g., "HIIT should play a central role in health activity guidelines" because it can "maximize the benefits of physical activity globally," p. 5216) [70], and are unblinded to treatment allocation, finding a standardized mean difference of 4.59 standard deviations in favor of HIIT becomes a plausible occurrence.

Such methodological sources of bias are beyond the scope of the present analysis. Here, we focus on statistical mechanisms that can produce similarly extraordinary (and likely non-replicable) results. For example, meta-analyses have reported that HIIT interventions have produced standardized mean differences that exceeded 2.5 standard deviations [71, 72]. Closer inspection of the characteristics of the studies that produced these large effect sizes [73–75] reveals certain notable commonalities: (i) small sample sizes (e.g., 10–20 participants per group), resulting in wide confidence intervals and low statistical power to detect even large effects, (ii) long lists of dependent variables, covering several multidimensional domains (e.g., anthropometric characteristics, inflammatory or immune markers, indices of cardiac, vascular, cardiorespiratory, or metabolic function), (iii) absence of pre-registration that could have allayed concerns about selective reporting, (iv) absence of designation of dependent variables as primary versus secondary, and (v) numerous statistical tests, each evaluated with the criterion of $p < 0.05$. Because of sampling variability and the lack of precision associated with small samples, estimates of population values (means, standard deviations) and, therefore, the associated p values "dance around" (p. 1720), as Gandevia [76] put it. Given a long enough list of dependent variables, it becomes almost inevitable that some means will happen to show exaggerated differences, thus resulting in extraordinarily large effect sizes. With a lax criterion such as $p < 0.05$, one or more comparisons will cross the threshold of 'statistical significance,' increasing the likelihood of publication. A cynic might argue that this approach could be used, deliberately or unwittingly, as a recipe for producing seemingly 'significant' and possibly novel or intriguing results, albeit results that are probably non-reproducible.

These basic statistical mechanics are explained in undergraduate and postgraduate university courses on research methodology. It is, therefore, surprising and disheartening that studies with the aforementioned characteristics, and attendant risk of producing untenable results, continue to be commonplace in large sections of exercise-science research [77], including research on HIIT.

Nosek et al. [57] criticized the "disciplinary incentives" that tend to "inflate the rate of false effects in published science" and "favor novelty over replication" (p. 615). In the following sections, we elaborate on several aspects of this problem.

3.1.1 Multiplicity

Methodologically strong studies, including most well-designed randomized controlled trials, have one outcome variable designated as 'primary' and, accordingly, test one main hypothesis, typically using the criterion of $p < 0.05$. Moreover, methodologically strong studies are pre-registered, which eliminates concerns about outcome switching (i.e., replacing the primary outcome of interest if it did not reach statistical significance with a different one that did) or selective reporting (i.e., only reporting the outcome that happened to reach the threshold of statistical significance out of a larger set of tested outcomes). However, in several domains of research, including studies investigating the effects of HIIT, pre-registration remains rare, and researchers report results pertaining to numerous dependent variables, each tested using the criterion of $p < 0.05$. This scenario is problematic insofar as it can raise the risk of type I errors (or 'false positives'), namely rejecting the null hypothesis when it is true.

Besides pre-registration, it is important for the tested hypotheses to be precise (e.g., "it is hypothesized that HIIT will improve outcome X as measured by test Y because of reason Z"). Instead, in the HIIT literature, studies often claim to have demonstrated the 'effectiveness' of HIIT relative to control treatments or relative to moderate-intensity continuous exercise (despite a smaller time commitment) by testing imprecise hypotheses that refer to broad concepts (e.g., cardiorespiratory fitness, endurance performance, muscle enzymes, blood pressure, glucose metabolism, inflammatory parameters, cardiometabolic health). In turn, each of these broad concepts is assessed by several variables (e.g., long lists of different indicators of cardiorespiratory fitness, endurance performance, muscle enzymes, and so on). If researchers explicitly follow a 'conjunction' approach [78], they need to reject all the constituent null hypotheses (e.g., one for each of the multiple inflammatory parameters) in order to claim that they rejected the joint null hypothesis (i.e., that HIIT has a stronger anti-inflammatory effect, in general, than moderate-intensity continuous exercise). The conjunction approach, because of the nature of the joint null hypothesis (i.e., all constituent tests must be significant), gives researchers only a single opportunity to reject the joint null hypothesis at the prespecified level of α (i.e., 5%) and, therefore, despite entailing multiple tests, it does not raise the overall risk of a type I error. On the other hand, the conjunction approach is characterized by low statistical power because researchers would fail to reject the joint null hypothesis if even one of the constituent tests yields a non-significant result. The low statistical power is the likely reason why the conjunction approach is rarely encountered in the research literature.

In contrast, in the 'disjunction' approach, it is only necessary to reject one of multiple constituent null hypotheses in order for researchers to be able to claim that they have rejected the joint null hypothesis [78]. For example, researchers may conclude that HIIT benefits muscle enzymes (or cardiometabolic health or arterial stiffness or cytokines) if only one or two of the variables that make up

this broad category, out of a larger set of tested variables, showed significant results in the expected direction. Consequently, the disjunction approach increases the risk of type I error because researchers have multiple opportunities to *incorrectly* reject the joint null hypothesis (i.e., each test of a constituent null hypothesis is also an opportunity to reject the joint null hypothesis).

For two independent events, the probability of observing both of these events together is given by the product of their (separate) probabilities. Therefore, if the probability of making a type I error is $\alpha = 0.05$, the probability of *not making* a type I error (i.e., erroneously rejecting the null hypothesis when it is true) on two independent simultaneous tests would be given by $(1-\alpha)\times(1-\alpha) = (1-\alpha)^2 = (1-0.05)^2 = 0.9025$. Conversely, the probability of *making* a type I error would be given by $1-(1-\alpha)^2 = 1-0.9025 = 0.0975$. Therefore, more broadly, the formula for the inflation of the type I error rate due to conducting multiple independent probability tests, often referred to as the Šidàk equation, is $\alpha^* = 1-(1-\alpha)^M$, where α^* is the inflated value of α as a result of conducting multiple independent tests, α is the conventionally defined probability of committing a type I error (typically, $\alpha = 0.05$), and M is the number of independent probability tests conducted at the level of α [79–81].

Applying this formula, one finds, for example, that conducting 14 independent tests following the disjunction approach results in $\alpha = 0.51$, namely > 10 times the nominal rate of 0.05. This means that, if 14 independent tests were to be conducted, one should expect the probability of making at least one type I error to be > 0.50. According to a statistical textbook: "It is especially important to realize that failing to control for multiple testing may play a major role in contributing to a disappointing failure rate in attempts to replicate published studies" (p. 216) [82].

As noted, the aforementioned formula relies on the simplifying assumption that the multiple probability tests are independent of each other. This assumption, however, is usually false in practice since, in a common example, several variables within the same data set may examine various facets of the same phenomenon (e.g., different parameters of glucose metabolism, immune function, or health-related quality of life), and will, therefore, probably be intercorrelated. To account for this dependence, researchers have proposed variations of the Šidák equation [83–86]. For example, an approach that originated in the field of genetics [87, 88] suggests that, when conducting 14 tests, instead of α rising to 0.51 when the tests are independent, α would rise to 0.48, 0.42, and 0.32 when the variables are intercorrelated $r = 0.30$, $r = 0.50$, and $r = 0.70$, respectively. Thus, while the formula $\alpha^* = 1-(1-\alpha)^M$ represents only the 'worst-case scenario,' it is nevertheless a useful reminder of the possible deleterious consequences of conducting multiple tests without consideration of the inflation of the type I error rate.

With pre-registration still being a rarity in exercise science [63], there is no guarantee that the dependent variables listed in an article represent a complete accounting of all the variables measured or analyzed. Even with this caveat in mind, it is common in the HIIT literature to encounter studies that follow the disjunction approach, hypothesizing joint null hypotheses, each consisting of numerous constituent tests, each tested at $p < 0.05$ [89–92]. This practice can increase the risk of type I error to high levels (see Fig. 2), even compared with other research within exercise science [55], thus raising serious concerns about the validity and reproducibility of any reported effects.

3.1.2 Sampling Variability and the Instability of *p* Values

To compound the problem of multiplicity described in the previous section, the samples used in the HIIT literature tend to be small (e.g., with as few as five individuals per group). The combination of long lists of dependent variables and small samples creates a statistical 'perfect storm,' a recipe for non-replicable science [43, 44, 46, 93]. Due to sampling variability, small samples produce highly volatile and imprecise estimates of the 'true' population values (e.g., means, standard deviations, intermean differences, and p values). The combination of instability and imprecision with an extremely lax criterion for determining 'statistical significance,' given a large enough number of tests, essentially guarantees two outcomes: (i) at least some of the tests will cross the liberal threshold of 'statistical significance' and (ii) these findings will have a high likelihood of being non-replicable in different samples.

The small samples have occasionally been justified on the basis of the argument that the studies are 'pilot' trials that were "not designed to be powered to detect statistically significant differences in small or moderate effects" (p. 2072) [94]. Instead, their purpose is portrayed as estimating "the magnitude of effect to lay the foundation for a fully powered efficacy trial" (p. 2072). It should be emphasized, however, that this rationale, although commonly encountered, is flawed, due to the inability of small-sample studies to accurately estimate population parameters [95, 96]. This lack of precision can lead to considerable over- or underestimations of the true effect size, with potentially devastating consequences for the design of subsequent larger trials.

As noted earlier (Sect. 2), although some researchers operate under the assumption that a finding of $p < 0.05$ entails 95% confidence that the same result would re-occur in a subsequent replication study, this is not the case. This misconception has been termed the 'replication fallacy' or 'replication delusion' [61]. In actuality, following an initial finding of $p < 0.05$, a subsequent (hypothetical) 'perfect' replication study drawing an equal number of participants from the same population has only about 50% chance of resulting

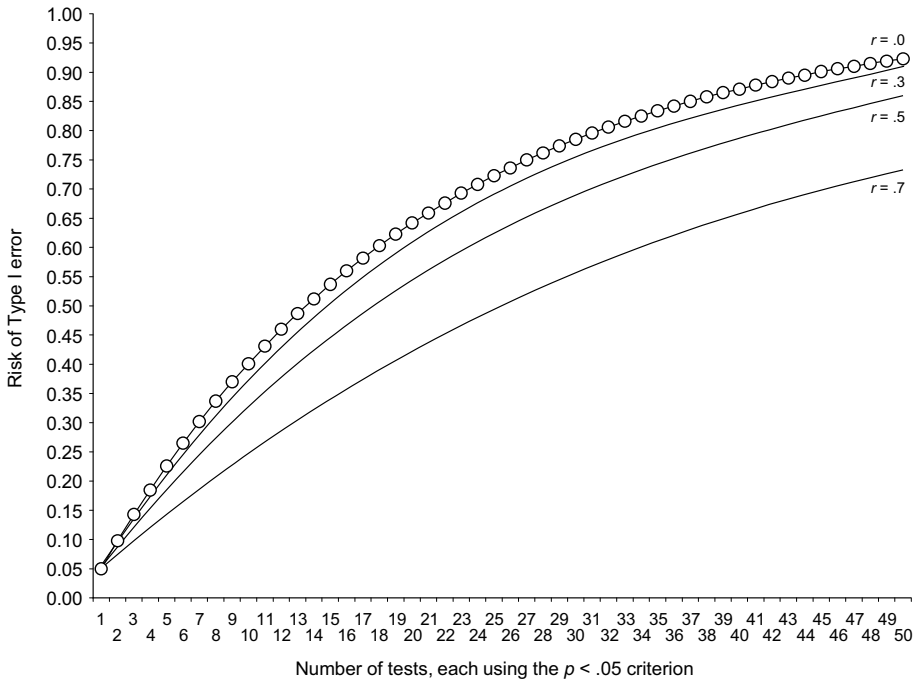

Fig. 2 The inflation of the risk of type I error as a function of the number of probability tests (at $p < 0.05$). The estimates shown include the theoretical case of statistically independent (uncorrelated) variables (using the Šidàk equation), as well as hypothetical cases in which the variables being analyzed are intercorrelated at levels of $r = 0.3$, $r = 0.5$, and $r = 0.7$ (using the M_{eff} method) [87, 88]

in a finding of $p < 0.05$ with the intergroup difference in the same direction [43]. Based on an empirical analysis of 45,955 observed effects derived from the Cochrane Database of Systematic Reviews, van Zwet and Goodman [97] put the estimate considerably lower, at 29%. Many researchers may find these figures surprising, despite numerous relevant warnings having been issued in applied literatures, including in psychology [46], physiology [76, 98, 99], medicine [93, 100], and pharmacology [101].

In an effort to understand the implications of p values for replication, statisticians have been analyzing the behavior of p values under various conditions, including different hypothetical population effect sizes, the level of α, and sample size [43, 102–105]. These efforts have resulted in formulas that enable researchers to calculate the probability of obtaining statistically significant results (e.g., $p < 0.05$) in subsequent replication studies [46]. One realization that has emerged from these investigations is that sampling variability renders p values extremely unstable and, therefore, an unreliable basis for drawing inferences about experimental effects in most applied-research contexts (given typical effect sizes and sample sizes), especially inferences regarding the replicability of findings [44, 46, 106].

To illustrate the implications for the HIIT literature, we examined the 48-study database used in a meta-analysis by Mattioni Maturana et al. [65], which concluded that HIIT "was superior to [moderate-intensity continuous training] in improving $\dot{V}O_{2max}$" (p. 559). In this meta-analysis, the median sample size was $N = 10$ per group, and the pooled effect size for $\dot{V}O_{2max}$ (i.e., the most extensively studied outcome) in comparison to moderate-intensity continuous training was $d = 0.40$. Assuming that the pooled effect size approximates the 'true' population effect size δ, the combination of these two numbers results in a noncentrality parameter $z = \delta\sqrt{(N/2)} = 0.40\sqrt{(10/2)} = 0.894$, which corresponds to an expected p value of 0.371 (the observed mean p value was slightly lower, at 0.323, for reasons that will be explained in Sect. 3.1.4).

Under these conditions ($N = 10$ per group, $\alpha = 0.05$, $\delta = 0.40$), statistical power (1-β) is only 0.14 (i.e., 14% of p values are expected to be below 0.05), much lower than the 0.80 conventionally considered adequate. As shown in Fig. 3, while 80% of the studies with 1-$\beta = 0.81$ will yield p values of 0.047 or less, 80% of the studies with 1-$\beta = 0.14$ will yield p values of 0.707 or less (which also means that 20% of studies will yield p values higher than 0.707). Indeed, 39 of the 48 p values (81.25%) associated with the studies in the meta-analysis by Mattioni Maturana et al. [65] were lower than 0.707, whereas 9 of 48 (18.75%) were larger than 0.707.

As a demonstration of the volatility of p values one can expect from this combination of effect sizes and sample sizes, Fig. 4 shows that the 48 p values related to $\dot{V}O_{2max}$ [65] covered the range from $p = 0.00000004$ to $p = 1.000$, and effect sizes exhibited an astounding range of 5.33 standard deviations, from −0.74 to +4.59. In other words, assuming that the effect size of the phenomenon under investigation is in the range between small and medium, attempting to study it with approximately 10 participants per group can lead to any outcome [46].

Moreover, as noted earlier and illustrated in Fig. 5 and Table 1, if an initial study yields $p < 0.05$, there is a

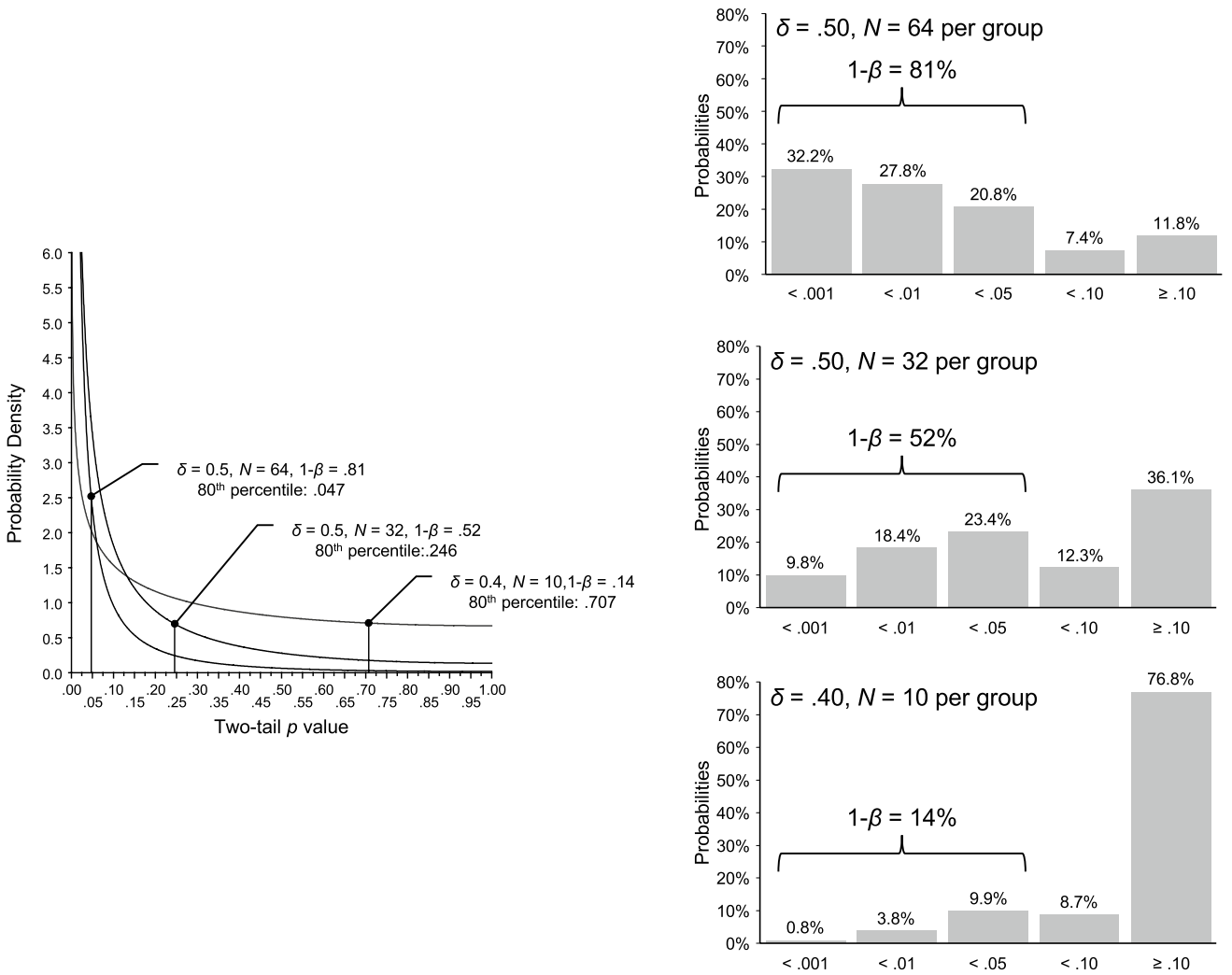

Fig. 3 The probability distribution of two-tailed p for three hypothetical studies: **i** an adequately powered study, with population effect size $\delta = 0.5$ and $N = 64$ per group ($1-\beta = 0.81$), **ii** the example shown by Cumming [46] (p. 289), with population effect size $\delta = 0.5$ and $N = 32$ per group ($1-\beta = 0.52$), and **iii** an example consistent with the studies included in the meta-analysis by Mattioni Maturana et al. [65], with population effect size $\delta = 0.4$ and $N = 10$ per group ($1-\beta = 0.14$). The 80th percentiles indicate that 80% of the area under each curve (the probability of two-tail p values) lies to the left of the marker and the figure indicated is the upper limit of the 80% percentile p interval (with a lower limit of zero). The probabilities associated with conventional intervals of p (i.e., 0.05, 0.01, 0.001) are shown as percentages in the histograms

50% chance that a subsequent replication will also yield $p < 0.05$, regardless of whether the population effect size is assumed to be 'known' or 'unknown.' However, if the initial study yields a p value of 0.371 (i.e., the p value expected from studies with the characteristics of those in the meta-analysis by Mattioni Maturana et al. [65]), the probability that a subsequent replication would yield $p < 0.05$ is only 14.6%. In other words, 85.4% of direct and exact replications (i.e., without any changes to research protocols, including sample size) would likely yield $p > 0.05$. Moreover, as noted by Cumming [46] and shown in Table 1, to have 90% confidence that a replication would yield $p < 0.05$, the initial study would have to produce $p < 0.00054$.

As shown in Table 2 and Fig. 6, the p intervals are extremely wide. The two-sided p interval, from the 10th to the 90th percentile, extends from 0.006 to 0.828, whereas the one-sided p interval from zero to the 80th percentile extends to 0.662. This means that 80% of replication two-tail p values would fall between 0.006 and 0.828 or between 0.000 and 0.662. Indeed, 85.42% of the two-tail p values associated with the studies in the meta-analysis by Mattioni Maturana et al. [65] were between 0.006 and 0.828, and 79.17% were between 0.000 and 0.662. For comparison (see Table 2), in a hypothetical literature in which one can expect

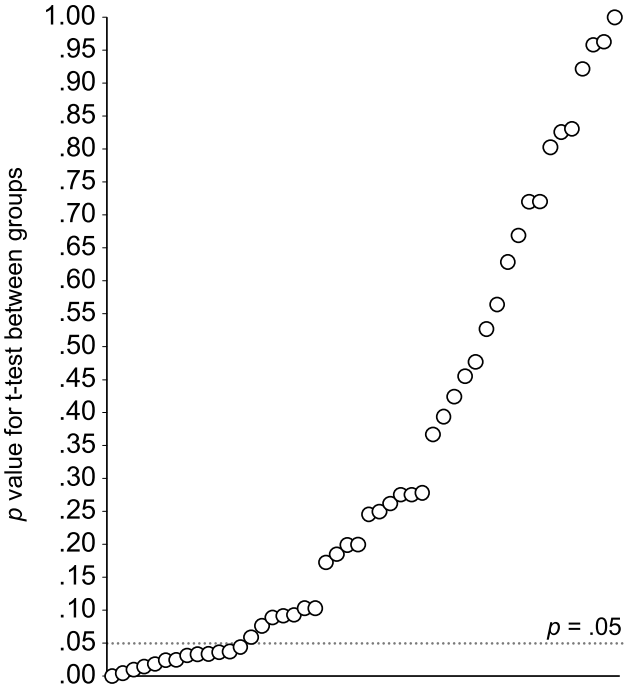

Fig. 4 The p values associated with the 48 studies comparing maximum oxygen consumption ($\dot{V}O_{2max}$) between high intensity interval training (HIIT) and moderate-intensity continuous exercise groups that were included in the meta-analysis by Mattioni Maturana et al. [65], illustrating the range from 0.000 to 1.000

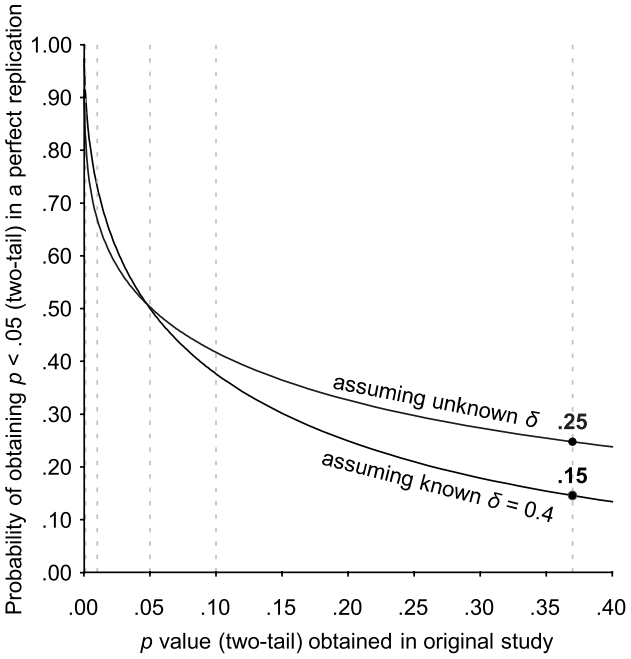

Fig. 5 Probability (y axis) that a hypothetical 'perfect' replication study (i.e., drawing samples of equal size from the same population as the original, and applying identical treatment and assessment methods) would obtain $p<0.05$, as a function of the p value obtained in the original study (under two assumptions: that the population effect size is known, and equal to the effect size obtained in the initial study, or not). It can be seen that if the initial study yielded $p<0.05$, there is only a 50% chance that a replication would also obtain $p<0.05$. If the initial study yielded $p=0.371$ (i.e., the p value expected from studies with the characteristics of those included in the meta-analysis by Mattioni Maturana et al. [65], given $\delta=0.40$ and $N=10$ per group), the probability of obtaining $p<0.05$ from a replication would be only 0.15 and 0.25, respectively

a study to yield $p=0.001$, the two-sided p interval for a replication study, from the 10th to the 90th percentile, extends from 0.0000005 to 0.139, whereas the one-sided p interval from zero to the 80th percentile extends to 0.036 (or to 0.018 in the case of a one-tail test).

3.1.3 Positive Predictive Value and False Positive Risk

Positive predictive value (PPV) is defined as the probability that a 'positive' research finding (e.g., $p<0.05$) represents a true effect (i.e., that the finding is a true positive). PPV can be estimated by the formula [107, 108]:

$$PPV = \frac{(1-\beta)R}{(1-\beta)R+\alpha}$$

where $1-\beta$ is statistical power, R indicates the prestudy odds (i.e., the odds that an effect is indeed non-null prior to the study being conducted, based on prior evidence), and α is the probability of a type I error. Although R is difficult to estimate, the highest value one can reasonably assume when there are no prior studies on a given topic is 50% (i.e., a 50–50 chance). Even in the unrealistic scenario of $R=0.50$, using the above formula shows, for example, that conducting 19, 23, 32, or 41 independent tests in underpowered studies (e.g., $1-\beta=0.14$) will result in only 7–10% probability of a true positive (see Fig. 7). Under the more realistic scenarios of 1-in-4 or 1-in-5 odds (i.e., $R=0.25$ or 0.20), the probability of a true positive drops to 3–5%.

As noted in the previous section, in the meta-analysis by Mattioni Maturana et al. [65], the median sample size was 10 per group (the mean was 13.2) and the pooled effect was $d=0.40$. As shown in Fig. 8, assuming that this effect size approximates the 'true' population effect (although this is likely an overestimate for reasons explained in Sect. 3.1.5), the median study exhibited only 14% statistical power (the mean of 16% was slightly higher due to one study with 75% power). This level of power is even lower than the median power of 21% highlighted as undermining the reliability of neuroscience [107]. Researchers have found that between 43 and 57% of studies in different domains of biomedicine have statistical power in the 0–20% range [109]. Of the 48 studies on $\dot{V}O_{2max}$ included in the Mattioni Maturana et al. [65] meta-analysis, considering the pooled effect size of $d=0.40$ as the effect size

Table 1 Probability of obtaining $p < 0.05$ from a replication as a function of the p value obtained in an initial experiment (p obt) under two assumptions (i.e., that the population effect size is known, and equal to the effect size obtained in the initial study, or not). The column labeled "Goodman" contains the values calculated by Goodman [43] (Table 1, p. 877), presented here as evidence of validation. The p value of 0.371 (i.e., the expected p value from the meta-analysis by Mattioni Maturana et al. [65], given $\delta = 0.40$ and $N = 10$ per group) is also included, to highlight the low probabilities of obtaining $p < 0.05$ from a replication study

p obt	Assuming δ is known ($\delta = d$)			Assuming δ is unknown		
	2-tail	Goodman	1-tail	2-tail	Goodman	1-tail
0.001	0.908	0.91	0.950	0.827	0.78	0.878
0.005	0.802	0.80	0.877	0.726	0.71	0.794
0.010	0.731	0.73	0.824	0.669	0.66	0.745
0.030	0.583	0.58	0.700	0.561	0.56	0.645
0.050	0.500	0.50	0.624	0.503	0.50	0.588
0.100	0.376	0.37	0.500	0.417	0.41	0.500
0.200	0.249		0.358	0.327		0.399
0.371	**0.146**		**0.227**	**0.247**		**0.298**
0.400	0.134		0.211	0.238		0.285
0.600	0.082		0.131	0.195		0.214

Table 2 Two-sided (extending from the 10th to the 90th percentile) and one-sided (extending from zero to the 80th percentile) p intervals for two- and one-tail single-study replications as a function of the p value obtained in an initial (two-tail) study (p obt). P intervals indicate the probability of obtaining $p < 0.05$ in a single, identical replication study. Compare to the values calculated by Cumming [46] (Table 1, p. 292) for validation. As noted by Cumming [46], "for the 90% p interval [one-tail] to be [0, 0.05], p obt must equal 0.00054" (p. 293). The p value of 0.371 (i.e., the expected p value from the studies included in the meta-analysis by Mattioni Maturana et al. [65], given $\delta = 0.40$ and $N = 10$ per group) is also included, to highlight the extraordinarily wide p interval associated with it

p obt	10–90th percentile interval, two-tail	10–90th percentile interval, one-tail	0–80th percentile interval, two-tail	0–80th percentile interval, one-tail
0.00054	[0.0000005, 0.099]	[0.0000001, 0.050]	[0.000, 0.023]	[0.000, 0.011]
0.001	[0.0000005, 0.139]	[0.0000005, 0.070]	[0.000, 0.036]	[0.000, 0.018]
0.010	[0.000012, 0.408]	[0.000006, 0.223]	[0.000, 0.162]	[0.000, 0.083]
0.020	[0.000035, 0.517]	[0.000018, 0.304]	[0.000, 0.242]	[0.000, 0.128]
0.050	[0.000162, 0.648]	[0.000081, 0.441]	[0.000, 0.379]	[0.000, 0.221]
0.100	[0.000544, 0.728]	[0.000273, 0.567]	[0.000, 0.491]	[0.000, 0.325]
0.200	[0.001924, 0.789]	[0.000988, 0.702]	[0.000, 0.591]	[0.000, 0.464]
0.371	**[0.005998, 0.828]**	**[0.003397, 0.821]**	**[0.000, 0.662]**	**[0.000, 0.616]**
0.400	[0.006848, 0.832]	[0.003978, 0.834]	[0.000, 0.669]	[0.000, 0.636]
0.600	[0.013091, 0.849]	[0.009726, 0.901]	[0.000, 0.701]	[0.000, 0.747]

of interest, 42 (88%) had statistical power in the 0–20% range and all but one (47 of 48, or 98%) were in the 0–33% range. The combination of the type I error rate (α) being allowed to escalate and the extraordinarily small (i.e., severely underpowered) studies can easily (i.e., in common, entirely realistic scenarios) lead to false discovery rates that approach 100%.

A complementary way to think of this problem is in terms of the false positive risk (FPR), namely the probability that a 'significant' result (e.g., $p < 0.05$) represents a false positive. The FPR can be estimated by the formula [60]:

$$FPR = \frac{p(1-R)}{p(1-R) + (1-\beta)R}$$

where p is the p value of a study, R indicates the prestudy odds (i.e., the odds that an effect is indeed non-null prior to the study being conducted, based on prior evidence), and $1-\beta$ is the statistical power of the study. The FPR is related to efforts [40–42], reviewed in Sect. 2, to associate the p value from a single study to the lower bound of the long-run risk of type I error (α). Applying the formula to the studies on $\dot{V}O_{2max}$ that were included in the Mattioni Maturana et al. [65] meta-analysis, and assuming that $R = 0.50$, shows that only three

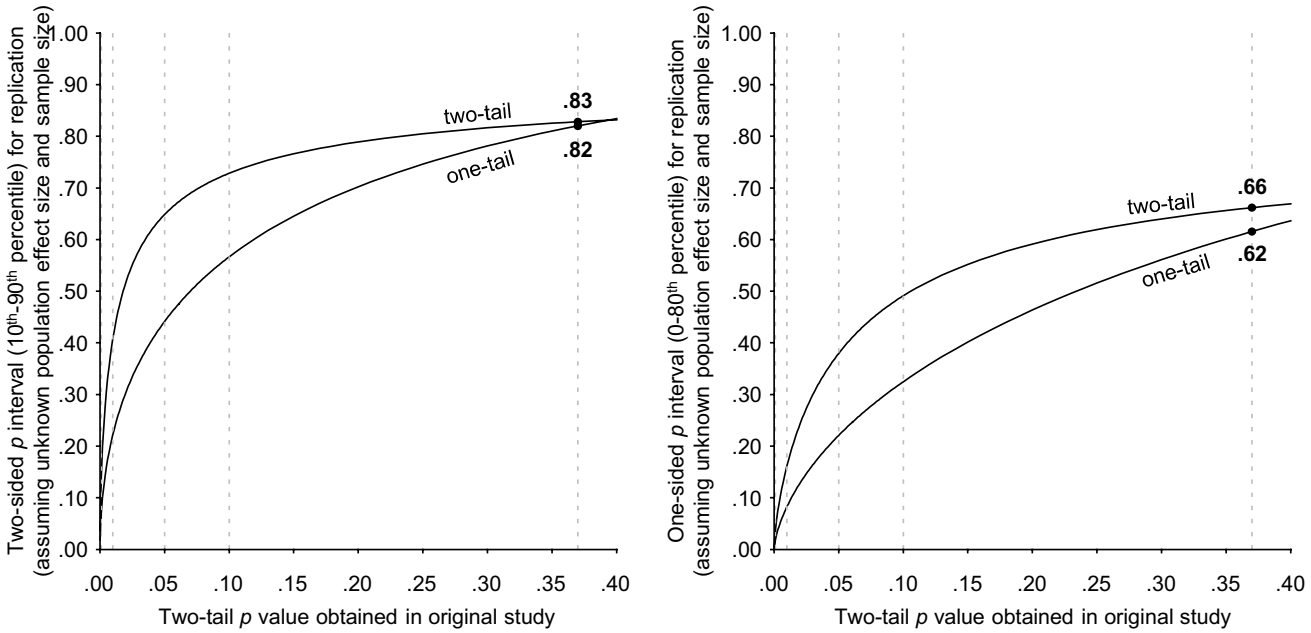

Fig. 6 p intervals estimated to indicate the probability of obtaining $p<0.05$ in a replication study as a function of the (two-tail) p value in an initial study. The two-sided p intervals, extending from the 10th to the 90th percentile, are shown on the left, whereas the one-sided p intervals, extending from zero to the 80th percentile, are shown on the right. Estimates are shown for both two-tail and one-tail tests in the replication study. The upper limits of the 90th percentile (left) and 80th percentile (right) p intervals associated with an initial study yielding $p=0.371$ (i.e., the p value expected from studies with the characteristics of those included in the meta-analysis by Mattioni Maturana et al. [65], given $\delta=0.40$ and $N=10$ per group) are highlighted

of the 48 studies produced FPR lower than 0.05 (see Fig. 9). Given their low level of statistical power (median 0.155, mean 0.169), even under the unrealistic assumption of $R=0.50$, the FPR of the 13 studies that produced $p<0.05$ was as high as 0.245, with a mean of 0.130 and a median of 0.123 (recall that the risk of type I error associated with $p=0.05$ has been estimated as at least 0.289).

3.1.4 Excess of 'Significant' Results

Assuming that the null hypothesis is false (e.g., that there is a difference between HIIT and moderate-intensity continuous training in terms of improving $\dot{V}O_{2max}$), and the effect size is $\delta=0.40$, samples of 10 per group are expected to reject the false null hypothesis in only 14% of the cases (i.e., statistical power of 14%). Instead, as shown in Fig. 10, 13 of the 48 studies (27.1%) included in the meta-analysis by Mattioni Maturana et al. [65], nearly double the expected rate, produced results with $p<0.05$.

This rate indicates an 'excess of significant findings' according to the test proposed by Ioannidis and Trikalinos [110]. This is a χ^2 statistic calculated as:

$$A = \left[(O-E)^2/E + (O-E)^2/(n-E)\right]$$

where O is the number of studies reporting 'statistically significant' results ($p<0.05$), E is the sum of the levels of statistical power in all the studies in the sample to detect the population effect size (assumed here to equal the pooled effect size from the meta-analysis, namely $d=0.40$), and n is the number of studies in the sample. For the studies in the meta-analysis by Mattioni Maturana et al. [65], E is 7.851, $O=13$, and $n=48$. Therefore, $\chi^2(1)=4.038$, $p=0.044$, indicating the presence of an excessive proportion of 'statistically significant' results.

Various mechanisms may account for this phenomenon [111]. One category includes 'researcher degrees of freedom' [112], some of which may be questionable (e.g., 'p-hacking,' selective outcome reporting, selective removal of data points, failing to account for multiplicity) and some of which may reflect publication bias (e.g., the 'file drawer' problem, namely the low probability of studies reporting non-significant results being accepted for publication) [113].

3.1.5 'Winner's Curse'

An additional problem, named 'winner's curse' [114, 115], emerges from underpowered studies. The 'winner's curse' refers to the fact that, when an underpowered study happens to correctly reject a null hypothesis, the estimate of

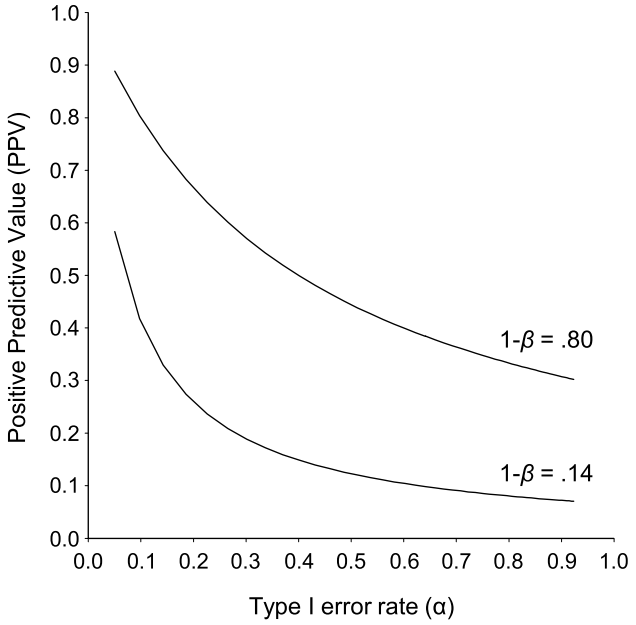

Fig. 7 Positive predictive value (PPV), namely the probability that a 'positive' research finding represents a true effect (i.e., that the finding is a true positive), as a function of the type I error rate (α), when statistical power $(1-\beta)$ is sufficient (i.e., $1-\beta=0.80$) and when it is the median of the power of studies included in the meta-analysis by Mattioni Maturana et al. [65] comparing high intensity interval training (HIIT) and moderate-intensity continuous training on maximum oxygen consumption ($\dot{V}O_{2max}$) (i.e., $1-\beta=0.14$). When α is allowed to escalate to high levels, even under the unrealistic scenario of $R=0.50$, the PPV drops to < 0.10

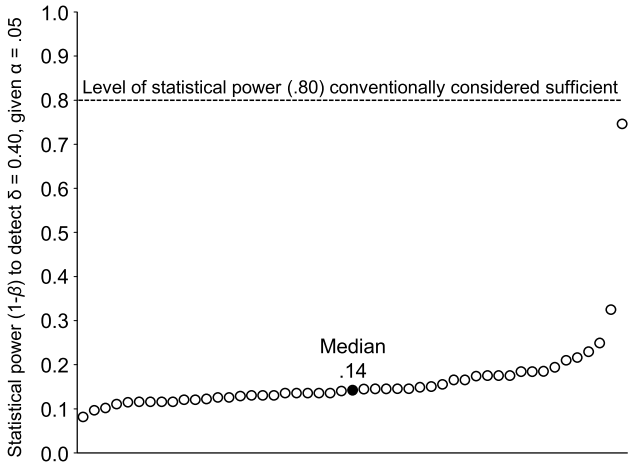

Fig. 8 Levels of statistical power $(1-\beta)$ for each of the 48 studies included in the Mattioni Maturana et al. [65] meta-analysis comparing the effects of high intensity interval training (HIIT) and moderate-intensity continuous exercise on maximum oxygen consumption ($\dot{V}O_{2max}$). Power was calculated from the reported sample sizes, assuming that the pooled effect ($d=0.40$) represents the 'true' population effect and $\alpha=0.05$. The median study exhibited 14% statistical power, 42 of 48 studies (88%) had statistical power in the 0–20% range and all but one (47 of 48, or 98%) were in the 0–33% range

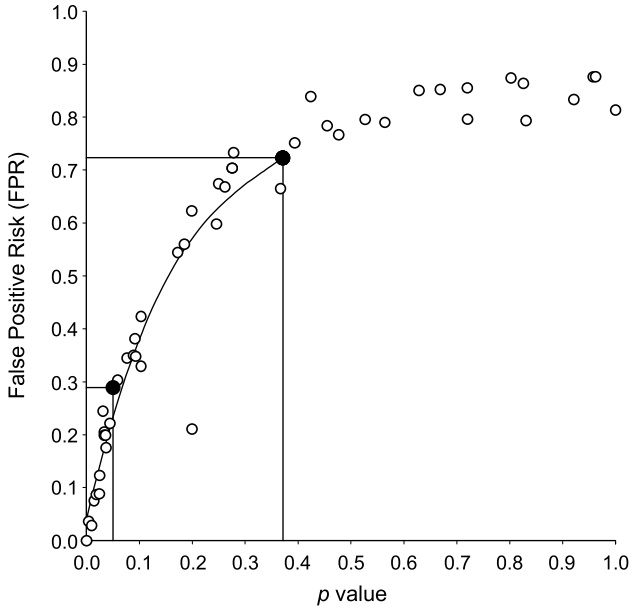

Fig. 9 The estimated false-positive risk (FPR) of the studies on maximum oxygen consumption ($\dot{V}O_{2max}$) that were included in the Mattioni Maturana et al. [65] meta-analysis, assuming $R=0.50$. Only 3 of the 48 studies (6.25%) produced FPR < 0.05. The FPR of the 13 studies that produced $p<0.05$ was as high as 0.245, with a mean of 0.130 and a median of 0.123. Two related figures are highlighted for reference: **i** the minimum risk of type I error (α) associated with $p=0.05$ has been estimated as 0.289; **ii** the relationship between p values and α holds until $p<1/e$, namely $p<0.368$, after which α reaches a plateau

the magnitude of the effect derived from such a study will likely be exaggerated. This is because, for a result to satisfy the criterion of statistical significance (even the uncorrected $p<0.05$) in an underpowered study, the effect will have to be unusually large. Young et al. [115] described the problem as follows:

> *The average result from multiple studies yields a reasonable estimate of a "true" relationship. However, the more extreme, spectacular results (the largest treatment effects, the strongest associations, or the most unusually novel and exciting biological stories) may be preferentially published. Journals serve as intermediaries and may suffer minimal immediate consequences for errors of over- or mis-estimation, but it is the consumers of these laboratory and clinical results (other expert scientists; trainees choosing fields of endeavour; physicians and their patients; funding agencies; the media) who are "cursed" if these results are severely exaggerated—overvalued and unrepresentative of the true outcomes of many similar experiments* (p. 1418).

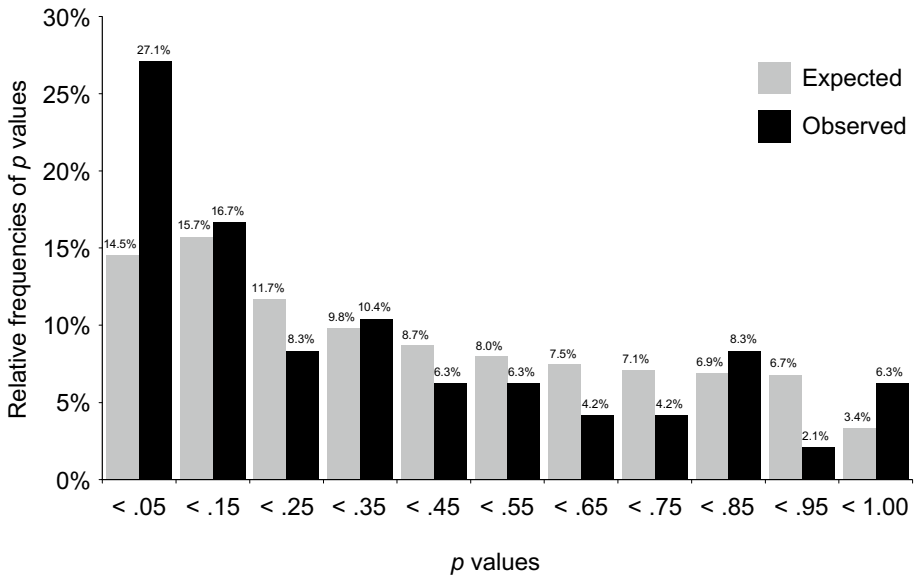

Fig. 10 The expected and observed frequencies of p values, in intervals ranging from $p < 0.05$ to $0.95 < p < 1.00$, resulting from the studies on maximum oxygen consumption ($\dot{V}O_{2max}$) included in the meta-analysis by Mattioni Maturana et al. [65], illustrating the presence of an excessive proportion of studies with $p < 0.05$

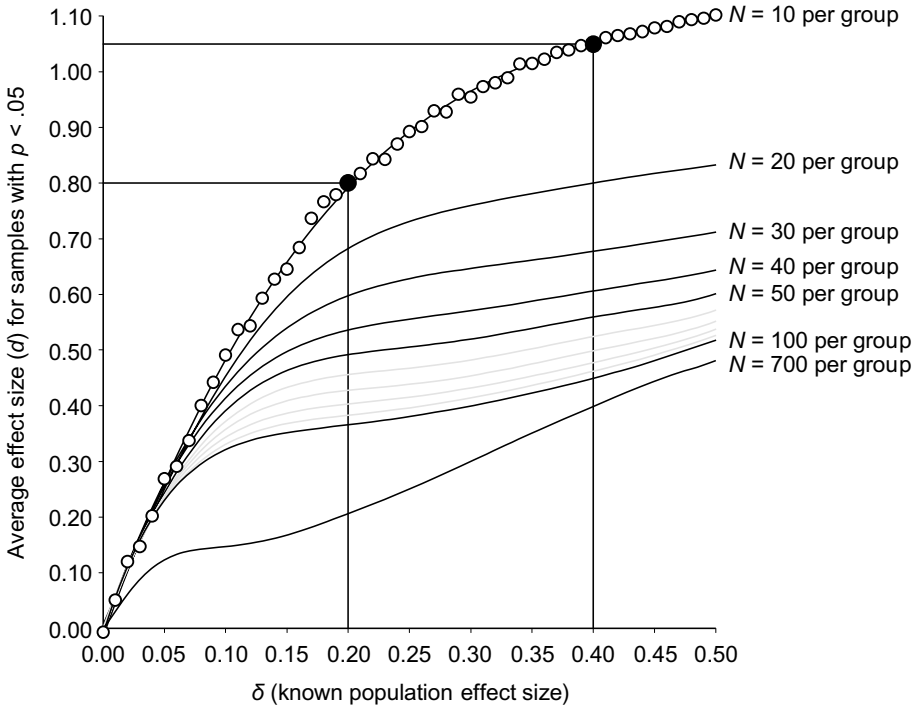

Fig. 11 Results of simulated experiments (100,000 simulated tests per data point) illustrating the phenomenon of 'winner's curse,' namely the inflation of the apparent effect size (d) compared with the known population effect size (δ) from studies with various sample sizes resulting in $p < 0.05$. For sample sizes of 10 per group, namely the median sample size of the 48 studies on maximum oxygen consumption ($\dot{V}O_{2max}$) included in the meta-analysis by Mattioni Maturana et al. [65], a small effect ($\delta = 0.20$) can appear as large ($d = 0.80$), while a population effect size of $\delta = 0.40$ (the pooled effect from the meta-analysis by Mattioni Maturana et al. [65]) can appear highly exaggerated, namely $d = 1.04$. Notice that samples of $N = 100$ per group suffice to eliminate the inflation of medium population effect sizes ($\delta = d = 0.50$) but samples of $N = \sim 700$ per group are required to eliminate the inflation for small population effect sizes ($\delta = d = 0.20$)

The 'winner's curse' can be shown by simulation, following the procedure proposed by Colquhoun [116]. If we consider the pooled effect size reported by Mattioni Maturana et al. [65], namely $d = 0.40$, and run 100,000 simulated 'experiments' by drawing random samples of 100 per group from populations designed to differ by $d = 0.40$ (i.e.,

experiments with 80% statistical power), we find that (i) consistent with the theoretical power level of 80.36%, 80.38% of the comparisons satisfy the $p < 0.05$ criterion of statistical significance, and (ii) importantly, the average observed effect size is $d = 0.45$, which approximates the given effect size of $d = 0.40$. On the other hand, if one runs 100,000 simulated experiments with the same effect size but sample sizes of 10 per group, namely the median sample size of the 48 studies on $\dot{V}O_{2max}$ included in the meta-analysis by Mattioni Maturana et al. [65], (i) the statistical power of 13.66% approximates the theoretical value of 13.55% but (ii) the average observed effect size is highly exaggerated, namely $d = 1.04$ instead of the given $\delta = 0.40$ (see Fig. 11). Indeed, after excluding an apparent outlier with a nearly fivefold effect size [66], the average effect size of the remaining 12 studies on $\dot{V}O_{2max}$ in the meta-analysis by Mattioni Maturana et al. [65] that produced $p < 0.05$ was 1.01. In general, larger sample sizes enable the estimation of the population effects with greater precision, whereas small samples increase the risk of greatly exaggerated estimates of effects.

3.1.6 Accuracy of Population Estimates

Davis-Stober and Dana [117] have proposed an index of the accuracy of population estimates produced by the conventional method of ordinary least squares (used in most of the commonly employed statistical tests, including tests of comparisons between sample means) compared against a 'benchmark' method of estimation that uses random estimates for both the direction and the magnitude of treatment effects (called 'random least squares'). The index, called the v-statistic, can range from zero to one, with a value of one indicating that the conventional method of estimation (ordinary least squares) is consistently more accurate than the random method, and a value of zero indicating that the random method of estimation is consistently more accurate than ordinary least squares. The values of the v-statistic are influenced by (i) the sample sizes, (ii) the magnitude of the effect being investigated, and (iii) the number of parameters that need to be estimated (i.e., two means in the case of a *t*-test). Preempting the criticism that comparing the accuracy of statistical tests against a 'benchmark' of random guessing sets a meaninglessly 'low bar,' Davis-Stober and Dana [117] wrote:

> *If one's estimates are less accurate than our guessing benchmark more than half of the time, there is little point in using them to establish treatment effects. As low as this hurdle may seem, we show that v < 0.5, or even v = 0, can happen surprisingly often, particularly when researching effect sizes conventionally categorized as small and medium* (p. 6)

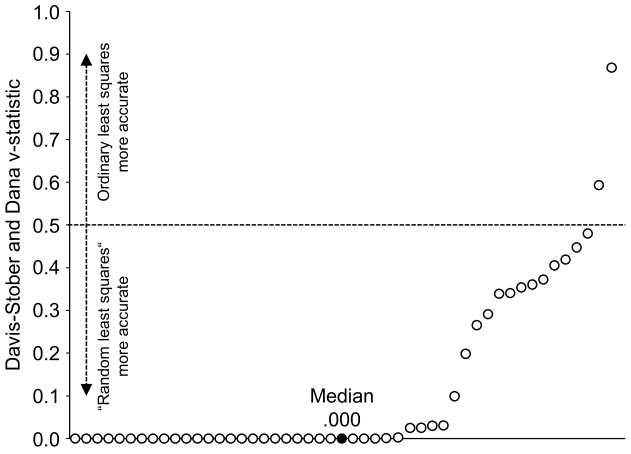

Fig. 12 Values of the v-statistic proposed by Davis-Stober and Dana [116] for each of the 48 studies on maximum oxygen consumption ($\dot{V}O_{2max}$) included in the meta-analysis by Mattioni Maturana et al. [65], comparing the effects of high intensity interval training (HIIT) and moderate-intensity continuous exercise. The v-statistic is an index of the relative accuracy of population estimates produced by the traditional method of ordinary least squares compared with 'random least squares' (i.e., random estimates for both the direction and the magnitude of treatment effects). The average v-statistic was 0.124 and the median was 0.000. Nearly all studies (46 of 48, or 96%) had values of the v-statistic below 0.500, and more than half (28 of 48, or 58%) had a v-statistic of zero, suggesting that random estimates were consistently more accurate than estimates based on the observed data

This is precisely the scenario encountered in the HIIT literature: small- to medium-size effects are being studied with small samples. Therefore, to gauge the accuracy of estimates derived from the studies included in the meta-analysis by Mattioni Maturana et al. [65], comparing the effects of HIIT and moderate-intensity continuous exercise on $\dot{V}O_{2max}$, the v-statistic for each study was calculated following the computational method outlined by Lakens and Evers [118]. The average v-statistic was 0.124 and the median was 0.000. Nearly all studies (46 of 48, or 96%) had values of the v-statistic below 0.500, and more than half (28 of 48, or 58%) had a v-statistic of zero (see Fig. 12). In the words of Lakens and Evers [118], "obviously, if a random estimator is more accurate than the estimator based on the observed data (indicated by a v-statistic smaller than 0.5), a study does not really reduce the uncertainty about whether the hypothesis is true" (p. 283).

3.1.7 Summary

When judged by conventional statistical standards, most studies investigating the effects of HIIT on fitness or health have limited informational yield. This is because they are examining small-to-medium effects with small samples, and commonly test a plethora of dependent variables. Estimates

of small-to-medium effects derived from small, underpowered studies are characterized by such imprecision and volatility that, given a large enough number of tests, some will probably cross the conventional threshold of statistical significance. Such 'statistically significant' results will likely reflect chance and, therefore, entail a low probability of replication. In addition, even if they represent true effects, such results likely overestimate the magnitude of the underlying effects.

3.2 The 'Is As Effective As' Problem

As noted in Sect. 2, statisticians commonly emphasize that "absence of evidence is not evidence of absence" [53, 54]. The principle behind this motto is that $p > 0.05$ (i.e., 'absence of evidence') provides no indication that the null effect, namely $\mu_1 - \mu_2 = 0$, is the most likely result (i.e., 'evidence of absence'). In other words, finding $p > 0.05$ for a comparison between two sample means (such as the mean of a group participating in HIIT and a group participating in moderate-intensity continuous exercise training) only permits a researcher to decide not to reject the null hypothesis. Such a result cannot be taken as a basis for *accepting* the null hypothesis (i.e., to conclude that there is 'no difference' or that the two treatments being compared have effects that are 'same,' 'equal,' 'similar,' 'equivalent,' or 'comparable').

Establishing the 'equivalence' of two interventions requires a different hypothesis, different design, different power calculations, and a different statistical approach [50–52]. An equivalence study begins with the difficult decision of determining a difference between the treatments that represents the smallest effect size of interest (e.g., smaller than any effect that can be considered clinically relevant, meaningful, or worthwhile). Then, the null hypothesis is formulated, stating that the difference between the two treatment means, or part of its surrounding confidence interval, falls outside the prespecified margin (i.e., suggesting that the treatments may not be equivalent, or one may be meaningfully more effective than the other). The alternative hypothesis would be that the difference between the treatments, and its surrounding confidence interval, are within the prespecified margin (i.e., that the treatments are equivalent, or one is as effective as the other). Power calculations for an equivalence study are based on the largest treatment difference considered to be practically irrelevant or inconsequential. The hypothesis of equivalence can be tested by specialized procedures, such as the two one-sided tests (TOST) method [119–121].

Most researchers carefully avoid the use of the adjectives 'similar' or 'comparable' (let alone 'equal' or 'same') to describe treatment means following a finding of $p > 0.05$. This is because a very common scenario is that tests fail to reject the null hypothesis, even though it is false, because of low statistical power (e.g., having too few participants to detect an effect given the magnitude of that effect). Yet, the HIIT literature contains numerous claims that various HIIT protocols have 'similar' or 'comparable' effects to more time-consuming moderate-intensity continuous exercise. Invariably, these claims are made on the basis of findings of $p > 0.05$ from studies that are underpowered to detect small ($d = 0.20$, requiring $N = 394$ per group), medium ($d = 0.50$, requiring $N = 64$ per group), or even large effects ($d = 0.80$, requiring $N = 26$ per group). As noted earlier, of the 48 studies included in the Mattioni Maturana et al. meta-analysis [65] comparing HIIT to moderate-intensity continuous exercise on $\dot{V}O_{2max}$, all but one (47 of 48, or 98%) had statistical power in the 0–33% range. Examples of claims made on the basis of underpowered studies include claims of 'equal' changes across a wide range of physiological parameters (samples of 8 and 8) [92], 'similar' changes in aerobic capacity (samples of 7 and 7) [122], 'similar' metabolic adaptations (samples of 10 and 10) [89], 'similar' changes in arterial stiffness (samples of 10 and 10) [123], 'similar' cardiometabolic changes (samples of 9, 10, and 6) [90], 'similar' cardiorespiratory adaptations in patients with heart failure (samples of 8 and 8) [124], 'similar' changes in body composition and fitness (samples of 16, 16, and 14) [125], 'similar' muscular and performance changes (samples of 8 and 8) [126], and 'similar' enjoyment and adherence (samples of 9 and 8) [127]. Likewise, such claims are made on the basis of findings of $p > 0.05$ from studies using within-subject designs that are also underpowered to detect small ($d = 0.20$, requiring $N = 199$), medium ($d = 0.50$, requiring $N = 34$), or even large effects ($d = 0.80$, requiring $N = 15$). Examples include claims of 'similar' adaptations in signaling molecules associated with mitochondrial biogenesis ($N = 10$) [128], 'similar' mitochondrial function ($N = 8$) [129], 'similar' 24-h oxygen consumption ($N = 8$) [130], 'similar' energy expenditure ($N = 9$) [131], 'similar' increases in serum brain-derived neurotrophic factor ($N = 8$) [132], and 'similar' enjoyment levels ($N = 7$ [133]; $N = 11$ [134]). To reiterate the essential point, claims of 'similar' or 'comparable' effects are unjustified on the basis of 'non-significant' comparisons between means ($p > 0.05$). Claims of 'similar' or 'comparable' effects can only be justified if appropriate hypotheses and associated tests (i.e., of equivalence or non-inferiority) are used [119–121].

3.2.1 Poor Reporting of Power Calculations

By using $p > 0.05$ as a criterion for establishing equivalence, there is no end to the extraordinary discoveries that researchers can claim. One common approach has been using severely underpowered comparative studies in conjunction with the $p > 0.05$ criterion in a race to discover the smallest duration or amount of exercise that can still be claimed to

Table 3 Synopsis of the sample-size calculations of the 11 studies included in the review by Sabag et al. [135], comparing the effects of low-volume HIIT to traditional HIIT or moderate-intensity continuous exercise

Study	Samples	Verbatim section on power	Comment
Tjønna et al., 2013 [136]	13 & 13	Prior experience suggests a SD of about 2.0–3.0 mL/kg/min. According to sample size tables for clinical studies, we needed 10 subjects in each group (we included 13 in case of drop out). With a standardized within-group difference of 1.0, differences may be detected using a paired t-test with 80% power, at a significance level of 5%. Clinically, this corresponds to a detectable difference for $\dot{V}O_{2max}$ of 3 mL/kg/min (p. 3)	While the calculations for a matched-pair t-test with $d=1.0$ indeed yields a required sample size of $N=10$, the cited source did not yield an effect size of $d=1.0$. Also, the focus of this study was not on 'within group' changes but rather inter-group comparisons, and the analysis was not based on a matched-pair t-test but rather on "mixed linear model analyses with group and time interaction"
Ramos et al., 2017 [137]	21 & 22	Sample size for the substudy was calculated using an anticipated mean difference in MetS z-score reduction of 0.60 (power = 0.80, alpha = 0.05 for two-tailed test) between HIIT and MICT groups. This was based on a previous study showing a similar mean difference in reduction of MetS z-score between HIIT and MICT (Supplementary material)	The information provided lacks standard deviation. The cited source does not report a mean difference in MetS z-score reduction 'similar' to 0.60 but rather 0.46 ± 1.55, and $d=0.29$. This entails a total sample size of $N=(188+188)=376$
Oh et al., 2017 [138]	20 & 13	Our study design did not consider sampling size calculation to estimate the effect of sample size. Therefore, the small sample size might have limited the statistical power of the study (p. 10)	No sample-size calculation
Winding et al., 2018 [139]	13 & 12	A limitation of the present study is the relatively small number of participants, which may have masked differences between HIIT and END (p. 1138)	No sample-size calculation
Abdelbasset et al., 2020 [140]	16 & 15	For sample size estimation, an initial power analysis was applied (2-tailed test with statistical power of 0.80, a error = 0.05, and effect size = 0.5). Estimates of mean difference and standard deviation for the [intrahepatic triglycerides] value from the previous study assessed 19 patients who received aerobic exercise. According to that study measures, 13 patients were required in each group. Forty-eight patients were included [for three groups] in the study to account for the dropout rate of 20% (p. 3)	Given the cited assumptions ($d=0.5$, $\alpha=0.05$, power = 0.80), the required sample is $N=64$ per group (128 for two groups, 192 for three groups, 230 with 20% oversampling for dropout). However, the cited source (which did not include power calculations) did not yield $d=0.5$ for the comparison between exercise and placebo for hepatic triglyceride concentration but rather $d=0.3$. This entails $N=170$ per group, (340 for two groups, 510 for three, 612 with 20% oversampling for dropout)
Poon et al., 2020 [141]	12 & 12	Based on a meta-analysis that compared HIIT with continuous endurance training on maximal oxygen update ($\dot{V}O_{2max}$) improvements in adults, the estimated standardized mean difference (Cohen's d) between HIIT and MICT was approximately 0.4. Therefore, it was anticipated that a sample size of 12 participants per group was adequate to detect this difference between groups on our primary outcome (i.e., $\dot{V}O_{2max}$), with a power of 0.8 at an alpha level of 0.05 (pp. 1998–1999)	The cited source (meta-analysis) reported non-standardized results (i.e., not Cohen's *d*). When converted to *d* using the information given (mean difference, 95% confidence limits), *d* was not 0.4. More importantly, the sample size required for $d=0.4$, $\alpha=0.05$, power = 0.8 is $N=100$ per group, not 12
Sabag et al., 2020 [142]	12 & 12	An a priori, two-tailed power calculation at an α of 0.05 and β of 0.8 gave an actual power of 0.813 for a sample size of 11 in each group. This calculation was determined using the effect size (ES) of 1.28 of a similar exercise intervention from a previous study, which detected significant improvements in liver fat within groups (p. 2373)	Besides confusing β and $1-\beta$ (power), the researchers referred to an effect size 'within groups' as the basis for power calculations for a between-groups comparison (also, the reported effect size for high-intensity, low volume exercise was 1.42 for intrahepatic lipids, not 1.28). In the cited source, the effect size for the comparison between high-intensity, low-volume exercise and low-intensity high-volume exercise was $d=0.19$, requiring $N=(436+436)=872$
Ryan et al., 2020 [143]	16 & 14		No sample-size calculation

Table 3 (continued)

Study	Samples	Verbatim section on power	Comment
Matsuo et al., 2014 [144]	14 & 14	A priori power analysis was performed to determine the sample size. The primary outcome variable of this study was the increase of $\dot{V}O_{2max}$ achieved through three types of exercise intervention. On the basis of data from both a previous study and our preliminary study on changes in $\dot{V}O_{2max}$, we assumed a 15% difference in the training effect between the three groups with an SD estimate of 10%. With an alpha error rate of 0.017 (with Bonferroni adjustment for post hoc tests) and statistical power of 80%, the minimal sample size in each group was estimated to be 11 subjects (33 subjects in total). Assuming subject attrition such as dropout, we recruited 14 subjects for each group (42 subjects in total) in this study (p. 46)	Assuming a large effect $d = 1.5$ with an adjusted $\alpha = (0.05/3) = 0.017$ indeed requires only $N = 11$ per group. However, the cited preliminary study only reported within-subjects changes in $\dot{V}O_{2max}$ in two participants, not intergroup differences or standard deviations. Moreover, the 'previous study' was conducted on a patient population (heart failure), with low baseline levels of $\dot{V}O_{2max}$ and there is no indication of a "15% difference in the training effect" (the cited study reported increases of 46% vs 14%, for interval and moderate continuous exercise, respectively)
Wilson et al., 2019 [145]	11 & 5		No sample-size calculation
Way et al., 2020 [146]	12 & 12	Sample size was calculated based on a projected change in peripheral arterial stiffness [pulse wave velocity] with [moderate-intensity continuous training] in adults with [type 2 diabetes] similar to the [moderate-intensity continuous training] protocol in our study. A priori, two-tailed power calculation of $\alpha = 0.05$ and $\beta = 0.20$ gave a power of 0.82 for a total sample size of 45 ($n = 15$ per group) (p. 150)	The researchers did not cite an anticipated effect size, so the calculations cannot be verified. Solving for the missing effect size shows that the study was sufficiently powered only for a large between-group effect ($d = 1.2$). The researchers reported basing their calculations on within-group changes but their analyses were for inter-group comparisons. The cited source reported $d = 0.80$ (radial) and $d = 0.50$ (femoral) for within-group changes and $d = 1.10$ (radial) and $d = 0.84$ (femoral) for inter-group comparisons

HIIT high-intensity interval training, *MICT* moderate-intensity continuous training, *SD* standard deviation, $\dot{V}O_{2max}$ maximum oxygen consumption

be 'as effective as' (or 'similar' or 'comparable' to) either 'traditional' HIIT or moderate-intensity continuous exercise. These minimalist forms have been termed 'low-volume HIIT,' 'very low volume HIIT,' or 'reduced exertion HIIT,' among other labels.

To illustrate the problems associated with this approach, we examined the studies included in a recent systematic review of 'low-volume HIIT,' which concluded that it "can induce similar, and at times greater, improvements in cardiorespiratory fitness, glucose control, blood pressure, and cardiac function when compared to more traditional forms of aerobic exercise training including high-volume HIIT and moderate intensity continuous training, despite requiring less time commitment and lower energy expenditure" (p. 1013) [135]. This is a remarkable claim because 'low-volume HIIT' was said to differ from regular HIIT solely by entailing a lower total duration of high-intensity intervals (< 15 min). Otherwise, the two modalities of training were said to share common features (e.g., intensity of 80–100% $\dot{V}O_{2max}$ or HRmax, duration of each high-intensity interval of 1–4 min, work-to-rest ratio of 1:1 to 1:2). In other words, the review concluded that, contrary to conventional wisdom, doing less exercise is 'as effective as' (or, remarkably, even 'more effective than') doing more exercise while holding other important aspects of the exercise 'dose' constant.

The review was based on 11 studies (see Table 3) and used the adjective 'comparable' to describe the results of the comparisons between the minimalist versions of HIIT and the comparator groups in 9 of the 11 cases [135]. Predictably, the studies had the common denominator of being underpowered (sample size range: 5–22 per group, mean: 13.5, mode: 12). Using a two-tail test, a two-group comparative study with $N = 12$ per group has 7.6%, 21.6%, and 46.6% statistical power to detect a small ($d = 0.20$), medium ($d = 0.50$), and large ($d = 0.80$) effect, respectively.

Researchers might wonder how this is possible since item 7a of the CONSORT checklist explicitly states that authors must explain "how sample size was determined" [147]. Given the sample size range of 5–22 per group, it is unsurprising that the claimed adequacy of the sample size could not be verified in any of the 11 studies. In four, no information was provided for how the sample size was determined. In the remaining studies, the irregularities ranged from not providing complete information (e.g., not stating the anticipated effect size), citing nonverifiable or incorrect information (e.g., citing effect sizes for within-group changes from previous studies but aiming to conduct between-group comparisons), citing the effect size from an early study [66] that has been identified as an outlier [148], to reporting the required information but claiming that the sample size needed to be only a fraction of what the calculations indicated in order to reach the desired level of statistical power. As one example:

Based on a meta-analysis that compared HIIT with continuous endurance training on maximal oxygen update ($\dot{V}O_{2max}$max) improvements in adults, the estimated standardized mean difference (Cohen's d) between HIIT and [moderate-intensity continuous training] was approximately 0.4. Therefore, it was anticipated that a sample size of 12 participants per group was adequate to detect this difference between groups on our primary outcome (i.e., $\dot{V}O_{2max}$), with a power of 0.8 at an alpha level of 0.05 (pp. 1998–1999) [141].

To reach 80% statistical power given an effect size $d = 0.4$ requires 100 participants per group rather than 12. Bonafiglia et al. [149] similarly found that 21 of 27 studies included in a meta-analysis comparing the effects of sprint interval training and continuous training either did not report sample-size calculations or did not provide full information. The reporting of power calculations is suboptimal both in the medical literature [150] and within exercise and sport science [151]. According to Charles et al. [150], only 34% of trials published in medical journals reported all data required to calculate the sample size, had accurate calculations, and were based on accurate assumptions. Of the remaining, 43% did not report all the required parameters to allow readers to verify the calculation, and 5% did not report sample size calculations. Within exercise and sport science, the situation appears worse. An analysis of 120 manuscripts submitted to a prominent disciplinary journal [151] shows that the median sample size was 19. Only 12 of the manuscripts (10%) included any sample-size calculations and, of them, four did not provide a justification for the cited effect size. Similar to the situation in the HIIT literature discussed in this section [135], none of the 12 manuscripts provided all the information required to enable the correct reproduction of the cited sample-size goal (i.e., the statistical test to be conducted, the targeted effect size, the level of α, and the desired level of statistical power). This situation is of grave concern and necessitates urgent change [77].

4 A Crisis of Confidence, a Looming Trainwreck, or an Opportunity for Reform?

Over the past 15 years, the research literature on HIIT has produced some extraordinary claims, which, upon closer inspection, are backed by surprisingly fragile evidence. This phenomenon can be analyzed from several angles. Perhaps the striking discrepancy between the boldness of the claims and the limitations of the experimental evidence is a reflection of a field eager for a scientific breakthrough. As noted in Sect. 2, journal editors and peer reviewers may, consciously or subconsciously, "apply lower standards" (p. 4) [62] when evaluating manuscripts reporting findings that seem highly intriguing or novel. Likewise, the willingness of the press to disseminate, and occasionally amplify, the extraordinary claims surrounding HIIT also suggests that the public at large may be eager for a breakthrough from exercise science, some miraculous discovery that would magnify and accelerate the benefits of exercise while requiring less effort [152].

An equally fascinating question pertains to the apparent willingness of exercise science as a research field to enter a state of 'suspension of disbelief,' accepting and propagating claims that defy conventional wisdom and research choices that directly contradict established methodological and statistical best practices. Like other scientific fields, exercise science will inevitably, sooner or later, have to confront its own crisis of replication and confidence [63]. Postponing this conversation will not help avert it. Therefore, it seems ironic that, while a push for more stringent methodologies [112, 153] and more responsible reporting [154] is sweeping the scientific landscape, one of the most prominent research lines within exercise science is characterized by a preponderance of studies with questionable statistical standards.

In the previous sections, it was shown that most samples in the HIIT literature are small, and thus the studies are underpowered to detect small, medium, or even large effects. This is important because the effect sizes, in most cases (especially when HIIT is compared against moderate-intensity continuous exercise rather than a no-exercise control), are likely to be small. It was also shown that most studies do not have one outcome designated as primary but rather tend to include long lists of dependent variables, all of which are tested at $p < 0.05$, without consideration for the inflation of α. There is also great flexibility in designs, definitions, outcomes, and analytic approaches, from the definition of HIIT to the selection of variables to represent various domains of physiological function (e.g., metabolism). Moreover, extraordinary claims related to the effectiveness of HIIT, along with claims that HIIT addresses "the most commonly cited reason for not exercising" (p. 212) [155] or "the primary reason for [the] failure to exercise on a regular basis" (p. 61) [156], namely lack of time, stimulate the interest or curiosity of the public (e.g., the narrative that, contrary to current recommendations, one only needs to exercise for a few seconds per day). The intense interest from the media may encourage or incentivize researchers to produce research results that support compelling narratives but may have low replicability. In particular, claims that smaller and smaller amounts of exercise were found to be 'effective' for improving fitness and health are bound to capture the interest of the general public. For example, recent media reports have highlighted that repeated 4-s spurts of exercise, totaling no more than 2 min per day [157], or a single 3-s muscular contraction per day [158] have been found to result in 'significant' gains in aerobic capacity (by 13%) and muscular

strength (by 12%), respectively (based on samples of 11 and 13, respectively).

Arguably, there is a striking similarity between the patterns seen in the HIIT literature and what was unfolding in the research field investigating phenomena of behavioral priming within psychology in the 2000s. The literature was being inundated with findings that have been described as "implausible" (p. 13) [159], "spectacular" (p. 19) [160], "fascinating" (p. 20) [161], and "eye-catching and counter-intuitive… the kind of sexy research that popular science writers love to describe" (p. 6) [161]. Failed attempts to replicate several of these widely publicized results led to an ongoing 'replication crisis' [162] or 'crisis of confidence' [163] in psychology. In response, Nobel laureate Daniel Kahneman wrote an open letter to researchers involved in research on priming, in which he encouraged them to try to remove the question mark that had been attached to their field [164]. He emphasized: "Your problem is not with the few people who have actively challenged the validity of some priming results. It is with the much larger population of colleagues who in the past accepted your surprising results as facts when they were published." Reminding readers that "a posture of defiant denial is self-defeating," Kahneman pointed out what was at stake: "I see a train wreck looming. I expect the first victims to be young people on the job market. Being associated with a controversial and suspicious field will put them at a severe disadvantage in the competition for positions. Because of the high visibility of the issue, you may already expect the coming crop of graduates to encounter problems."

Although undertaking the kind of radical reforms advocated by Kahneman is unlikely to be universally appreciated or endorsed, psychology has, to some extent, entered a period of critical self-reflection. Many authors have argued that the replication crisis can be seen as an opportunity for positive change [165–167]. This perspective has grown into a movement [168] that has even been characterized, perhaps optimistically or prematurely, as a 'renaissance' [169]. The winds of change are reaching other fields, even beyond the social sciences, such as cancer biology and drug development, which are coming to terms with the fact that they, too, are facing a replication crisis [170, 171].

The replication crisis in psychology offers a potential blueprint for how exercise science could proceed. Arguing that there is no problem is certainly a comforting option but, to echo Kahneman, "a posture of defiant denial is self-defeating." Continuing to overlook the fundamental principles of statistics in pursuit of implausible results that will capture the next headline will predictably lead to poor long-term outcomes. The exorbitant claims in the HIIT literature could serve as a clarion call that should inspire a period of critical self-reflection and positive reform. Recognizing the pitfalls, returning to, and respecting the fundamentals could have a lasting positive influence on the integrity, societal value, and reputation of exercise science.

It is, therefore, encouraging that the first signs of reform within exercise science have started to appear. Statistical experts [23, 77] and journal editors [76, 99, 151, 172] are making strong cases about the need to improve the quality of research designs and statistical analyses. Newly created organizations, such as the Consortium for Transparency in Exercise Science [63] and the Society for Transparency, Openness, and Replication in Kinesiology, are spearheading educational initiatives aimed at promoting stronger research practices. In psychology, arguably one of the most consequential reform efforts has been the push to expand the practice of study preregistration [173–176]. Therefore, the growing number of journals within exercise science that encourage preregistration and welcome registered reports represents a particularly promising development [177]. Beyond these efforts, curricular reforms will be necessary, with the goal of significantly improving statistical literacy at both the undergraduate and postgraduate levels. At the undergraduate level, courses intended to promote critical appraisal skills, specifically designed for consumers of research information (i.e., future exercise professionals), should be considered a necessity for a field aspiring to fully transition to a model of evidence-based practice. At the postgraduate level, where most students are prospective producers of research information, the teaching of statistical skills should be combined with efforts to cultivate a mindset that welcomes openness and transparency while resisting the "disciplinary incentives" to "favor novelty over replication" (p. 615) [57]. Finally, an important issue that the extraordinary claims surrounding HIIT have brought to the surface is that the field of exercise science must critically reexamine its relationship with the mass media. Researchers, university press offices, and journal editors should also resist the temptation to construct and disseminate media-friendly narratives that are based on statistically questionable or fragile evidence.

Declarations

Conflict of Interest The authors declare no competing financial or other interests.

Author Contributions PE conceived and drafted the manuscript and responses to peer review and editorial comments. PS and NBT revised and edited the original and revised manuscript, each contributing additional, original intellectual content.

Funding No funding was received for the preparation and/or publication of this manuscript.

References

1. Haskell WL. Health consequences of physical activity: understanding and challenges regarding dose-response. Med Sci Sports Exerc. 1994;26(6):649–60. https://doi.org/10.1249/00005768-199406000-00001.
2. Pate RR. Physical activity and health: dose-response issues. Res Q Exerc Sport. 1995;66(4):313–7. https://doi.org/10.1080/02701367.1995.10607917.
3. Pate RR, Pratt M, Blair SN, Haskell WL, Macera CA, Bouchard C, Buchner D, Ettinger W, Heath GW, King AC, et al. Physical activity and public health: a recommendation from the Centers for Disease Control and Prevention and the American College of Sports Medicine. JAMA. 1995;273(5):402–7. https://doi.org/10.1001/jama.273.5.402.
4. US Department of Health and Human Services. Physical activity and health: a report of the Surgeon General. Atlanta, Georgia: US Department of Health and Human Services, Public Health Service, Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion; 1996.
5. Leon AS. Physical activity and cardiovascular health: a national consensus. Champaign: Human Kinetics; 1997.
6. NIH Consensus Development Panel on Physical Activity and Cardiovascular Health. Physical activity and cardiovascular health. JAMA. 1996;276(3):241–246. https://doi.org/10.1001/jama.1996.03540030075036
7. Blair SN, LaMonte MJ, Nichaman MZ. The evolution of physical activity recommendations: how much is enough? Am J Clin Nutr. 2004;79(5):913S-920S. https://doi.org/10.1093/ajcn/79.5.913S.
8. Dishman RK, Buckworth J. Increasing physical activity: a quantitative synthesis. Med Sci Sports Exerc. 1996;28(6):706–19. https://doi.org/10.1097/00005768-199606000-00010.
9. Troiano RP, Berrigan D, Dodd KW, Mâsse LC, Tilert T, McDowell M. Physical activity in the United States measured by accelerometer. Med Sci Sports Exerc. 2008;40(1):181–8. https://doi.org/10.1249/mss.0b013e31815a51b3.
10. Metzger JS, Catellier DJ, Evenson KR, Treuth MS, Rosamond WD, Siega-Riz AM. Patterns of objectively measured physical activity in the United States. Med Sci Sports Exerc. 2008;40(4):630–8. https://doi.org/10.1249/MSS.0b013e3181620ebc.
11. Tudor-Locke C, Brashear MM, Johnson WD, Katzmarzyk PT. Accelerometer profiles of physical activity and inactivity in normal weight, overweight, and obese U.S. men and women. Int J Behav Nutr Phys Act. 2010;7:60. https://doi.org/10.1186/1479-5868-7-60.
12. Winett RA. Developing more effective health-behavior programs: analyzing the epidemiological and biological bases for activity and exercise programs. Appl Prev Psychol. 1998;7(4):209–24. https://doi.org/10.1016/S0962-1849(98)80025-5.
13. Swain DP, Franklin BA. Comparison of cardioprotective benefits of vigorous versus moderate intensity aerobic exercise. Am J Cardiol. 2006;97(1):141–7. https://doi.org/10.1016/j.amjcard.2005.07.130.
14. O'Donovan G, Shave R. British adults' views on the health benefits of moderate and vigorous activity. Prev Med. 2007;45(6):432–5. https://doi.org/10.1016/j.ypmed.2007.07.026.
15. Haskell WL, Lee IM, Pate RR, Powell KE, Blair SN, Franklin BA, Macera CA, Heath GW, Thompson PD, Bauman A, American College of Sports Medicine; American Heart Association. Physical activity and public health: updated recommendation for adults from the American College of Sports Medicine and the American Heart Association. Circulation. 2007;116(9):1081–93. https://doi.org/10.1161/CIRCULATIONAHA.107.185649.
16. O'Donovan G, Blazevich AJ, Boreham C, Cooper AR, Crank H, Ekelund U, Fox KR, Gately P, Giles-Corti B, Gill JM, Hamer M, McDermott I, Murphy M, Mutrie N, Reilly JJ, Saxton JM, Stamatakis E. The ABC of Physical Activity for Health: a consensus statement from the British Association of Sport and Exercise Sciences. J Sports Sci. 2010;28(6):573–91. https://doi.org/10.1080/02640411003671212.
17. Burgomaster KA, Hughes SC, Heigenhauser GJ, Bradwell SN, Gibala MJ. Six sessions of sprint interval training increases muscle oxidative potential and cycle endurance capacity in humans. J Appl Physiol. 2005;98(6):1985–90. https://doi.org/10.1152/japplphysiol.01095.2004.
18. Coyle EF. Very intense exercise-training is extremely potent and time efficient: a reminder. J Appl Physiol. 2005;98(6):1983–4. https://doi.org/10.1152/japplphysiol.00215.2005.
19. Thompson WR. Worldwide survey of fitness trends for 2014. ACSM Health Fitness J. 2013;17(6):10–20.
20. Gray SR, Ferguson C, Birch K, Forrest LJ, Gill JM. High-intensity interval training: key data needed to bridge the gap from laboratory to public health policy. Br J Sports Med. 2016;50(20):1231–2. https://doi.org/10.1136/bjsports-2015-095705.
21. Steen RG. Misinformation in the medical literature: what role do error and fraud play? J Med Ethics. 2011;37(8):498–503. https://doi.org/10.1136/jme.2010.041830.
22. Viana RB, Naves JPA, Coswig VS, de Lira CAB, Steele J, Fisher JP, Gentil P. Is interval training the magic bullet for fat loss? A systematic review and meta-analysis comparing moderate-intensity continuous training with high-intensity interval training (HIIT). Br J Sports Med. 2019;53(10):655–64. https://doi.org/10.1136/bjsports-2018-099928.
23. Sainani KL, Borg DN, Caldwell AR, Butson ML, Tenan MS, Vickers AJ, Vigotsky AD, Warmenhoven J, Nguyen R, Lohse KR, Knight EJ, Bargary N. Call to increase statistical collaboration in sports science, sport and exercise medicine and sports physiotherapy. Br J Sports Med. 2021;55(2):118–22. https://doi.org/10.1136/bjsports-2020-102607.
24. Nickerson RS. Null hypothesis significance testing: a review of an old and continuing controversy. Psychol Methods. 2000;5(2):241–301. https://doi.org/10.1037/1082-989x.5.2.241.
25. Sterne JA, Davey SG. Sifting the evidence: what's wrong with significance tests? BMJ. 2001;322(7280):226–31. https://doi.org/10.1136/bmj.322.7280.226.
26. Christensen R. Testing Fisher, Neyman, Pearson, and Bayes. Am Stat. 2005;59(2):121–6. https://doi.org/10.1198/000313005X20871.
27. Lakens D. The practical alternative to the *p* value is the correctly used p value. Perspect Psychol Sci. 2021;16(3):639–48. https://doi.org/10.1177/1745691620958012.
28. Hubbard R, Bayarri MJ. Confusion over measures of evidence (p's) versus errors (α's) in classical statistical testing. Am Stat. 2003;57(3):171–8. https://doi.org/10.1198/0003130031856.
29. Fisher RA. Statistical methods for research workers. 5th ed. Edinburgh: Oliver and Boyd; 1934.
30. Neyman J, Pearson ESIX. On the problem of the most efficient tests of statistical hypotheses. Philos Trans R Soc Lond A. 1933;231:289–337. https://doi.org/10.1098/rsta.1933.0009.
31. Szucs D, Ioannidis JPA. When null hypothesis significance testing is unsuitable for research: a reassessment. Front Hum Neurosci. 2017;11:390. https://doi.org/10.3389/fnhum.2017.00390.
32. Neyman J, Pearson ES. The testing of statistical hypotheses in relation to probabilities a priori. Math Proc Camb Philos Soc. 1933;29(4):492–510. https://doi.org/10.1017/S030500410001152X.

33. Fisher RA. The design of experiments. Edinburgh: Oliver and Boyd; 1935.
34. Fisher R. Statistical methods and scientific induction. J R Stat Soc Ser B Methodol. 1955;17:69–78. https://doi.org/10.1111/j.2517-6161.1955.tb00180.x.
35. Lehmann EL. The Fisher, Neyman-Pearson theories of testing hypotheses: one theory or two? J Am Stat Assoc. 1993;88(424):1242–9. https://doi.org/10.1080/01621459.1993.10476404.
36. Lehmann EL. Fisher, Neyman, and the creation of classical statistics. New York: Springer; 2011. https://doi.org/10.1007/978-1-4419-9500-1.
37. Goodman S. A dirty dozen: twelve p-value misconceptions. Semin Hematol. 2008;45(3):135–40. https://doi.org/10.1053/j.seminhematol.2008.04.003.
38. Greenland S, Senn SJ, Rothman KJ, et al. Statistical tests, p values, confidence intervals, and power: a guide to misinterpretations. Eur J Epidemiol. 2016;31(4):337–50. https://doi.org/10.1007/s10654-016-0149-3.
39. Goodman SN. Toward evidence-based medical statistics. 1: the p value fallacy. Ann Intern Med. 1999;130(12):995–1004. https://doi.org/10.7326/0003-4819-130-12-199906150-00008.
40. Berger JO, Sellke T. Testing a point null hypothesis: the irreconcilability of p values and evidence. J Am Stat Assoc. 1987;82(397):112–22. https://doi.org/10.1080/01621459.1987.10478397.
41. Sellke T, Bayarri MJ, Berger JO. Calibration of p values for testing precise null hypotheses. Am Stat. 2001;55(1):62–71. https://doi.org/10.1198/000313001300339950.
42. Berger JO. Could Fisher, Jeffreys and Neyman have agreed on testing? Stat Sci. 2003;18(1):1–32. https://doi.org/10.1214/ss/1056397485.
43. Goodman SN. A comment on replication, p-values and evidence. Stat Med. 1992;11(7):875–9. https://doi.org/10.1002/sim.4780110705.
44. Halsey LG, Curran-Everett D, Vowler SL, Drummond GB. The fickle *p* value generates irreproducible results. Nat Methods. 2015;12(3):179–85. https://doi.org/10.1038/nmeth.3288.
45. Lazzeroni LC, Lu Y, Belitskaya-Lévy I. Solutions for quantifying *p*-value uncertainty and replication power. Nat Methods. 2016;13(2):107–8. https://doi.org/10.1038/nmeth.3741.
46. Cumming G. Replication and p intervals: *p* values predict the future only vaguely, but confidence intervals do much better. Perspect Psychol Sci. 2008;3(4):286–300. https://doi.org/10.1111/j.1745-6924.2008.00079.x.
47. Killeen PR. An alternative to null-hypothesis significance tests. Psychol Sci. 2005;16(5):345–53. https://doi.org/10.1111/j.0956-7976.2005.01538.x.
48. Lecoutre B, Lecoutre MP, Poitevineau J. Killeen's probability of replication and predictive probabilities: how to compute, use, and interpret them. Psychol Methods. 2010;15(2):158–71. https://doi.org/10.1037/a0015915.
49. Sanabria F, Killeen PR. Better statistics for better decisions: rejecting null hypotheses statistical tests in favor of replication statistics. Psychol Sch. 2007;44(5):471–81. https://doi.org/10.1002/pits.20239.
50. Amrhein V, Greenland S, McShane B. Scientists rise up against statistical significance. Nature. 2019;567(7748):305–7. https://doi.org/10.1038/d41586-019-00857-9.
51. Hoekstra R, Finch S, Kiers HA, Johnson A. Probability as certainty: dichotomous thinking and the misuse of p values. Psychon Bull Rev. 2006;13(6):1033–7. https://doi.org/10.3758/bf03213921.
52. Smith RJ. P > 0.05: The incorrect interpretation of "not significant" results is a significant problem. Am J Phys Anthropol. 2020;172(4):521–7. https://doi.org/10.1002/ajpa.24092.
53. Alderson P. Absence of evidence is not evidence of absence. BMJ. 2004;328(7438):476–7. https://doi.org/10.1136/bmj.328.7438.476.
54. Altman DG, Bland JM. Absence of evidence is not evidence of absence. BMJ. 1995;311(7003):485. https://doi.org/10.1136/bmj.311.7003.485.
55. Speed HD, Andersen MB. What exercise and sport scientists don't understand. J Sci Med Sport. 2000;3(1):84–92. https://doi.org/10.1016/s1440-2440(00)80051-1.
56. Vadillo MA, Konstantinidis E, Shanks DR. Underpowered samples, false negatives, and unconscious learning. Psychon Bull Rev. 2016;23(1):87–102. https://doi.org/10.3758/s13423-015-0892-6.
57. Nosek BA, Spies JR, Motyl M. Scientific utopia: II. Restructuring incentives and practices to promote truth over publishability. Perspect Psychol Sci. 2012;7(6):615–31. https://doi.org/10.1177/1745691612459058.
58. Anderson SF. Misinterpreting p: the discrepancy between p values and the probability the null hypothesis is true, the influence of multiple testing, and implications for the replication crisis. Psychol Methods. 2020;25(5):596–609. https://doi.org/10.1037/met0000248.
59. Colling LJ, Szucs D. Statistical inference and the replication crisis. Rev Phil Psychol. 2021;12(1):121–47. https://doi.org/10.1007/s13164-018-0421-4.
60. Colquhoun D. The reproducibility of research and the misinterpretation of p-values. R Soc Open Sci. 2017;4(12): 171085. https://doi.org/10.1098/rsos.171085.
61. Gigerenzer G. Statistical rituals: the replication delusion and how we got there. Adv Methods Pract Psychol Sci. 2018;1(2):198–218. https://doi.org/10.1177/2515245918771329.
62. Serra-Garcia M, Gneezy U. Nonreplicable publications are cited more than replicable ones. Sci Adv. 2021;7(21): eabd1705. https://doi.org/10.1126/sciadv.abd1705.
63. Caldwell AR, Vigotsky AD, Tenan MS, Radel R, Mellor DT, Kreutzer A, Lahart IM, Mills JP, Boisgontier MP, Consortium for Transparency in Exercise Science (COTES) Collaborators. Moving sport and exercise science forward: a call for the adoption of more transparent research practices. Sports Med. 2020;50(3):449–59. https://doi.org/10.1007/s40279-019-01227-1.
64. Bauer N, Sperlich B, Holmberg HC, Engel FA. Effects of high-intensity interval training in school on the physical performance and health of children and adolescents: a systematic review with meta-analysis. Sports Med Open. 2022;8(1):50. https://doi.org/10.1186/s40798-022-00437-8.
65. MattioniMaturana F, Martus P, Zipfel S, Nieß AM. Effectiveness of HIIE versus MICT in improving cardiometabolic risk factors in health and disease: a meta-analysis. Med Sci Sports Exerc. 2021;53(3):559–73. https://doi.org/10.1249/MSS.0000000000002506.
66. Wisløff U, Støylen A, Loennechen JP, Bruvold M, Rognmo Ø, Haram PM, Tjønna AE, Helgerud J, Slørdahl SA, Lee SJ, Videm V, Bye A, Smith GL, Najjar SM, Ellingsen Ø, Skjaerpe T. Superior cardiovascular effect of aerobic interval training versus moderate continuous training in heart failure patients: a randomized study. Circulation. 2007;115(24):3086–94. https://doi.org/10.1161/CIRCULATIONAHA.106.675041.
67. Andreacci JL, LeMura LM, Cohen SL, Urbansky EA, Chelland SA, Von Duvillard SP. The effects of frequency of encouragement on performance during maximal exercise testing. J Sports Sci. 2002;20(4):345–52. https://doi.org/10.1080/02640410275357612 5.
68. Halperin I, Pyne DB, Martin DT. Threats to internal validity in exercise science: a review of overlooked confounding

variables. Int J Sports Physiol Perform. 2015;10(7):823–9. https://doi.org/10.1123/ijspp.2014-0566.
69. Midgley AW, Marchant DC, Levy AR. A call to action towards an evidence-based approach to using verbal encouragement during maximal exercise testing. Clin Physiol Funct Imaging. 2018;38(4):547–53. https://doi.org/10.1111/cpf.12454.
70. Wisløff U, Coombes JS, Rognmo Ø. CrossTalk proposal: high intensity interval training does have a role in risk reduction or treatment of disease. J Physiol. 2015;593(24):5215–7. https://doi.org/10.1113/JP271041.
71. Khalafi M, Symonds ME. The impact of high-intensity interval training on inflammatory markers in metabolic disorders: a meta-analysis. Scand J Med Sci Sports. 2020;30(11):2020–36. https://doi.org/10.1111/sms.13754.
72. Solera-Martínez M, Herraiz-Adillo Á, Manzanares-Domínguez I, De La Cruz LL, Martínez-Vizcaíno V, Pozuelo-Carrascosa DP. High-intensity interval training and cardiometabolic risk factors in children: a meta-analysis. Pediatrics. 2021;148(4): e2021050810. https://doi.org/10.1542/peds.2021-050810.
73. Gerosa-Neto J, Antunes BM, Campos EZ, et al. Impact of long-term high-intensity interval and moderate-intensity continuous training on subclinical inflammation in overweight/obese adults. J Exerc Rehabil. 2016;12(6):575–80. https://doi.org/10.12965/jer.1632770.385.
74. Oh S, So R, Shida T, et al. High-intensity aerobic exercise improves both hepatic fat content and stiffness in sedentary obese men with nonalcoholic fatty liver disease. Sci Rep. 2017;7:43029. https://doi.org/10.1038/srep43029.
75. Paahoo A, Tadibi V, Behpoor N. Effectiveness of continuous aerobic versus high-intensity interval training on atherosclerotic and inflammatory markers in boys with overweight/obesity. Pediatr Exerc Sci. 2021;33(3):132–8. https://doi.org/10.1123/pes.2020-0138.
76. Gandevia S. Publications, replication and statistics in physiology plus two neglected curves. J Physiol. 2021;599(6):1719–21. https://doi.org/10.1113/JP281360.
77. Sainani K, Chamari K. Wish list for improving the quality of statistics in sport science. Int J Sports Physiol Perform. 2022;17(5):673–4. https://doi.org/10.1123/ijspp.2022-0023.
78. Rubin M. When to adjust alpha during multiple testing: a consideration of disjunction, conjunction, and individual testing. Synthese. 2021;199(3–4):10969–1000. https://doi.org/10.1007/s11229-021-03276-4.
79. Albers C. The problem with unadjusted multiple and sequential statistical testing. Nat Commun. 2019;10(1):1921. https://doi.org/10.1038/s41467-019-09941-0.
80. Forstmeier W, Wagenmakers EJ, Parker TH. Detecting and avoiding likely false-positive findings: a practical guide. Biol Rev Camb Philos Soc. 2017;92(4):1941–68. https://doi.org/10.1111/brv.12315.
81. Streiner DL. Best (but oft-forgotten) practices: the multiple problems of multiplicity - whether and how to correct for many statistical tests. Am J Clin Nutr. 2015;102(4):721–8. https://doi.org/10.3945/ajcn.115.113548.
82. Maxwell SE, Delaney HD, Kelley K. Designing experiments and analyzing data: a model comparison perspective. 3rd ed. Oxfordshire: Routledge; 2018.
83. Blakesley RE, Mazumdar S, Dew MA, Houck PR, Tang G, Reynolds CF 3rd, Butters MA. Comparisons of methods for multiple hypothesis testing in neuropsychological research. Neuropsychology. 2009;23(2):255–64. https://doi.org/10.1037/a0012850.
84. Sankoh AJ, Huque MF, Dubey SD. Some comments on frequently used multiple endpoint adjustment methods in clinical trials. Stat Med. 1997;16(22):2529–42. https://doi.org/10.1002/(sici)1097-0258(19971130)16:22%3c2529::aid-sim692%3e3.0.co;2-j.
85. Vickerstaff V, Omar RZ, Ambler G. Methods to adjust for multiple comparisons in the analysis and sample size calculation of randomised controlled trials with multiple primary outcomes. BMC Med Res Methodol. 2019;19(1):129. https://doi.org/10.1186/s12874-019-0754-4.
86. Sankoh AJ, D'Agostino RB Sr, Huque MF. Efficacy endpoint selection and multiplicity adjustment methods in clinical trials with inherent multiple endpoint issues. Stat Med. 2003;22(20):3133–50. https://doi.org/10.1002/sim.1557.
87. Cheverud JM. A simple correction for multiple comparisons in interval mapping genome scans. Heredity. 2001;87(Pt 1):52–8. https://doi.org/10.1046/j.1365-2540.2001.00901.x.
88. Nyholt DR. A simple correction for multiple testing for single-nucleotide polymorphisms in linkage disequilibrium with each other. Am J Hum Genet. 2004;74(4):765–9. https://doi.org/10.1086/383251.
89. Burgomaster KA, Howarth KR, Phillips SM, Rakobowchuk M, Macdonald MJ, McGee SL, Gibala MJ. Similar metabolic adaptations during exercise after low volume sprint interval and traditional endurance training in humans. J Physiol. 2008;586(1):151–60. https://doi.org/10.1113/jphysiol.2007.142109.
90. Gillen JB, Martin BJ, MacInnis MJ, Skelly LE, Tarnopolsky MA, Gibala MJ. 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. 2016;11(4): e0154075. https://doi.org/10.1371/journal.pone.0154075.
91. Robinson E, Durrer C, Simtchouk S, Jung ME, Bourne JE, Voth E, Little JP. Short-term high-intensity interval and moderate-intensity continuous training reduce leukocyte TLR4 in inactive adults at elevated risk of type 2 diabetes. J Appl Physiol. 2015;119(5):508–16. https://doi.org/10.1152/japplphysiol.00334.2015.
92. Cocks M, Shaw CS, Shepherd SO, Fisher JP, Ranasinghe A, Barker TA, Wagenmakers AJ. 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. J Physiol. 2016;594(8):2307–21. https://doi.org/10.1113/jphysiol.2014.285254.
93. McGiffin DC, Cumming G, Myles PS. The frequent insignificance of a "significant" p-value. J Card Surg. 2021;36(11):4322–31. https://doi.org/10.1111/jocs.15960.
94. Locke SR, Bourne JE, Beauchamp MR, Little JP, Barry J, Singer J, Jung ME. High-intensity interval or continuous moderate exercise: a 24-week pilot trial. Med Sci Sports Exerc. 2018;50(10):2067–75. https://doi.org/10.1249/MSS.0000000000001668.
95. Albers C, Lakens D. When power analyses based on pilot data are biased: inaccurate effect size estimators and follow-up bias. J Exp Soc Psychol. 2018;74:187–95. https://doi.org/10.1016/j.jesp.2017.09.004.
96. Kraemer HC, Mintz J, Noda A, Tinklenberg J, Yesavage JA. Caution regarding the use of pilot studies to guide power calculations for study proposals. Arch Gen Psychiatry. 2006;63(5):484–9. https://doi.org/10.1001/archpsyc.63.5.484.
97. van Zwet EW, Goodman SN. How large should the next study be? Predictive power and sample size requirements for replication studies. Stat Med. 2022;41(16):3090–101. https://doi.org/10.1002/sim.9406.
98. Curran-Everett D. Explorations in statistics: statistical facets of reproducibility. Adv Physiol Educ. 2016;40(2):248–52. https://doi.org/10.1152/advan.00042.2016.
99. Gandevia S, Cumming G, Amrhein V, Butler A. Replication: do not trust your p-value, be it small or large. J Physiol. 2021;599(11):2989–90. https://doi.org/10.1113/JP281614.

100. Gorroochurn P, Hodge SE, Heiman GA, Durner M, Greenberg DA. Non-replication of association studies: "pseudo-failures" to replicate? Genet Med. 2007;9(6):325–31. https://doi.org/10.1097/gim.0b013e3180676d79.
101. Gibson EW. The role of p-values in judging the strength of evidence and realistic replication expectations. Stat Biopharm Res. 2021;13(1):6–18. https://doi.org/10.1080/19466315.2020.1724560.
102. Boos DD, Stefanski LA. P-value precision and reproducibility. Am Stat. 2011;65(4):213–21. https://doi.org/10.1198/tas.2011.10129.
103. Hung HM, O'Neill RT, Bauer P, Köhne K. The behavior of the P-value when the alternative hypothesis is true. Biometrics. 1997;53(1):11–22. https://doi.org/10.2307/2533093.
104. Sackrowitz H, Samuel-Cahn E. P values as random variables: expected p values. Am Stat. 1999;53(4):326–31. https://doi.org/10.1080/00031305.1999.10474484.
105. Shao J, Chow SC. Reproducibility probability in clinical trials. Stat Med. 2002;21(12):1727–42. https://doi.org/10.1002/sim.1177.
106. Amrhein V, Korner-Nievergelt F, Roth T. The earth is flat (p > 0.05): significance thresholds and the crisis of unreplicable research. PeerJ. 2017;5: e3544. https://doi.org/10.7717/peerj.3544.
107. Button KS, Ioannidis JP, Mokrysz C, Nosek BA, Flint J, Robinson ES, Munafò MR. Power failure: why small sample size undermines the reliability of neuroscience. Nat Rev Neurosci. 2013;14(5):365–76. https://doi.org/10.1038/nrn3475.
108. Ioannidis JP. Why most published research findings are false. PLoS Med. 2005;2(8): e124. https://doi.org/10.1371/journal.pmed.0020124.
109. Dumas-Mallet E, Button KS, Boraud T, Gonon F, Munafò MR. Low statistical power in biomedical science: a review of three human research domains. R Soc Open Sci. 2017;4(2): 160254. https://doi.org/10.1098/rsos.160254.
110. Ioannidis JP, Trikalinos TA. An exploratory test for an excess of significant findings. Clin Trials. 2007;4(3):245–53. https://doi.org/10.1177/1740774507079441.
111. Masicampo EJ, Lalande DR. A peculiar prevalence of p values just below 0.05. Q J Exp Psychol. 2012;65(11):2271–9. https://doi.org/10.1080/17470218.2012.711335.
112. Simmons JP, Nelson LD, Simonsohn U. False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. Psychol Sci. 2011;22(11):1359–66. https://doi.org/10.1177/0956797611417632.
113. Rosenthal R. The file drawer problem and tolerance for null results. Psychol Bull. 1979;86(3):638–41. https://doi.org/10.1037/0033-2909.86.3.638.
114. Ioannidis JP. Why most discovered true associations are inflated. Epidemiology. 2008;19(5):640–8. https://doi.org/10.1097/EDE.0b013e31818131e7.
115. Young NS, Ioannidis JP, Al-Ubaydli O. Why current publication practices may distort science. PLoS Med. 2008;5(10): e201. https://doi.org/10.1371/journal.pmed.0050201.
116. Colquhoun D. An investigation of the false discovery rate and the misinterpretation of p-values. R Soc Open Sci. 2014;1(3): 140216. https://doi.org/10.1098/rsos.140216.
117. Davis-Stober CP, Dana J. Comparing the accuracy of experimental estimates to guessing: a new perspective on replication and the "crisis of confidence" in psychology. Behav Res Methods. 2014;46(1):1–14. https://doi.org/10.3758/s13428-013-0342-1.
118. Lakens D, Evers ER. Sailing from the seas of chaos into the corridor of stability: practical recommendations to increase the informational value of studies. Perspect Psychol Sci. 2014;9(3):278–92. https://doi.org/10.1177/1745691614528520.
119. Lakens D. Equivalence tests: a practical primer for t tests, correlations, and meta-analyses. Soc Psychol Personal Sci. 2017;8(4):355–62. https://doi.org/10.1177/1948550617697177.
120. Parkhurst DF. Statistical significance tests: equivalence and reverse tests should reduce misinterpretation. Bioscience. 2001;51(12):1051–7. https://doi.org/10.1641/0006-3568(2001)051[1051:SSTEAR]2.0.CO;2.
121. Mazzolari R, Porcelli S, Bishop DJ, Lakens D. Myths and methodologies: the use of equivalence and non-inferiority tests for interventional studies in exercise physiology and sport science. Exp Physiol. 2022;107(3):201–12. https://doi.org/10.1113/EP090171.
122. McRae G, Payne A, Zelt JG, Scribbans TD, Jung ME, Little JP, Gurd BJ. Extremely low volume, whole-body aerobic-resistance training improves aerobic fitness and muscular endurance in females. Appl Physiol Nutr Metab. 2012;37(6):1124–31. https://doi.org/10.1139/h2012-093.
123. Rakobowchuk M, Tanguay S, Burgomaster KA, Howarth KR, Gibala MJ, MacDonald MJ. Sprint interval and traditional endurance training induce similar improvements in peripheral arterial stiffness and flow-mediated dilation in healthy humans. Am J Physiol Regul Integr Comp Physiol. 2008;295(1):R236–42. https://doi.org/10.1152/ajpregu.00069.2008.
124. Iellamo F, Manzi V, Caminiti G, Vitale C, Castagna C, Massaro M, Franchini A, Rosano G, Volterrani M. Matched dose interval and continuous exercise training induce similar cardiorespiratory and metabolic adaptations in patients with heart failure. Int J Cardiol. 2013;167(6):2561–5. https://doi.org/10.1016/j.ijcard.2012.06.057.
125. Martins C, Kazakova I, Ludviksen M, Mehus I, Wisloff U, Kulseng B, Morgan L, King N. High-intensity interval training and isocaloric moderate-intensity continuous training result in similar improvements in body composition and fitness in obese individuals. Int J Sport Nutr Exerc Metab. 2016;26(3):197–204. https://doi.org/10.1123/ijsnem.2015-0078.
126. Gibala MJ, Little JP, van Essen M, Wilkin GP, Burgomaster KA, Safdar A, Raha S, Tarnopolsky MA. Short-term sprint interval versus traditional endurance training: similar initial adaptations in human skeletal muscle and exercise performance. J Physiol. 2006;575(Pt 3):901–11. https://doi.org/10.1113/jphysiol.2006.112094.
127. Vella CA, Taylor K, Drummer D. High-intensity interval and moderate-intensity continuous training elicit similar enjoyment and adherence levels in overweight and obese adults. Eur J Sport Sci. 2017;17(9):1203–11. https://doi.org/10.1080/17461391.2017.1359679.
128. Bartlett JD, Hwa Joo C, Jeong TS, Louhelainen J, Cochran AJ, Gibala MJ, Gregson W, Close GL, Drust B, Morton JP. Matched work high-intensity interval and continuous running induce similar increases in PGC-1a mRNA, AMPK, p38, and p53 phosphorylation in human skeletal muscle. J Appl Physiol. 2012;112(7):1135–43. https://doi.org/10.1152/japplphysiol.01040.2011.
129. Trewin AJ, Parker L, Shaw CS, Hiam DS, Garnham A, Levinger I, McConell GK, Stepto NK. Acute HIIE elicits similar changes in human skeletal muscle mitochondrial H2O2 release, respiration, and cell signaling as endurance exercise even with less work. Am J Physiol Regul Integr Comp Physiol. 2018;315(5):R1003–16. https://doi.org/10.1152/ajpregu.00096.2018.
130. Hazell TJ, Olver TD, Hamilton CD, Lemon PWR. Two minutes of sprint-interval exercise elicits 24-hr oxygen consumption similar to that of 30 min of continuous endurance exercise. Int J Sport Nutr Exerc Metab. 2012;22(4):276–83. https://doi.org/10.1123/ijsnem.22.4.276.

131. Skelly LE, Andrews PC, Gillen JB, Martin BJ, Percival ME, Gibala MJ. High-intensity interval exercise induces 24-h energy expenditure similar to traditional endurance exercise despite reduced time commitment. Appl Physiol Nutr Metab. 2014;39(7):845–8. https://doi.org/10.1139/apnm-2013-0562.
132. Saucedo Marquez CM, Vanaudenaerde B, Troosters T, Wenderoth N. High-intensity interval training evokes larger serum BDNF levels compared with intense continuous exercise. J Appl Physiol. 2015;119(12):1363–73. https://doi.org/10.1152/japplphysiol.00126.2015.
133. Sagelv EH, Hammer T, Hamsund T, Rognmo K, Pettersen SA, Pedersen S. High intensity long interval sets provides similar enjoyment as continuous moderate intensity exercise: the Tromsø Exercise Enjoyment Study. Front Psychol. 2019;10:1788. https://doi.org/10.3389/fpsyg.2019.01788.
134. Crisp NA, Fournier PA, Licari MK, Braham R, Guelfi KJ. Optimising sprint interval exercise to maximise energy expenditure and enjoyment in overweight boys. Appl Physiol Nutr Metab. 2012;37(6):1222–31. https://doi.org/10.1139/h2012-111.
135. Sabag A, Little JP, Johnson NA. Low-volume high-intensity interval training for cardiometabolic health. J Physiol. 2022;600(5):1013–26. https://doi.org/10.1113/JP281210.
136. Tjønna AE, Leinan IM, Bartnes AT, Jenssen BM, Gibala MJ, Winett RA, Wisløff U. Low- and high-volume of intensive endurance training significantly improves maximal oxygen uptake after 10-weeks of training in healthy men. PLoS One. 2013;8(5):e65382. https://doi.org/10.1371/journal.pone.0065382.
137. Ramos JS, Dalleck LC, Borrani F, Beetham KS, Wallen MP, Mallard AR, Clark B, Gomersall S, Keating SE, Fassett RG, Coombes JS. Low-volume high-intensity interval training is sufficient to ameliorate the severity of metabolic syndrome. Metab Syndr Relat Disord. 2017;15(7):319–28. https://doi.org/10.1089/met.2017.0042.
138. Oh S, So R, Shida T, Matsuo T, Kim B, Akiyama K, Isobe T, Okamoto Y, Tanaka K, Shoda J. High-intensity aerobic exercise improves both hepatic fat content and stiffness in sedentary obese men with nonalcoholic fatty liver disease. Sci Rep. 2017;7:43029. https://doi.org/10.1038/srep43029.
139. Winding KM, Munch GW, Iepsen UW, Van Hall G, Pedersen BK, Mortensen SP. The effect on glycaemic control of low-volume high-intensity interval training versus endurance training in individuals with type 2 diabetes. Diabetes Obes Metab. 2018;20(5):1131–9. https://doi.org/10.1111/dom.13198.
140. Abdelbasset WK, Tantawy SA, Kamel DM, Alqahtani BA, Elnegamy TE, Soliman GS, Ibrahim AA. Effects of high-intensity interval and moderate-intensity continuous aerobic exercise on diabetic obese patients with nonalcoholic fatty liver disease: a comparative randomized controlled trial. Medicine. 2020;99(10):e19471. https://doi.org/10.1097/MD.0000000000019471.
141. Poon ET, Little JP, Sit CH, Wong SH. The effect of low-volume high-intensity interval training on cardiometabolic health and psychological responses in overweight/obese middle-aged men. J Sports Sci. 2020;38(17):1997–2004. https://doi.org/10.1080/02640414.2020.1766178.
142. Sabag A, Way KL, Sultana RN, Keating SE, Gerofi JA, Chuter VH, Byrne NM, Baker MK, George J, Caterson ID, Twigg SM, Johnson NA. The effect of a novel low-volume aerobic exercise intervention on liver fat in type 2 diabetes: a randomized controlled trial. Diabetes Care. 2020;43(10):2371–8. https://doi.org/10.2337/dc19-2523.
143. Ryan BJ, Schleh MW, Ahn C, Ludzki AC, Gillen JB, Varshney P, Van Pelt DW, Pitchford LM, Chenevert TL, Gioscia-Ryan RA, Howton SM, Rode T, Hummel SL, Burant CF, Little JP, Horowitz JF. Moderate-intensity exercise and high-intensity interval training affect insulin sensitivity similarly in obese adults. J Clin Endocrinol Metab. 2020;105(8):e2941–59. https://doi.org/10.1210/clinem/dgaa345.
144. Matsuo T, Saotome K, Seino S, Shimojo N, Matsushita A, Iemitsu M, Ohshima H, Tanaka K, Mukai C. Effects of a low-volume aerobic-type interval exercise on VO_2max and cardiac mass. Med Sci Sports Exerc. 2014;46(1):42–50. https://doi.org/10.1249/MSS.0b013e3182a38da8.
145. Wilson GA, Wilkins GT, Cotter JD, Lamberts RR, Lal S, Baldi JC. HIIT improves left ventricular exercise response in adults with type 2 diabetes. Med Sci Sports Exerc. 2019;51(6):1099–105. https://doi.org/10.1249/MSS.0000000000001897.
146. Way KL, Sabag A, Sultana RN, Baker MK, Keating SE, Lanting S, Gerofi J, Chuter VH, Caterson ID, Twigg SM, Johnson NA. The effect of low-volume high-intensity interval training on cardiovascular health outcomes in type 2 diabetes: a randomised controlled trial. Int J Cardiol. 2020;320:148–54. https://doi.org/10.1016/j.ijcard.2020.06.019.
147. Schulz KF, Altman DG, Moher D, CONSORT Group. CONSORT 2010 Statement: updated guidelines for reporting parallel group randomised trials. J Clin Epidemiol. 2010;63(8):834–40. https://doi.org/10.1016/j.jclinepi.2010.02.005.
148. Pattyn N, Beulque R, Cornelissen V. Aerobic interval vs. continuous training in patients with coronary artery disease or heart failure: an updated systematic review and meta-analysis with a focus on secondary outcomes. Sports Med. 2018;48(5):1189–205. https://doi.org/10.1007/s40279-018-0885-5.
149. Bonafiglia JT, Islam H, Preobrazenski N, Gurd BJ. Risk of bias and reporting practices in studies comparing VO_2max responses to sprint interval vs. continuous training: a systematic review and meta-analysis. J Sport Health Sci. 2021. https://doi.org/10.1016/j.jshs.2021.03.005.
150. Charles P, Giraudeau B, Dechartres A, Baron G, Ravaud P. Reporting of sample size calculation in randomised controlled trials: review. BMJ. 2009;338: b1732. https://doi.org/10.1136/bmj.b1732.
151. Abt G, Boreham C, Davison G, Jackson R, Nevill A, Wallace E, Williams M. Power, precision, and sample size estimation in sport and exercise science research. J Sports Sci. 2020;38(17):1933–5. https://doi.org/10.1080/02640414.2020.1776002.
152. Cheval B, Boisgontier MP. The theory of effort minimization in physical activity. Exerc Sport Sci Rev. 2021;49(3):168–78. https://doi.org/10.1249/JES.0000000000000252.
153. Ioannidis JP. How to make more published research true. PLoS Med. 2014;11(10): e1001747. https://doi.org/10.1371/journal.pmed.1001747.
154. Fanelli D. Redefine misconduct as distorted reporting. Nature. 2013;494(7436):149. https://doi.org/10.1038/494149a.
155. Gibala MJ. High-intensity interval training: a time-efficient strategy for health promotion? Curr Sports Med Rep. 2007;6(4):211–3.
156. Gibala MJ, McGee SL. Metabolic adaptations to short-term high-intensity interval training: a little pain for a lot of gain? Exerc Sport Sci Rev. 2008;36(2):58–63. https://doi.org/10.1097/JES.0b013e318168ec1f.
157. Satiroglu R, Lalande S, Hong S, Nagel MJ, Coyle EF. Four-second power cycling training increases maximal anaerobic power, peak oxygen consumption, and total blood volume. Med Sci Sports Exerc. 2021;53(12):2536–42. https://doi.org/10.1249/MSS.0000000000002748.
158. Sato S, Yoshida R, Murakoshi F, Sasaki Y, Yahata K, Nosaka K, Nakamura M. Effect of daily 3-s maximum voluntary isometric, concentric, or eccentric contraction on elbow flexor strength. Scand J Med Sci Sports. 2022;32(5):833–43. https://doi.org/10.1111/sms.14138.

159. Wagenmakers EJ. Defiant denial is self-defeating. Psychol Inq. 2021;32(1):12–6. https://doi.org/10.1080/1047840X.2021.1889314.
160. Harris C, Rohrer D, Pashler H. A train wreck by any other name. Psychol Inq. 2021;32(1):17–23. https://doi.org/10.1080/1047840X.2021.1889317.
161. Sherman JW, Rivers AM. There's nothing social about social priming: derailing the "train wreck." Psychol Inq. 2021;32(1):1–11. https://doi.org/10.1080/1047840X.2021.1889312.
162. Wiggins BJ, Christopherson CD. The replication crisis in psychology: an overview for theoretical and philosophical psychology. J Theoret Philos Psychol. 2019;39(4):202–17. https://doi.org/10.1037/teo0000137.
163. Pashler H, Wagenmakers EJ. Editors' introduction to the special section on replicability in psychological science: a crisis of confidence? Perspect Psychol Sci. 2012;7(6):528–30. https://doi.org/10.1177/1745691612465253.
164. Yong E. Nobel laureate challenges psychologists to clean up their act. Nature. 2012. https://doi.org/10.1038/nature.2012.11535.
165. Rodgers JL, Shrout PE. Psychology's replication crisis as scientific opportunity: a précis for policymakers. Policy Insights Behav Brain Sci. 2018;5(1):134–41. https://doi.org/10.1177/2372732217749254.
166. Sharpe D, Goghari VM. Building a cumulative psychological science. Can Psychol. 2020;61(4):269–72. https://doi.org/10.1037/cap0000252.
167. Shrout PE, Rodgers JL. Psychology, science, and knowledge construction: broadening perspectives from the replication crisis. Annu Rev Psychol. 2018;69:487–510. https://doi.org/10.1146/annurev-psych-122216-011845.
168. Munafò MR, Nosek BA, Bishop DVM, Button KS, Chambers CD, du Sert NP, Simonsohn U, Wagenmakers EJ, Ware JJ, Ioannidis JPA. A manifesto for reproducible science. Nat Hum Behav. 2017;1:0021. https://doi.org/10.1038/s41562-016-0021.
169. Nelson LD, Simmons J, Simonsohn U. Psychology's renaissance. Annu Rev Psychol. 2018;69:511–34. https://doi.org/10.1146/annurev-psych-122216-011836.
170. Begley CG, Ellis LM. Drug development: raise standards for preclinical cancer research. Nature. 2012;483(7391):531–3. https://doi.org/10.1038/483531a.
171. Begley CG, Ioannidis JP. Reproducibility in science: improving the standard for basic and preclinical research. Circ Res. 2015;116(1):116–26. https://doi.org/10.1161/CIRCRESAHA.114.303819.
172. Impellizzeri FM, McCall A, Meyer T. Registered reports coming soon: our contribution to better science in football research. Sci Med Football. 2019;3(2):87–8. https://doi.org/10.1080/24733938.2019.1603659.
173. Nosek BA, Ebersole CR, DeHaven AC, Mellor DT. The preregistration revolution. Proc Natl Acad Sci USA. 2018;115(11):2600–6. https://doi.org/10.1073/pnas.1708274114.
174. Nosek BA, Hardwicke TE, Moshontz H, et al. Replicability, robustness, and reproducibility in psychological science. Annu Rev Psychol. 2022;73:719–48. https://doi.org/10.1146/annurev-psych-020821-114157.
175. Scheel AM, Schijen MRMJ, Lakens D. An excess of positive results: comparing the standard psychology literature with registered reports. Adv Meth Pract Psychol Sci. 2021. https://doi.org/10.1177/25152459211007467.
176. Schäfer T, Schwarz MA. The meaningfulness of effect sizes in psychological research: differences between sub-disciplines and the impact of potential biases. Front Psychol. 2019;10:813. https://doi.org/10.3389/fpsyg.2019.00813.
177. Abt G, Boreham C, Davison G, Jackson R, Wallace E, Williams AM. Registered reports in the journal of sports sciences. J Sports Sci. 2021;39(16):1789–90. https://doi.org/10.1080/02640414.2021.1950974.