Replication concerns in sports science: a narrative review of selected methodological issues in the field

Cristian Mesquida[1], Jennifer Murphy[1], Daniël Lakens[3], Joe Warne[1,2]

[1]Centre of Applied Science for Health, Technological University Dublin, Tallaght, Dublin, Ireland
[2]Setanta College, Thurles Chamber of Commerce, Tipperary, Ireland.
[3]Human-Technology Interaction Group, Eindhoven University of Technology, Eindhoven, The Netherlands

ORCIDs
Cristian Mesquida – 0000-0002-1542-8355
Jennifer Murphy – 0000-0001-8624-3828
Daniël Lakens – 0000-0002-0247-239X
Joe Warne – 0000-0002-4359-8132

Correspondence Cristian Mesquida; Centre of Applied Science for Health**,** Technological University Dublin at Tallaght, Dublin, D24 FKT9, Ireland; X00180647@mytudublin.ie

Statements
All authors have read and approved this version of the manuscript.
This is a preprint, not a peer reviewed manuscript.

Please cite as: Mesquida, C., Murphy, J., Lakens, D., & Warne, J. (2022). Replication concerns in sports science: a narrative review of selected methodological issues in the field. *SportRχiv*

Abstract
Known methodological issues such as publication bias, questionable research practices (QRPs) and studies with underpowered designs are known to decrease the replicability of scientific findings. The presence of such issues has been widely established across different research fields, especially in psychology. Their presence raised the first concerns that the replicability of scientific findings could be low and led researchers to conduct large replication projects. These replication projects revealed that a significant portion of original studies could not be replicated, giving rise to the conceptualization of the Replication Crisis. Although previous research in the field of sports and exercise science has identified the first warning signs, such as an overwhelming proportion of significant findings, small sample sizes and lack of open science practices, their possible consequences for the replicability of our field have been overlooked. Furthermore, the presence of publication bias, QRPs and studies with underpowered designs, which are known to increase the number of false positives in a body literature, has yet to be examined. In this review we aim to explore the prevalence of these issues by conducting a z-curve analysis in applied studies published in the *Journal of Sports Sciences*. Overall, we have observed evidence of publication bias and studies with underpowered designs raising the possibility that a portion of findings in our field may not replicate. We discuss the consequences of the above issues on the replicability of our field and offer potential solutions to improve replicability.

Key Points

Several methodological issues such as publication bias, questionable research practices, studies with underpowered designs and lack of open science practices have the potential to hinder the replicability of sports and exercise science findings.

We provide evidence of publication bias and studies with underpowered designs in a large set of studies within one journal. It is likely that a portion of published findings cannot be replicated due to the above practices, which are likely to increase the number of false positives and produce overestimated effect sizes and wide confidence intervals.

We advocate for the adoption of Open Science practices including preregistration/Registered Reports, sample size planning based on power or accuracy of the parameter estimate, data sharing, and replications.

1. Introduction

Null hypothesis significance testing (NHST) is a method of statistical inference where the probability of observed or more extreme data is compared against a null model (H_0). In a Neyman-Pearson approach to hypothesis testing, the *p*-value is compared with a pre-established significance threshold to either corroborate or reject the H_0. One interesting observation is that over 90% of published articles using NHST in biomedicine and psychology reported statistically significant results (i.e., $p < .05$) [1–3]. Similarly, it has been observed that between 70% and 82% of published articles in sports science journals reported statistically significant results [4,5]. One conclusion that can be drawn based on this data is that researchers in these disciplines plan and design experiments that usually reject H_0 because their studies examine predominantly true effects with high statistical power (henceforth, power).

However, it is unlikely that the high proportion of significant results in these fields are solely due to high quality research designs and testing true effects. One key fact that should render researchers skeptical about the replicability of prior findings is when a literature body produce more significant results than expected based on the power of the studies [3,6]. For instance, while in psychology over 90% of published studies reported significant results, the average power to detect a medium effect size (ES) has been estimated to barely reach 50% [7,8] or even lower [9,10]. An excess of significant results is problematic, and indicates that other factors play a role that bias the proportion of significant results in the published literature. Three main factors identified in the literature are publication bias, including reviewer bias and the file-drawer problem [11,12]; questionable research practices (QRPs), including HARKing and *p*-hacking [8,12–14; see also 15,16 for researchers' degrees of freedom], and studies with underpowered designs [8,9,18,19], among others [20–22] (see **Table 1** for definitions). Together, these factors contribute to the probability that a published statistically significant research finding is actually a false positive, and consequently, the systematic presence of these issues in a body of literature is likely to hinder its replicability.

Table 1 Definitions of key concepts

Excess of significance findings The phenomenon whereby a set of studies produce a higher percentage of significant results than it should be expected given the average power of these studies.
Statistical power The probability of a statistical test rejecting H_0 when it is false, that is, the probability of obtaining a significant result. It depends on the given effect size in the population, the chosen significance level, and the number of participants tested [23].
Publication bias It is publishing behaviors that gives studies which find support for their tested hypotheses a higher chance of being published, as opposed to the publication of replication studies and non-significant results. These behaviors include editors and reviewers selectively publishing studies with significant findings (i.e., review bias; [11]) and researchers deciding not to submit studies with non-significant results (i.e., file-drawering) [12].
Questionable Research Practices (QRPs) QRPs describe a set of research behaviours that can spuriously increase the probability of finding evidence in support of a hypothesis [14]. Some forms of QRPs are HARKing and *p*-hacking [13,14].
HARKing A form of QRP that involves the post hoc formulation of the Hypothesis After the Results are Known [15].
***P*-hacking** A form of QRPs that exploits flexibility in data analysis to obtain statistically significant results [14]. Examples of *p*-hacking include optional stopping, the inclusion or exclusion of data on the basis of *post hoc* criteria and multiple testing [13,14].
Observed Discovery Rate (ODR) The relative frequency of significant results [24]. If a set of 10 studies produce 4 significant results, the ODR is 40%. The ODR does not distinguish between true and false findings.
Expected Discovery Rate (EDR) The percentage of statistically significant results that were obtained in all studies. In other words, the average power of all studies with both significant and non-significant results. EDR can be used to assess the size of the file-drawer problem and estimate the maximum number of false positive results [24].

Expected Replication Rate (ERR)
The average power of only significant results. It represents the estimated proportion of significant studies that would yield another significant effect if subjected to a direct replication [24].

These aforementioned issues raise concerns about the credibility of scientific findings and sparked interest in replicability across scientific fields, especially in psychology [25–29]. One of the first attempts to systematically replicate original studies was the Open Science Collaboration Project [25] which set out to replicate 100 primary findings published in three high-impact psychology journals; strikingly, although 97% of the original studies reported significant results, only 37% of the replication studies yielded a statistically significant result. This project was followed by other replication attempts in psychology [30], social sciences [26] and economics [27] with replication rates of 54%, 62% and 61%, respectively. Despite these developments in other fields, replication studies are still very rare in sports science [31]. This might be in part not only due to the difficulties in conducting replication studies observed across disciplines [26,31,32], but also due to particular features of sports science research. Firstly, it is practically impossible to conduct replications of original studies that require long-term observations/interventions (e.g., multiple exposures to altitude training), expensive equipment and samples with unusual traits (e.g., elite athletes). Secondly, replication studies may require expertise that only a few researchers have, such as the study of motoneuron adaptations to resistance training by using high-density electromyography analysis [33]. Finally, limited availability of original raw data, inaccurate explanation of procedures or methods, and poor reporting practices in the original study hinder the accuracy of replication studies. Before performing a large-scale replication project in sports science, it seems reasonable to first evaluate the extent to which methodological issues that may influence the replicability of the published literature.

To date, few studies have investigated the presence of the aforementioned methodological issues in sports science [4,34–37]. Their findings have raised the first warning signs that our scientific field is likely to face a problem with replicability due to an overwhelming proportion of significant findings, small sample sizes and lack of data availability [4,34–37]. However, the specific presence of methodological issues such as publication bias, QRPs, and studies with underpowered designs, which are known to increase the number of false positives in the published literature, has yet to be examined. One method to examine the presence of the above issues is by conducting a z-curve analysis. Briefly, the z-curve method is a meta-analytic tool that estimates two interpretable measures (i.e., Expected Discovery Rate and Expected Replication Rate; see **Table 1** for definitions) for the reliability of scientific literature based on test-statistics of published studies [38,39]. The purpose of the current review is twofold. First, we aim to explore the prevalence of publication bias and studies with underpowered designs by conducting a z-curve in applied studies published in the *Journal of Sports Sciences*. Second, we aim to discuss the potential consequences of these aforementioned methodological issues on the replicability of sports science findings and offer potential solutions to improve it. We hope that this review will encourage other researchers to examine the presence of these and other methodological issues in larger literature bodies, conduct replication studies where needed, and increase the adoption of Open Science practices, such as conducting pre-study power calculations, and making research data available to facilitate replicability.

2. Methodological issues

In line with previous findings in biomedicine and psychology [2,40], Büttner et al. [4] reported that out of 129 studies from sports and exercise medicine journals, 106 (82.2%) reported statistically significant results [4]. For this percentage to be a true representation of the studies performed in the field, both power and the proportion of true hypothesis tested must exceed 80% [6]. In other words, nearly all hypotheses that sports scientists test must examine a true effect, and either the effects investigated or the sample sizes used must be consistently large to achieve the desired power (i.e., ≥ 80%) [6]. In the following sections we will discuss why 82% significant results in the literature should be interpreted with caution.

2.1. Publication bias and questionable research practices

One way to objectively examine the reliability of a set of findings is to quantify the evidential value of a body of literature [41]. Evidential value is determined by the number of studies examining true and false effects, the power of the studies that examine true effects, the frequency of type I error rates (and how they are inflated by *p*-hacking) and publication bias [42–44]. Fortunately, issues relating to power of the studies, *p*-hacking, and publication bias can be explored via the distribution of reported *p*-values [43,44]. For example, when H_0 is true (i.e., there is no significant effect to be found) *p*-values between a [0, 1] interval should be equally likely in a two-sided hypothesis test regardless of the sample size, yielding a uniform distribution [42,44,45] (**Fig. 1**). In other words, when the null-hypothesis is true, a *p*-value of 0.01 is just about as likely to be observed as a *p*-value of 0.9.

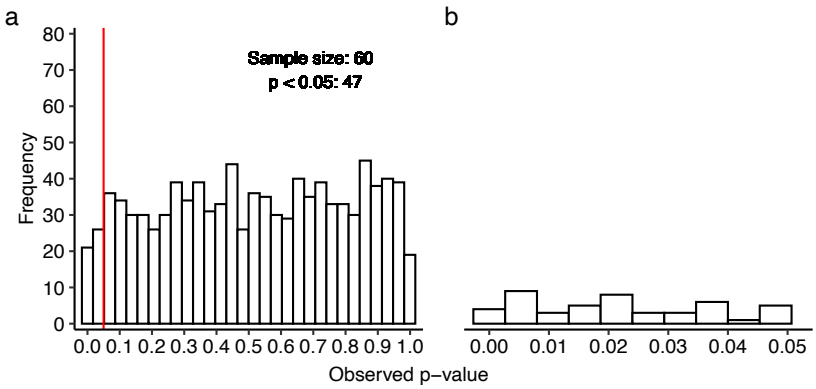

Fig. 1 Distribution of *p*-values over: **a** [0-1] interval and **b** over [0-0.05] interval when H_0 is true. 1000 *p*-values were generated for simulated comparisons with an unpaired *t*-test for statistical difference between two samples of 60 participants each. Red line denotes statistical significance at $p < 0.05$ and the number of significant p-values representing type I errors

However, when the alternative hypothesis (H_1) is true, the distribution of *p*-values becomes a function of power, thus, the study sample size and the true (but always unknown) effect size [45,46]. Sample size is therefore an important factor when evaluating the distribution of *p*-values in a literature. Suppose there is a true difference between two populations with a Cohen's *d* ES (ES *d*) of 0.5 and we perform an unpaired *t*-test to test this difference in three different sample sizes (i.e., 10, 30 and 60 participants per group). As we can see in **Fig. 2a**, a sample size of 10 per group and a true ES of 0.5 yields a power of 18%, which means that out of 1000 replications, only 180 should be expected to reach statistical significance (in the long run), even though there is a true effect to be found.

With a sample size of 60 participants per group, power is as high as 78%, meaning that 780 out of 1000 replications reach statistical significance in the long run (**Fig. 2c**). In studies with high power and where a true effect is examined, the likelihood of observing a small *p*-value (e.g., $p = 0.01$) is higher compared to a large *p*-value (e.g., $p = 0.4$) [45,46]. Moreover, as power increases even more, most of the p-values are below 0.01 and there are relatively fewer *p*-values between 0.01 and 0.05 (**Fig. 2**). For instance, while there are 235 *p*-values below 0.01 with a power of 48%, there are as many as 562 with a power of 78%. Consequently, the *p*-value distribution (in sufficiently powered research) follows a right-skewed distribution which decreases monotonically in the absence of *p*-hacking and publication bias [47]. For this reason, the distribution of *p*-values can be used not only to determine whether a set of homogeneous studies investigates true or false effects, but it can also be used to estimate the average power of the set of studies. Altogether, it should be clear that the small sample sizes observed in sports science [34,37] may be a reason for concern given the high proportion of significant findings that are observed [4,5].

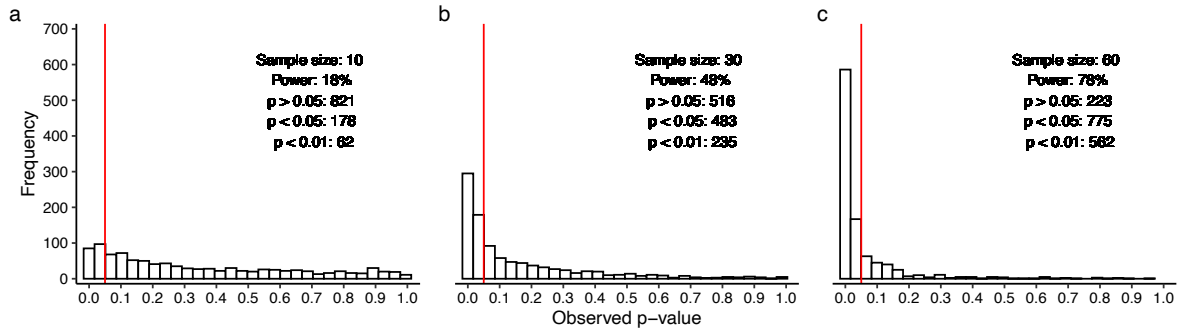

Fig. 2 Power affects the distribution of *p*-values when H_0 is false. 1000 *p*-values were generated for simulated comparisons with an unpaired *t*-test for each sample size. The number of *p*-values below 0.05 and 0.01 and above 0.05 are shown. The power is the percentage of simulations in which the *p*-value reaches significance (i.e., $p < 0.05$) given that H_1 is true. Vertical red line denotes statistical significance at $p < 0.05$

Whilst the above assumes an unbiased frequentist observation, one explanation for an excess of significant findings in a set of studies that has been raised in the literature is publication bias and *p*-hacking [14,48,49]. In the presence of publication bias (where non-significant results are less likely to get published), researchers have incentives to explore *post hoc* analyses to find a significant *p*-value (i.e., *p*-hacking). If *p*-hacking occurs in a literature, the distribution of reported significant *p*-values adopts different shapes [42]; For instance, when researchers resort to optional stopping (when H_0 is true), the distribution of reported significant *p*-values is right-skewed (i.e., there will be a greater number of *p*-values between .04 and .05 than between .00 and .01; see **Fig. 3**). The *p*-value distribution can also be used to examine a bias to publishing statistically significant results. The lack of a continuous distribution of *p*-values below the default alpha level of 0.05 and above this threshold indicates the presence of bias in favour of statistically significant results in the published literature, and the presence of a file-drawer. Therefore, by examining the distribution of *p*-values, it can be determined whether published findings contain evidential value of a true effect, and the extent to which findings in the literature are affected by publication bias and/or *p*-hacking [43,44].

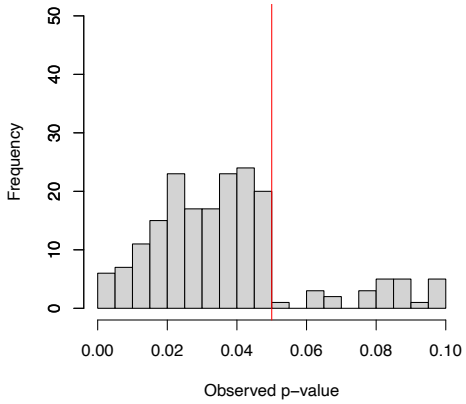

Fig. 3 Distribution of *p*-values over [0, 0.1] interval when H_0 is true but in the presence of *p*-hacking. This would reflect the influence of collecting 10 participants and conduct an unpaired *t*-test after each addition until 100 participants are collected. Red line denotes statistical significance at $p < 0.05$

2.1.1. Z-curve

One method to assess the evidential value of a body of literature is a z-curve [38,39]. Z-curve is based on the concept that the average power of a set of studies can be derived from the distribution of z-scores. Z-curve converts significant and non-significant *p*-values reported in a literature into z-scores, and uses the distribution of z-scores within the range of 0 to 6 to calculate two estimates of average power (see [39] for details). First, the conditional mean power is computed by using only the *significant* results in the published studies. By using this estimate of average power, it is possible to calculate the Expected Replication Rate, that is, the expected success rate (in the long run) if these studies would be exactly replicated. Second, the unconditional average power is computed, which is an estimate of the power in studies that were not published because these studies yielded statistically non-significant findings, and remained in the file-drawer. The presence of publication bias can be calculated from the unconditional average power by examining whether the Estimated Discovery Rate is lower than the Observed Discovery Rate. Another way to examine the presence of publication bias is by using the Estimated Discovery Rate to calculate the file-drawer ratio which estimates how many non-significant results there might be for each significant result [38].

2.1.2. Methods

Given the high percentage of significant findings reported in sports and exercise medicine journals [4,5], we used the *z-curve* package in R to conduct a z-curve analysis [50] to examine the presence of publication bias in a set of studies (N = 119) published in the *Journal of Sports Sciences*. We selected this journal in light of the findings from Abt et al. [34] who reported that this journal published studies with a median sample size of 19. A study using a sample of 19 participants may not have sufficient power, especially to detect small and medium ES in between-subject designs [51]. The selection protocol for the studies to be included in the z-curve analysis is based on the *Selection Protocol for Replication in Sports and Exercise Science* [52]. Hence, only applied sports and exercise science studies in the subdisciplines of physiology, sports performance, physical activity, injury prevention and psychology published in the *Journal of Sports Sciences* (from Volume 39 (Issue 12) to Volume

37 (Issue 16)) were selected. Furthermore, applied studies had to use either an experimental or quasi-experimental design. Studies were selected if they contained an inference test such as a *t*-test and *F*-test. Z-curve method uses all *p*-values regardless of whether the *p*-value is yielded by a non-parametric test (i.e., Wilcoxon Rank-Sum tests, Mann-Whitney-U-Tests or Kruskal-Wallis one-way ANOVA). Therefore, *p*-values derived from the above non-parametric tests were also included. After study selection, only one *p*-value per independent experiment was extracted in order to meet the independence criteria [53]. The extracted *p*-value corresponded to the first or primary dependent variable stated in either the hypothesis or, in its absence, in the study aim. In case where there were multiple hypotheses/aims, the first or primary hypothesis/aim was considered. If the selected hypothesis/aim included multiple dependent variables, the first or primary dependent variable was considered. In case the selected dependent variable was operationalized using several outcome measures of the same construct (i.e., to be measured in several alternative ways), the first outcome measure reported was selected. Extracted *p*-values were recomputed when sufficient information was available (i.e., degrees of freedom and *F*-ratio or *t*-statistic). *P*-values were discarded under 5 circumstances; first, when the *p*-value was reported relatively (e.g., $p < 0.05$) and it could not be recomputed due to lack of sufficient information. Second, when studies tested an hypothesis for non-significance. Third, the described statistical test in the methods did not match the statistical test reported in the results section of the study. Fourth, the study did not report the effect of interest given the hypothesis stated in the introduction. Finally, the study expected to find a significant difference in one direction but occurred in the other one; the inclusion of this type of significant *p*-value in z-curve would be problematic because it could create bias in favor of statistical significance. The data extraction and coding was conducted by the primary author (CM) of the study and random verification of the coded data was carried out by a secondary author (DL). The materials including the study selection protocol, sample of studies, dataset generated and R code for z-curve are available at https://osf.io/y3482/.

2.1.3. Results

The results from the z-curve are presented as mean [95% CI] in **Fig.4**. The Observed Discovery Rate was 0.69 [0.60, 0.77], indicating that 69% of the studies reported significant findings. This is in agreement with Twomey et al. [5] who found that approximately 70% of the studies published in three flagship sports science journals reported significant findings. In addition, the point estimate of the Observed Discovery Rate (0.69) lies within the 95% CI of the Expected Discovery Rate of 0.48 [0.10, 0.70]; this suggests that, even though there is a discrepancy, there is not enough data to statistically reject the hypothesis that there is no evidence of publication bias. However, the confidence interval is rather wide due to the small sample of studies included from *Journal of Sports Sciences*, and absence of evidence is not evidence of absence. A visual inspection of the obtained results suggests there is a potential indication of publication bias (see **Fig. 4**); there is a steep drop from just statistically significant values (i.e., $z > 1.96$) compared with non-significant values. This figure suggests that, even when publication bias might not be extreme (i.e., a reasonable proportion of non-significant findings are published in this literature) there are still relatively less *p*-values just above the traditional alpha level of 5% than below this threshold. This provides tentative evidence for the presence of publication bias. One additional point of evidence to support this contention is that the file-drawer ratio has a point estimate of 1.1 [0.45, 8.61], and this indicates that the ratio between published and unpublished literature is approximately even, i.e., for each published significant result, there is on average 1.1 non-significant results in the "file-drawer" not published, which is concerning. The presence of

publication bias (including the file-drawer problem) has two negative consequences. First, publishing non-significant findings based on studies with high statistical power for the smallest effect of interest can be productive for refuting or challenging our current understanding of a given theory or phenomena. Second, meta-analyses are useful to synthetize a set of ES from studies investigating the same phenomena by providing an estimate of the true ES, and reducing the uncertainty around the confidence interval (CI). However, the presence of publication bias can result in overestimated ES, and a skew towards positive outcomes, thus undermining the evidential value of meta-analyses [9,54,55].

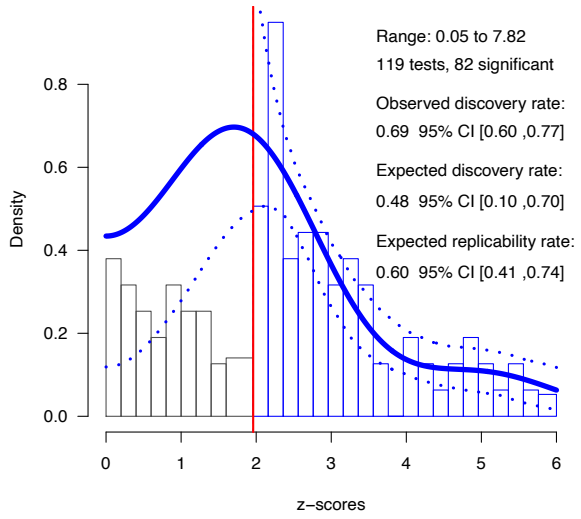

Fig. 4 Distribution of z-scores over [0-6] interval. The vertical red line refers to a z-score of 1.96, the critical value for statistical significance when using a two-tailed alpha of 0.05. The dark blue line is the density distribution for the inputted *p*-values (represented in the histogram as z-scores). The dotted lines represent the 95% CI for the density distribution. Range represents the minimum and maximum values of z-scores used to fit the z-curve. A total of 119 independent *p*-values (including 82 significant and 37 non-significant *p*-values) were converted into z-scores to fit the z-curve model.

2.2. Power

In a Neyman-Pearson approach, researchers should use the NHST framework under the assumption of two conditions [56]. First, H_0 should be plausible enough so that its rejection might be unexpected. Second, researchers should be willing to make a decision of a scientific claim for which the type I and type II error rates are adequately controlled. Researchers can limit the frequency of type I and type II errors by choosing the alpha level and conducting studies with high-power designs for effects of interest given that the type II error rate is defined as 1 – power (the higher the power, the lower the type II error rate). To ensure that studies have well-powered designs, researchers should conduct pre-study power calculations for a given sample size and ES (**Fig. 5**). The value of this approach is discussed below.

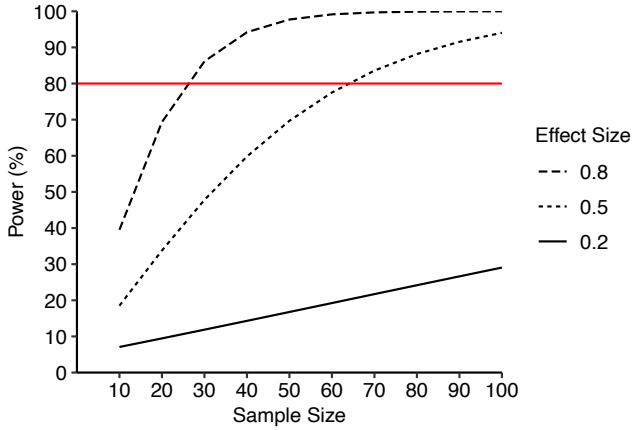

Fig. 5 Power of an unpaired *t*-test given a range of sample sizes and ES *d*. Red line denotes an adequate power of 80%.

2.2.1. Estimating power in sports sciences

Power has direct implications on replicability because, from a frequentist standpoint, power is also described as the long-run probability of obtaining a significant effect when there is a true effect to be found [57]. To date, most researchers are familiar with Cohen's suggestion [58] that experiments should have at least 80% power. Hence, a study is typically considered adequately powered if it finds a significant effect in 8 out of 10 replications when there is a true effect to be found (although one might argue that whenever feasible, a higher statistical power is desired). Moreover, according to Fisher [59], a good experiment should rarely produce a non-significant result when H_0 is false [59]. Therefore, if studies examining true effects are designed with high power, any researcher is more likely to find the same effect when replicating the same procedures with adequate power.

There is, however, concern that studies in the sports sciences are adequately powered for effects of interest. It is again worth highlighting the findings from two recent studies [4,34]; the high proportion (82.2%) of significant findings [4] and the small median sample sizes (N = 19) reported in the *Journal of Sports Sciences* [34] seem to indicate that, unless all examined effects are large, there might be relatively low power. As we will discuss in the following section, a median sample size of 19 is likely to yield underpowered designs, especially to detect small and medium ES. The main implication of underpowered designs is that the literature should be filled with a higher proportion of null findings since the published studies would have a low probability of detecting the studied effect [60], but this is not the reality. To our best knowledge, only one study has assessed the power of a set of studies in our field [61]. This study estimated the median observed power of 108 significance tests from 29 articles using fixed ES based on Cohen's benchmarks [58]. The median observed power was 14%, 65% and 97% for small, medium and large ES, respectively. Furthermore, moving beyond the median power, and looking at individual studies, it was found that no studies had adequate power to detect small ES, only 38% of studies had adequate power to detect a medium ES, and about 75% of studies had a power of at least 80% to detect large ES. However, one limitation of this method was the use of fixed ES based on Cohen's benchmarks which are derived from effects observed in behavioral science [58]. It is uncertain whether Cohen's benchmarks accurately represent ES observed in any given subfield of sport science [61–63]. For instance, Swinton [64] conducted a Bayesian hierarchical meta-analysis to identify specific ES benchmarks in strength and conditioning interventions and

reported that the benchmarks for small, medium and large ES were 0.12, 0.43 and 0.78, respectively. Therefore, sports science researchers should avoid the use of ES based on Cohen's benchmarks and use specific ES derived from related studies when attempting to estimate the sample size required given the ES of interest and the intended power.

To further elaborate, we provide observed power estimates in our field using two methods based on i) a typical ES estimate and sample size reported in previous research [34,65] and ii) z-curve method [24]. R code used for both methods is available at https://osf.io/y3482/. Regarding the first method using a specific ES benchmark and sample sizes in our field, there is reason for caution because of the use of small sample sizes in our field [34,37]. Four biomechanics and sports science journals had a mean sample size (standard deviation) of 21 (24),15 (19), 32 (32) and 20 (22) (of 188 articles published in 2009 [37]). Similarly, the *Journal of Sports Sciences* had a median sample size of 19 for 120 articles published between 2016 and 2019 [34]. In line with this previous finding [34], we found a median sample size of 20 for 119 papers included in the z-curve analysis. To see how sample size affects observed power, we will use an ES *d* of 0.43 which has been reported to be the medium ES benchmark for effects observed in 679 strength and conditioning intervention studies [64]. Suppose we conduct an experiment to find a true ES *d* of 0.43 with a sample size of 20 for a paired *t*-test. The reason why a paired *t*-test was selected is because 58% of studies (69 out of 119) from our sample used a within-subject design. This within-subject design would yield a power of 45%, implying that if 10 replications were to be conducted, only about 5 would find a significant effect. It is worth noting that for achieving 80% power, a sample size of 44 would be needed if the true ES was $d = 0.43$. Small sample sizes might be appropriate if the true ES being estimated is large enough to be reliably observed in such samples [18]; for instance, ES estimates from strength and conditioning interventions might be much larger than those observed in sports performance research [62,63]. However, studies with small samples in combination with selective reporting of statistically significant results are susceptible to overestimate ES estimates [66]. This means one should be cautious about the observed large ES estimates in the literature, if small studies are the sole source of these estimates [18]. Given the small samples reported in biomechanics and sport science journals [34,37], it might therefore be hypothesized that sports science faces a problem with underpowered designs, especially to detect small and medium ES. However, it should be noted that within-subject designs have higher power compared to between-subject designs given an ES and sample size [51]. The extent to which within-subject designs can increase power compared to between-subject designs is given by the correlation between observations [51]. This is because correlation is typically positive and higher in within-subject designs compared to between-subject designs. Hence, the higher the correlation between observations, the higher the power achieved. Therefore, between-subject designs may potentially have even less power to detect the ES of interest than the power estimated from a within-subject design.

Following up on the results of the z-curve reported earlier, we provided estimates of observed power for the same set of studies. The Expected Discovery Rate was 0.48 [0.10, 0.70] indicating an average power of 48% for studies reporting both significant and non-significant results. The Expected Replication Rate was 0.60 [0.41, 0.74] indicating that studies reporting significant results have an average power of 60%. This suggests that if we were going to conduct direct replications (with the statistical power, with the same effect size and sample size) of the studies reporting significant findings, only 60% of these studies would yield another significant effect. This is an

important finding to take into consideration for researchers who plan to build on published findings, and suggests that researchers will need a larger sample size, on average, in replication studies than were used in original studies. Taken together, we have presented two methods to provide estimates of observed power which seem to indicate the use of underpowered designs in our field. In the following section, we discuss the consequences of underpowered designs.

2.2.2. Consequences of underpowered designs

Whilst low power in itself is caused by low sample size or small ES, or both, the consequences of low power should be emphasized here. Firstly, underpowered designs are less likely to find a true effect even if the effect exists at the population level [8,67]. This is because small sample sizes contain a high sampling variance and therefore are less likely to not contain the true population parameters. This is demonstrated in **Fig. 6**, where even though there is a true difference between population A and B (i.e., a true ES *d* of 0.5), two of three of the studies do not find a significant effect and thus commit a type II error.

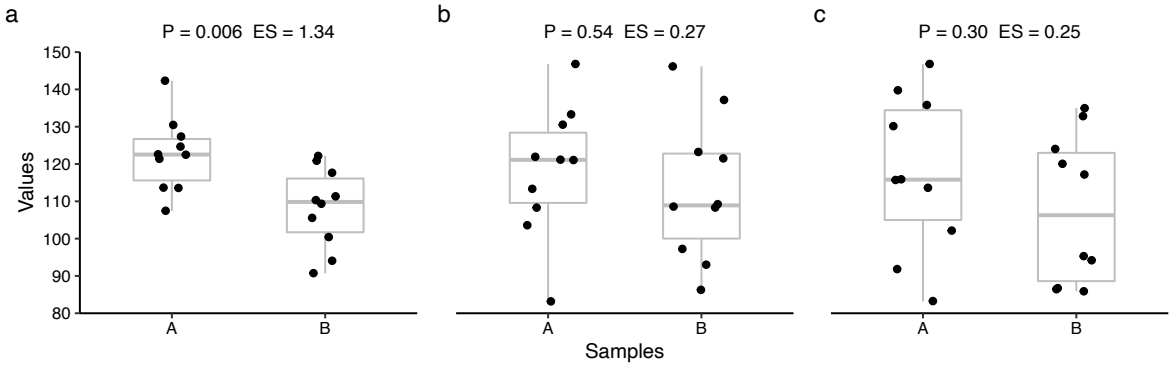

Fig.6 Small samples show substantial variation. To illustrate the variability of statistical outcomes derived from small samples, 6 samples of 10 values each at random were drawn from the same two populations as in **Fig. 2**. The true ES *d* between population A and B is 0.5. The estimated ES and *p*-value when sample pairs are compared are provided to demonstrate the variability of observed outcomes.

Secondly, underpowered designs also increase the proportion of false positives in a literature body where there is publication bias [8,67], which is known as the *positive predictive value*. To see how this plays out, let's assume that 20 sports science studies within the same scope have an average power of 45%, as we have calculated previously assuming a total sample size of 20 and a medium ES *d* of 0.43 for a paired *t*-test. In such a situation, approximately only 9 out of 20 studies (20 x 0.32) would find a significant effect even if all H_0 tested were false. The number of false positives with an alpha level of 0.05 would be 1 (20 x 0.05). Thus, the number of false positives relative to the total number of published significant findings is 10% (i.e., false positives/(false positives + correct hits) = 1 / (1 + 9)). On the other hand, let's consider how things would play out if the average power in a set of 20 studies is 80% instead of 20%. In this case, the number of significant findings when there is a true effect to be found would be 16 (20 x 0.8). Whilst the number of false positives would be the same (0.05 x 20 = 1), the proportion of false positives would be approximately 6% (1 / (1 + 16)). Comparatively speaking, although an unbiased literature can only be achieved by publishing all results, irrespective of the *p*-value, the reliability of a literature body is higher when the power is 80% rather than 20%. In fact, a set of underpowered studies investigating the same effect and all reporting significant findings is so unlikely that the findings become literally improbable [8]. Suppose that a set of 5 studies with an average power of 45% has reported significant effects

when H_0 was false. The probability of all 5 studies finding a significant effect would be 1.85% (0.45^5). Therefore, if the power observed in sports science studies is as low as hypothesized [34], we may expect an elevated number of false positives in sets of underpowered studies within the same scope. Given the observed high proportion of significant findings discussed [4], an elevated number of false positives seems a plausible explanation for a significant proportion of our research in this field.

Thirdly, the ES estimate provided by a study with an underpowered design in the presence of selection bias for significant results is likely to be overestimated [18,25,26,68]. As observed in **Fig. 6**, when a significance test has low power due to a small sample size, a significant ES will only be found when the ES is relatively extreme [68]. However, when power is augmented by taking more observations, the estimated ES becomes closer to the population value of the ES [68] (see **Fig.7**). For instance, both the Open Science Collaboration project [25] and the Social Science Replication Project [26] conducted replications with higher-power designs than the original studies; one of the main findings was that both replication projects observed that the mean ES of the replicated studies was approximately 50% of that reported in the original studies [25,26]. Because of the observed small sample sizes reported in sports sciences [34,37], it is likely that ES reported are overestimated, further compounding the issue with low power. Another consequence is that if published ES are overestimated and therefore do not reflect the true distribution of ES, meta-analyses are compromised [69].

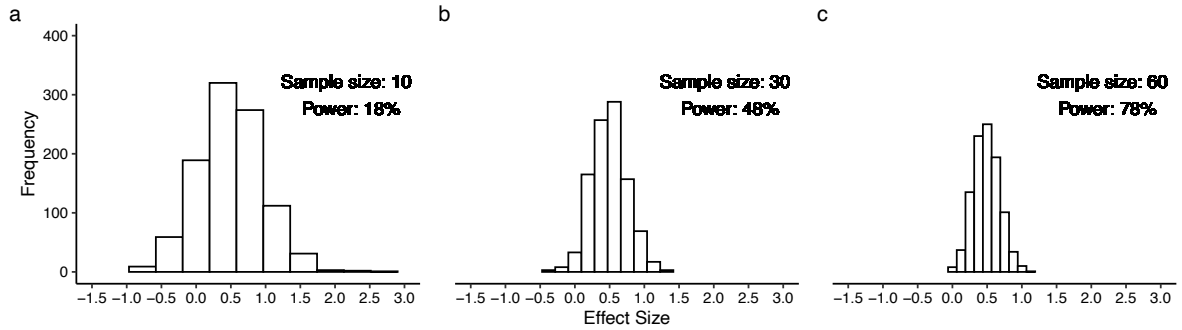

Fig. 7 Sample size affects the estimation of the true ES. Using the same data simulated as in **Fig. 2**, 1000 ES were computed. The histograms show the distribution of estimated ES for three different sample sizes. As sample size increases, the estimated ES becomes closer to the true ES *d* of 0.5.

In addition, the overestimation of ES is in itself a cause of concern when conducting pre-study power calculations [66,70]. The rationale for conducting a pre-study power calculation is to obtain an estimate of the sample size needed given any ES and intended power. However, if the ES is overestimated during power calculations, researchers may end up obtaining a smaller sample size estimate and thus eventually achieving less power than intended [66]. This is especially problematic when studies use small sample sizes and in the presence of publication bias because only overestimated ES will be published. For example, suppose a researcher wants to test the effect of a treatment on two independent samples and the true ES *d*, which is unknown, is 0.5. The researcher wants to obtain the sample size required to achieve 80% power and uses an overestimated ES *d* of 1.34 from a previous underpowered study (see **Fig. 6a**). Thus, the researcher finds out that a sample size of 20 (i.e., 10 participants per group) is needed to achieve 80% power and detect an ES *d* of 1.34 for an unpaired *t*-test. However, although the intended power was 80%, the overestimated ES (i.e., ES *d* = 1.34) yielded a true power of 19% (R code available at https://osf.io/y3482/). Thus, a researcher, who conducts a pre-study power calculation based on

the likely overestimated ES from an original small-sample study, may end up designing an experiment which has less power than intended, and to compound the issue the use of smaller sample sizes for a given power would ultimately yield overestimated ES. This situation does not only occur when conducting pre-study power calculations based on ES from previous studies with underpowered designs, but also when the ES of interest is derived from a pilot study (i.e., follow-up bias; see [70]). Consequently, researchers should take care when choosing the ES for a pre-study power calculation. As it is practically impossible to know the true effect size (and if it was known, there would be no need to collect additional data), researchers need to decide upon the expected effect size, for example based on the ES estimated from a meta-analysis, or based on the ES estimated from a previous study. However, in this case, researchers should use adjusting methods that account for the overestimation of ES due to small sample sizes and publication bias when conducting a pre-study power calculation [66,71]. A better approach is therefore to perform a power analysis based on a smallest effect size of interest [72].

Lastly, underpowered designs also increase the inaccuracy of parameter estimation [61,62; see **Fig. 8]**. This is because the width of CI around the parameter estimate depends on the standard deviation and the number of observations. Thus, larger sample sizes produce smaller standard deviations. The larger the CI around a parameter estimate, the less certain one can be that the estimate approximates the corresponding true population parameter [73]. As we can observe in **Fig. 8**, the width of a CI decreases as the sample size increases (which also increases the statistical power). ESs and CIs obtained with larger samples are more accurate than those obtained with smaller ones [73]. Similarly, it has been reported that out of a sample of 290 between-subject ES *d* estimates from 5 psychology journals, 83% of the ES sampled had CI widths that were larger than reported ES estimate and 26% were twice as large as the reported ES estimate [74]. As a consequence of the small samples sizes reported in sports sciences [34,37], it might be hypothesized that CI width might be larger than in other research areas with larger sample sizes such as psychology, further compounding potential issues with the precision of our observations.

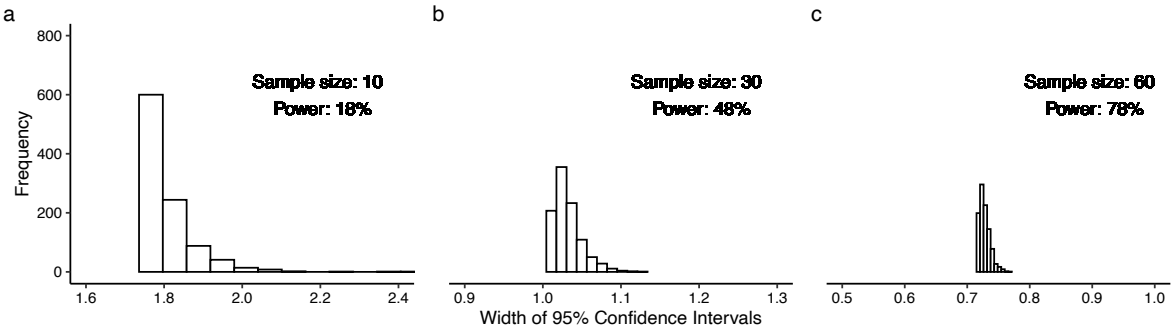

Fig. 8 Sample size affects the estimation of the CIs. Using the same data simulated as in **Fig. 2**, 1000 95% CI were computed. The histograms show the distribution of these 95% CI ranges for the same three different sample sizes. As sample size increases, both the range and the scatter of the CI decreases, reflecting increased power and greater precision from larger sample sizes.

2.2.3 Use of pre-study power calculations in sports science

Despite the core importance of power in NHST, the use of pre-study power calculations is still scarce in sports science [34]. In 2000, it was reported that of 40 articles published in the *Journal of Science and Medicine in Sport*, no study included a pre-study power calculation [61]. More recently, Abt et al. [34] reported that only 10% of

articles (12 out of 120) published in the *Journal of Sports Sciences* included such practice. Although this reflects an increased use of power analysis, it is clearly not a standard practice in our field. This is in marked contrast with the recent findings from Collins and Watt [75], who observed that 71% (152 out of 214) of psychologists self-reported to have used power analysis for sample size planning. There might be several reasons as to why pre-study power calculations are not standard practice in our field [73,75–77]. Firstly, researchers do not sufficiently understand this statistical concept and its importance in NHST [75]. This is reasonable to assume as all studies (12 out of 12) from Abt et al. [34] that included pre-study power calculations failed to disclose full information on the statistical test to be conducted to detect the chosen ES and 4 failed to include convincing rationale for why the given ES was chosen. It has been argued that if researchers do not have sufficient understanding of power, they cannot be expected to successfully calculate and accurately report power analysis [75]. Secondly, researchers may rely on intuition, rules of thumb or prior practices also known as heuristics to determine study sample sizes [76,77]. For instance, of 187 psychology researchers, 45 (23%) mentioned some rule of thumb (e.g., 20 subjects per condition) and 41 (21%) based their sample sizes on the common practice in their field of research [77]. These practices might be a major concern especially in scientific disciplines using small sample sizes and investigating small and medium ES because this combination would produce studies with underpowered designs as previously discussed. Thirdly, an common practice among researchers to determine the number of participants is optional stopping [13,14]. This practice involves stopping collecting data earlier than planned because a significant effect was found (**Fig.9**). This can occur in situations, for example, where a researcher who has already collected 30 observations per condition, and then tests for significance every 5 or 10 per condition observations [14]. However, such practice is considered a form of QRP because it leads to overestimated ES and increased type I error rates [14]. Instead, sample size planning should be based on a goal of achieving adequate power or accurate parameter estimates [67,73,76]. Therefore, given the scarce use of sample size planning based on power calculations and its lack of accurate reporting [34,75], it might be suggested that researchers in our field have a poor understanding of power and the consequences of low power designs on statistical results [18,19]. Furthermore, the scarcity of pre-study power calculations also suggests that sports science researchers may rely on either heuristics or optional stopping for sample size planning.

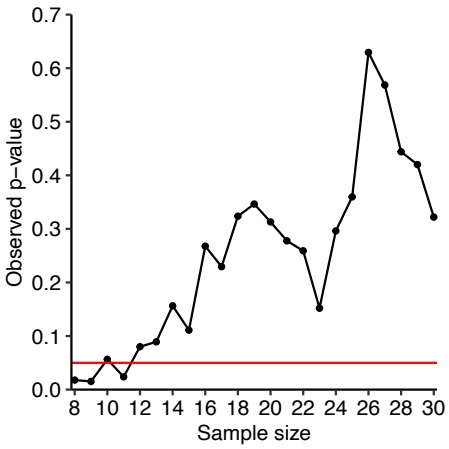

Fig. 9 Illustrative simulation of *p*-values obtained by a researcher who continuously adds a participant to each of two sample groups and conducts an unpaired *t*-test after each addition. The horizontal red line denotes statistical significance at $p < 0.05$. Note that sample size refers to number of participants per each of the two groups.

2.3. Availability of research data

2.3.1. Data sharing practices

Another methodological issue is our field is lack of data sharing. Empirical data shows that, in general, sports scientists are reluctant to engage in data sharing practices [35]. Indeed, Borg et al. [35] reported that only 13 of 299 articles published in 2019 in quartile one sports science journals shared data. Yet, this is not surprising given that only 5 of 286 articles stated that data was available upon request. The lack of data sharing practices might be problematic for several reasons. Firstly, it has been reported that about 50% of published studies in psychology contain at least one inconsistent *p*-value and about 13% contain a grossly inconsistent *p*-value [78,79]. Secondly, the willingness to share research data has been related to the strength of the statistical significance and a higher prevalence of reporting errors of statistical results [80]. Interestingly, *p*-values in the interval between .03 and .05 (which are less likely to occur when there is a true effect to be found), were more common in papers which did not share data (16.7%) than in papers which did (9.1%). Thirdly, integrity surveys among researchers have revealed that the prevalence of QRPs were in the range of 33-51% [81,82]. More serious forms of misconduct including fabrication and falsification of data or results have been reported to range between ~ 2 and 4% [81,82]. In light of these findings, there is a clear need to adopt data sharing practices that allow the research community to reproduce and replicate the published results.

2.3.2. Reporting practices

The *p*-value of a significance test is the main statistic used for rejecting H_0. However, researchers' poor understanding of the NHST often leads to the misconception that significance means a large effect whilst no significance means small effect or no effect [83–85]. In studies with underpowered designs, non-significant results are hardly indicative of the absence of an effect, and with large sample sizes, ES can be statistically significant but practically irrelevant [72]. It has been therefore recommended to combine the *p*-value along with ES estimates and their CIs [83,86]. An ES provides quantitative information about the magnitude of the relationship or effect studied and its CI indicate the uncertainty of that measure by presenting the range within which the true ES is likely to lie [68]. Furthermore, ES and their CIs allow findings from several studies to be combined in the form of meta-analysis to obtain more accurate ES estimates [68,87]. Despite this, the reporting of ES and CI is usually omitted in sports science [36,61]. For instance, Speed & Andersen [61] reported that only 14% (4 out of 29) of articles published in the *Journal of Science and Medicine in Sport* reported ES. Similarly, a more recent study observed that only 39% of studies published in the *Journal of Applied Biomechanics* in 2014 reported ES [36]. These findings suggest an overreliance on *p*-values to interpret the statistical results despite the consequences of small sample sizes on the reliability of statistical results [18,19].

Besides the quantitative information, reporting ES and their CI or at least including sufficient information to calculate them also contributes to improving the replicability of findings. For instance, a researcher attempting to replicate an original study with a higher-power design will need the original ES to estimate the sample size of the replication study. Similarly, a researcher might opt for a more conservative approach which is to use the lower CI bond of the original ES. Alternatively, researchers may use the accuracy in parameter estimation method, which also requires CIs, to identify the minimum sample size that would ensure a precise estimate of the population parameter [67]. Therefore, the omission of reporting ES and CI along with the lack of making data publicly

available may hinder any attempt of replication since other researchers might not be able to conduct a pre-study power calculation based on the original ES or CI.

However, reporting only ES and their CIs, and full information about the pre-study power calculations might not be enough. With the aim of facilitating cumulative scientific knowledge through meta-analysis [87] and the use of other statistical methods such as z-curve or power calculations [24,44,66], it has been suggested that besides sample size per condition, means, standard deviations (SDs) and exact *p*-values, studies should also disclose *F*-ratio or *t*-statistics, the type of design and the correlations between dependent observations for within-subjects designs [87], but it appears that this is rarely achieved. The compounding issues of poor reporting practices are easy to demonstrate with two examples; consider a within-subject design (i.e., pre vs. post) in which a study reports means and SDs but not the within-subject ES. Thus, a researcher attempting to conduct a meta-analysis, and assuming the study meets the inclusion criteria, should use the Hedges' g_{av} ES (ES g_{av}) from such study [88]. However, this researcher may well not be able to calculate ES g_{av} (see supplementary file in [77]) because the correlation between observations is never reported. Alternatively, as long as means, SDs, number of observations, t-statistic and exact *p*-value are reported, the researcher could use the user-friendly web application *within* [89] to estimate the correlation parameter, and then calculate ES g_{av}. However, again t-statistics and exact *p*-values are usually not reported. For instance, of 174 studies assessed for eligibility for our z-curve analysis, 45 (26%) were excluded because *p*-values were reported relatively (e.g., $p < 0.05$) and did not include sufficient information to be calculated and 2 (1.2%) because *p*-values were not reported at all. Finally, the researcher may opt to ask the study authors for the correlation, the t-statistic or the raw data so that researchers can calculate it themselves. Yet, given the reluctance of sports science researchers for sharing data [35], one possible outcome is that the researcher will not be able to get hold of this. Hence, the researcher may have to discard the study due to poor reporting practices and lack of data sharing. Second, a researcher attempting to conduct a pre-study power calculation using G*Power for a within-subject ANOVA will need the correlations between observations [90]. However, again this correlation is seldom reported. Taken together, these two hypothetical situations reflect some of the barriers that researchers have to overcome when attempting to conduct a meta-analysis or a pre-study power calculation.

Furthermore, the reporting of exact *p*-values and ES not only inform about the statistical significance, direction and magnitude of an effect, but also can be used to answer meta-scientific questions (e.g., how replicable is a particular set of findings?) by performing a z-curve/p-curve analysis, a meta-analysis or a meta-meta-analysis. Addressing meta-scientific questions may require the analysis of large datasets (see [10,47,91,92] for examples). This can be facilitated by the use of software to scan, select and analyze large sets of published data, where results should be machine-readable. The ultimate goal is to enhance the ability of computers to automatically find and use the data, in addition to supporting its reuse by researchers (i.e., FAIR principles; see [93]). This can be facilitated by the adoption of common reporting practices such as the reporting standards recommended by the American Psychological Association (APA). Following APA standards, statistic test results should be reported in the following order: the *F*-ratio or *t*-statistic and degrees of freedom (in parentheses) followed by the *p*-value (e.g., $F(1,35) = 5.45$, $p = 0.001$ or $t(85) = 2.86$, $p = 0.025$). However, this is not a common standard reporting practice in sports science. Thus, adopting common reporting practices, such as APA's reporting recommendation, would facilitate machine readability and data usability enabling the analysis of large sets of data containing *p*-values, ES

or CI. The reporting of statistical parameters is key to replicate original studies, assess the replication success and conduct additional statistical tests. However, the heterogeneity of our reporting practices in sports sciences make a full evaluation of replicability in our field problematic, to say the least.

2.4. Future recommendations for sports science: adoption of Open Science practices

As a consequence of above practices [8,18,19,34,73] and their effect on replicability rates reported by replication projects [25,26,26,28,29], Open Science practices are slowly being adopted within the research ecosystem. Open Science practices refer to a set of behaviors that enables research to be reproduced and replicated, with the aim of improving the reliability of scientific findings [73,94]. These practices may be especially important in research fields that reward publication of positive findings from studies with low power designs and exploiting, either intentionally or not, researchers' degrees of freedom [14,16,95]. We herein suggest a series of Open Science practices that could be adopted by researchers and journals to improve the replicability in our field [73,96,97].

One practice is preregistration which was conceived to mitigate QRPs by preventing HARKing and by reducing the risk of *p*-hacking via restricted flexibility in study design and data analysis [94,97]. In preregistered studies, authors register the protocol of their hypothesis, methods and analysis plan before data collection. Consequently, preregistered studies have been observed to produce smaller ES than non-preregistered studies due to the likely absence of publication bias and QRPs [98]. However, preregistration alone may still not be enough to prevent publication bias [99,100]. Alternatively, Registered Reports are considered a more effective format against publication bias [6,94,101,102]. For instance, Scheel et al. [6] found that 96% of non-registered studies reported statistically significant findings in comparison to 44% of Registered Reports. In a Registered Report, one submits a detailed plan of the research questions, hypotheses, methodology, and analysis to a scientific journal for review prior to collecting data. Once a Registered Report is accepted, the journal agrees to publish the study if the quality control criteria are met, regardless of the results. However, to date, only five sports science journals offer the Registered Report format, namely, *Journal of Experimental Physiology, Human Movement Science*, *Science and Medicine in Football* [103], *Psychology of Sport and Exercise*, *Reports in Sport and Exercise* and *Journal of Sports Sciences* [104]. Another practice that should be increasingly adopted is the use and report of pre-study power calculations for sample size planning to assure that studies are conducted with adequate power given the ES of interest [73,76]. In addition, low availability of research data reinforces the importance of sharing data including raw data, materials and code in public repositories, and improving transparency and quality of reporting practices [73,94]. Sharing research data alongside a manuscript increase the transparency of the research process because it allows both reviewers and readers to verify the results and therefore increase the reliability of the presented findings. Finally, researchers should conduct replications where needed and feasible [25–27,29,105–107]. Replication provides diagnostic evidence about a finding, and allows for exploring the boundaries of studied effects, and ultimately, the progression of science by confronting the existing understanding with new evidence [30,60,108,109].

3. Limitations of study selection protocol

Our investigation has a few limitations that should be addressed herein. Firstly, our selection is a convenience sample of original studies published in only one sports science journal. Thereby, our findings are far from a complete picture of the field of sports sciences. Furthermore, the small sample of studies included (N = 119) increased the uncertainty around the parameter estimates [110], and therefore this should be investigated in a larger study as opposed to the example case study presented here. Secondly, the protocol followed to select *p*-values for z-curve required us to make multiple subjective decisions because selected studies usually: a) tested vague and multiple hypotheses, b) measured dependent variables that were often operationalized using additional constructs of the same measure and c) used dependent variables that were measured in several alternative ways (see [16] for researchers' degrees of freedom). Third, although a secondary author undertook some random verification of the data selected, only the primary author extracted and coded data. This and the fact that data extraction was often difficult due to the researchers' degrees of freedom might have been a source of bias.

4. Conclusions

Based on previous findings in other research areas [8,19,25,26,40] and similarities to our own discipline [4,31,34,35], several methodological issues such as a high proportion of significant findings, studies with underpowered designs, and inaccurate reporting practices cast serious doubts about the replicability of sports science findings [4,31,34,35]. Firstly, there seems to be an excess of significant findings given the observed presence of publication bias and underpowered designs in the set of studies examined. Secondly, the small sample sizes reported in several biomechanics and sports science journals may also be a cause of concern, especially in studies using between-subject designs, for several reasons [8,18]. Small samples are likely to yield underpowered designs which are known to increase the proportion of false positives and false negatives, produce overestimated ES estimates and increase inaccuracy of parameter estimation (i.e., wide CI). Thirdly, there is clear evidence that most studies do not report enough statistical data, such as ES, CI, *F*-ratios, *t*-statistics and degrees of freedom, which directly impact the ability to evaluate methodological quality effectively. Altogether, the presence of these methodological issues suggests that there is clear room for improving our research standards and highlights the importance of increasingly adopting Open Science practices in sports science research.

Statements and Declarations

Conflicts of Interest

Cristian Mesquida, Jennifer Murphy, Daniël Lakens and Joe Warne declare that they have no conflict of interest.

Funding

Cristian Mesquida is funded by Technological University Dublin. Jennifer Murphy is a recipient of the Irish Research Council's Government of Ireland Postgraduate Scholarship Programme (project ID GOIPG/2020/1155).

Availability of data and material

Raw data, statistical analyses and R code (including source code for simulations and figures) used for this review are available on the Open Science Framework repository, https://osf.io/y3482/.

Author contributions

CM and JW conceived this review. CM conducted the literature review, screened the studies to be included in z-curve, and extracted and analyzed data. DL made substantial contributions to the analysis and interpretation of z-curve data. CM drafted the first version of the manuscript. JM, DL and JW edited and critically revised the manuscript for important intellectual content. All authors read and approved the final version of the manuscript.

5. References

1. Ioannidis JPA. What Have We (Not) Learnt from Millions of Scientific Papers with P Values? Am Stat. 2019; 73(1):20-25. https://doi.org/10.1080/00031305.2018.1447512

2. Fanelli D. "Positive" Results Increase Down the Hierarchy of the Sciences. PLoS One. 2010; 5(4): e10068. https://doi.org/10.1371/journal.pone.0010068

3. Sterling TD, Rosenbaum WL, Weinkam JJ. Publication Decisions Revisited: The Effect of the Outcome of Statistical Tests on the Decision to Publish and Vice Versa. Am Stat. 1995; 49(1):108-112. https://doi.org/10.2307/2684823

4. Büttner F, Toomey E, McClean S, Roe M, Delahunt E. Are questionable research practices facilitating new discoveries in sport and exercise medicine? The proportion of supported hypotheses is implausibly high. Br J Sports Med. 2020; 54(22):1365-1371.doi:10.1136/bjsports-2019-101863

5. Twomey R, Yingling V, Warne J, Schneider C, McCrum C, Atkins W, et al. The Nature of Our Literature: A Registered Report on the Positive Result Rate and Reporting Practices in Kinesiology. Commun Kinesiol. 2021; 1(3). https://doi.org/10.51224/cik.v1i3.43

6. Scheel AM, Schijen MRMJ, Lakens D. An Excess of Positive Results: Comparing the Standard Psychology Literature With Registered Reports. Adv Methods Pract Psychol Sci. 2021; 4(2):1-12. https://doi.org/10.1177/25152459211007467

7. Cohen J. The statistical power of abnormal-social psychological research: a review. J Abnorm Soc Psychol. 1962;65:145-53. https://doi.org/10.1037/h0045186

8. Fraley RC, Vazire S. The N-Pact Factor: Evaluating the Quality of Empirical Journals with Respect to Sample Size and Statistical Power. PLoS One. 2014; 9(10): e109019. https://doi.org/10.1371/journal.pone.0109019

9. Bakker M, van Dijk A, Wicherts JM. The Rules of the Game Called Psychological Science. Perspect Psychol Sci. 2012; 7(6):543-54. https://doi.org/10.1177/1745691612459060

10. Stanley TD, Carter EC, Doucouliagos H. What meta-analyses reveal about the replicability of psychological research. Psychol Bull. 2018; 144(12):1325-1346. https://doi.org/10.1037/bul0000169

11. Mahoney MJ. Publication prejudices: An experimental study of confirmatory bias in the peer review system. Cogn Ther Res. 1977; 1:161–175. https://doi.org/10.1007/BF01173636

12. Rosenthal R. The file drawer problem and tolerance for null results. Psychol Bull. 1979; 86(3), 638–641. https://doi.org/10.1037/0033-2909.86.3.638

13. John LK, Loewenstein G, Prelec D. Measuring the Prevalence of Questionable Research Practices With Incentives for Truth Telling. Psychol Sci. 2012; 23(5):524-532. https://doi.org/10.1177/0956797611430953

14. 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):1359-1366. https://doi.org/10.1177/0956797611417632

15. Kerr NL. HARKing: Hypothesizing After the Results are Known. Personal Soc Psychol Rev. 1998; 2(3):196-217. https://doi.org/10.1207/s15327957pspr0203_4

16. Wicherts JM, Veldkamp CLS, Augusteijn HEM, Bakker M, van Aert RCM, van Assen MALM. Degrees of Freedom in Planning, Running, Analyzing, and Reporting Psychological Studies: A Checklist to Avoid p-Hacking. Front Psychol. 2016; 7:1832. https://doi:10.3389/fpsyg.2016.01832

17. Gelman A, Loken E. The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time. 2013. Available from: http://stat.columbia.edu/~gelman/research/unpublished/forking.pdf

18. Button KS, Ioannidis JPA, Mokrysz C, Nosek BA, Flint J, Robinson ESJ, et al. Power failure: why small sample size undermines the reliability of neuroscience. Nat Rev Neurosci. 2013; 14(5):365-76. https://doi:10.1038/nrn3475

19. Maxwell SE. The Persistence of Underpowered Studies in Psychological Research: Causes, Consequences, and Remedies. Psychol Methods. 2004; 9(2):147-63. https://doi.org/10.1037/1082-989X.9.2.147

20. Bishop DV. The psychology of experimental psychologists: Overcoming cognitive constraints to improve research: The 47th Sir Frederic Bartlett Lecture. Q J Exp Psychol. 2020; 73(1):1-19. https://doi.org/10.1177/1747021819886519

21. Bird A. Understanding the Replication Crisis as a Base Rate Fallacy. Br J Philos Sci. 2020; 27(4). https://doi.org/10.1093/bjps/axy051

22. Oberauer K, Lewandowsky S. Addressing the theory crisis in psychology. Psychon Bull Rev. 2019; 26(5):1596-1618. https://doi.org/10.3758/s13423-019-01645-2

23. Cohen J. Statistical Power Analysis. Curr Dir Psychol Sci. 1992; 1(3):98-101. https://doi:10.1111/1467-8721.ep10768783

24. Bartoš F, Schimmack U. Z-curve. 2.0: Estimating replication rates and discovery rates. 2020. https://doi://10.31234/osf.io/urgtn

25. Collaboration OS. Estimating the reproducibility of psychological science. Science. 2015; 349(6251):aac4716. https://doi:10.1126/science.aac4716

26. Camerer CF, Dreber A, Holzmeister F, Ho T-H, Huber J, Johannesson M, et al. Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. Nat Hum Behav. 2018; 2(9):637-644. https://doi.org/10.1038/s41562-018-0399-z

27. Camerer CF, Dreber A, Forsell E, Ho T-H, Huber J, Johannesson M, et al. Evaluating replicability of laboratory experiments in economics. Science. 2016; 351(6280):1433-6. https://doi:10.1126/science.aaf0918

28. Klein RA, Ratliff KA, Vianello M, Adams RB, Bahník Š, Bernstein MJ, et al. Investigating Variation in Replicability. Soc Psychol. 2014; 45(3):142-152. htpps://doi.org/10.1027/1864-9335/a000178

29. Errington TM, Mathur M, Soderberg CK, Denis A, Perfito N, Iorns E, et al. Investigating the replicability of preclinical cancer biology. Pasqualini R, Franco E, editors. eLife. 2021; 10:e71601. https://doi:10.7554/eLife.71601

30. Klein RA, Vianello M, Hasselman F, Adams BG, Adams RB, Alper S, et al. Many Labs 2: Investigating Variation in Replicability Across Samples and Settings. Adv Methods Pract Psychol Sci. 2018; 1(4):443-490. https://doi:10.1177/2515245918810225

31. Halperin I, Vigotsky AD, Foster C, Pyne DB. Strengthening the Practice of Exercise and Sport-Science Research. Int J Sports Physiol Perform. 2018; 13(2):127-134. htpps://doi.org/10.1123/ijspp.2017-0322

32. Errington TM, Denis A, Perfito N, Iorns E, Nosek BA. Reproducibility in cancer biology: The challenges of replication. eLife. 2021; 10:e67995. https://doi.org/10.7554/eLife.23693

33. Del Vecchio A, Casolo A, Negro F, Scorcelletti M, Bazzucchi I, Enoka R, et al. The increase in muscle force after 4 weeks of strength training is mediated by adaptations in motor unit recruitment and rate coding. J Physiol. 2019; 597(7):1873-1887. https://doi:10.1113/JP277250

34. Abt G, Boreham C, Davison G, Jackson R, Nevill A, Wallace E, et al. Power, precision, and sample size estimation in sport and exercise science research. J Sports Sci. 2020; 38(17):1933-1935. https://doi:10.1080/02640414.2020.1776002

35. Borg DN, Bon J, Sainani KL, Baguley BJ, Tierney N, Drovandi C. Sharing Data and Code: A Comment on the Call for the Adoption of More Transparent Research Practices in Sport and Exercise Science. SportRxiv; 2020. https://doi:10.31236/osf.io/ftdgj

36. Vagenas G, Palaiothodorou D, Knudson D. Thirty-year Trends of Study Design and Statistics in Applied Sports and Exercise Biomechanics Research. Int J Exerc Sci. 2018;11: 239–259.

37. Knudson DV. Authorship and Sampling Practice in Selected Biomechanics and Sports Science Journals. Percept Mot Skills. 2011; 112(3):838-44. https://doi:10.2466/17.PMS.112.3.838-844

38. Bartoš F, Schimmack U. Z-Curve 2.0: Estimating Replication Rates and Discovery Rates. PsyArXiv. 2020. https://doi:10.31234/osf.io/urgtn

39. Brunner J, Schimmack U. Estimating Population Mean Power Under Conditions of Heterogeneity and Selection for Significance. Meta-Psychol. 2020. https://doi:10.15626/MP.2018.874

40. Ioannidis JPA. Why Most Published Research Findings Are False. PLoS Med. 2005; 2(8): e124. https://doi:10.1371/journal.pmed.0020124

41. Lakens D. Professors Are Not Elderly: Evaluating the Evidential Value of Two Social Priming Effects through P-Curve Analyses. PsyArXiv; 2017. https://doi:10.31234/osf.io/3m5y9

42. Lakens D. What p-hacking really looks like: A comment on Masicampo and LaLande (2012). Q J Exp Psychol. 2015; 68(4):829-32. https://doi:10.1080/17470218.2014.982664

43. Simmons JP, Simonsohn U. Power Posing: P-Curving the Evidence. Psychol Sci. 2017; 28(5):687-693. https://doi:10.1177/0956797616658563

44. Simonsohn U, Nelson LD, Simmons JP. P-curve: A key to the file-drawer. J Exp Psychol Gen. 2014; 143(2):534-47. https://doi:10.1037/a0033242

45. Hung HMJ, O'Neill RT, Bauer P, Kohne K. The Behavior of the P-Value When the Alternative Hypothesis is True. Biometrics. 1997; 53(1):11-22. https://doi:10.2307/2533093

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:10.1111/j.1745-6924.2008.00079.x

47. Hartgerink CHJ, van Aert RCM, Nuijten MB, Wicherts JM, van Assen MALM. Distributions of p-values smaller than .05 in psychology: what is going on? PeerJ. 2016; 4:e1935. https://doi:10.7717/peerj.1935

48. Francis G. Publication bias and the failure of replication in experimental psychology. Psychon Bull Rev. 2012; 19(6):975-91. https://doi:10.3758/s13423-012-0322-y

49. Franco A, Malhotra N, Simonovits G. Publication bias in the social sciences: Unlocking the file drawer. Science. 2014; 345(6203):1502-5. https://doi:10.1126/science.1255484

50. Bartoš F. zcurve. 2021. Available from: https://github.com/FBartos/zcurve/blob/a908ec8086fa440ad7dfaf6cec09a8b0343e52d8

51. Maxwell SE, Delaney HD, Kelley K. Designing experiments and analyzing data: A model comparison perspective. 3rd ed. Routledge. 2017.

52. Murphy J, Mesquida C, Caldwell AR, Earp BD, Warne J. Selection Protocol for Replication in Sports and Exercise Science. OSF Preprints; 2021. https://doi.org/10.31219/osf.io/v3wz4

53. Schimmack U, Brunner J. Z-Curve: A Method for the Estimating Replicability Based on Test Statistics in Original Studies. 2017. http://www.utstat.utoronto.ca/~brunner/zcurve2016/HowReplicable.pdf

54. Carter EC, McCullough ME. Publication bias and the limited strength model of self-control: has the evidence for ego depletion been overestimated? Front Psychol. 2014; 5:823. https://doi:10.3389/fpsyg.2014.00823

55. Friese M, Frankenbach J. p-Hacking and publication bias interact to distort meta-analytic effect size estimates. Psychol Methods. 2020; 25(4):456-471. https//:doi:10.1037/met0000246

56. Lakens D. The Practical Alternative to the p Value Is the Correctly Used p Value. Perspect Psychol Sci J Assoc Psychol Sci. 2021; 16(3):639-648. https://doi:10.1177/1745691620958012

57. Miller J. What is the probability of replicating a statistically significant effect? Psychon Bull Rev. 2009; 16:pages 617–640. https://doi:10.3758/PBR.16.4.617

58. Cohen J. Statistical power analysis for the behavioral sciences. 2nd ed. L. Erlbaum Associates; 1988.

59. Fisher RA. The arrangement of field experiments. J Minist Agric. 1926; 33:503-515. https://doi:10.23637/rothamsted.8v61q

60. Nosek BA, Errington TM. Making sense of replications. eLife. 2017; 6:e23383. https://doi:10.7554/eLife.23383

61. Speed HD, Andersen MB. What exercise and sport scientists don't understand. J Sci Med Sport. 2000; 3(1):84-92. https://doi:10.1016/S1440-2440(00)80051-1

62. Atkinson G, Nevill AM. Selected issues in the design and analysis of sport performance research. J Sports Sci. 2001; 19(10):811-27. https://doi:10.1080/026404101317015447

63. Rhea MR. Determining the magnitude of treatment effects in strength training research through the use of the effect size. J Strength Cond Res. 2004; 18(4):918-20. https://doi:10.1519/14403.1

64. Swinton P, Burgess K, Hall A, Greig L, Psyllas J, Aspe R, et al. A Bayesian approach to interpret intervention effectiveness in strength and conditioning Part 1: A meta-analysis to derive context-specific thresholds. SportRxiv; 2021. https://doi:10.51224/SRXIV.9

65. Knudson D. Confidence crisis of results in biomechanics research. Sports Biomech. 2017; 16(4):425-433. https://doi:10.1080/14763141.2016.1246603

66. Anderson SF, Kelley K, Maxwell SE. Sample-Size Planning for More Accurate Statistical Power: A Method Adjusting Sample Effect Sizes for Publication Bias and Uncertainty. Psychol Sci. 2017; 28(11):1547-1562. https://doi:10.1177/0956797617723724

67. Maxwell SE, Kelley K, Rausch JR. Sample size planning for statistical power and accuracy in parameter estimation. Annu Rev Psychol. 2008; 59:537-63. https://doi:10.1146/annurev.psych.59.103006.093735

68. 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:10.1038/nmeth.3288

69. Kvarven A, Strømland E, Johannesson M. Comparing meta-analyses and preregistered multiple-laboratory replication projects. Nat Hum Behav. 2020;4: 423–434. https://doi:10.1038/s41562-019-0787-z

70. 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, 187–195. https://doi:10.1016/j.jesp.2017.09.004

71. Simonsohn U, Nelson LD, Simmons JP. p-Curve and Effect Size: Correcting for Publication Bias Using Only Significant Results. Perspect Psychol Sci J Assoc Psychol Sci. 2014; 9(6):666–681. https://doi:10.1177/1745691614553988

72. Anvari F, Lakens D. Using anchor-based methods to determine the smallest effect size of interest. J Exp Soc Psychol. 2021; 96. https://doi:10.1016/j.jesp.2021.104159

73. Asendorpf JB, Conner M, Fruyt FD, Houwer JD, Denissen JJA, Fiedler K, et al. Recommendations for Increasing Replicability in Psychology. Eur J Personal. 2013; 27(2), 108–119. https://doi:https://doi.org/10.1002/per.1919

74. Brand A, Bradley MT. The Precision of Effect Size Estimation From Published Psychological Research: Surveying Confidence Intervals. Psychol Rep. 2016; 118(1):154-170. https://doi:10.1177/0033294115625265

75. Collins E, Watt R. Using and Understanding Power in Psychological Research: A Survey Study. Collabra Psychol. 2021; 7 (1): 28250. https://doi:10.1525/collabra.28250

76. Lakens D. Sample size justification. PsyArXiv. 2021. https://doi:10.31234/osf.io/9d3yf

77. Bakker M, Hartgerink CHJ, Wicherts JM, van der Maas HLJ. Researchers' Intuitions About Power in Psychological Research. Psychol Sci. 2016; 27(8):1069-77. https://doi:10.1177/0956797616647519

78. Bakker M, Wicherts JM. The (mis)reporting of statistical results in psychology journals. Behav Res Methods. 2011; 43(3):666-78. https://doi:10.3758/s13428-011-0089-5

79. Nuijten MB, Hartgerink CHJ, van Assen MALM, Epskamp S, Wicherts JM. The prevalence of statistical reporting errors in psychology (1985–2013). Behav Res Methods. 2016; 48:1205-1225. https://doi:10.3758/s13428-015-0664-2

80. Wicherts JM, Bakker M, Molenaar D. Willingness to Share Research Data Is Related to the Strength of the Evidence and the Quality of Reporting of Statistical Results. PLoS One. 2011; 6(11): e26828. https://doi:10.1371/journal.pone.0026828

81. Gopalakrishna G, Riet G ter, Vink G, Stoop I, Wicherts JM, Bouter LM. Prevalence of questionable research practices, research misconduct and their potential explanatory factors: A survey among academic researchers in The Netherlands. PLoS One. 2022;17:e0263023. https:doi:10.1371/journal.pone.0263023

82. Fanelli D. How Many Scientists Fabricate and Falsify Research? A Systematic Review and Meta-Analysis of Survey Data. PLoS One. 2009; 4(5):e5738. https://doi:10.1371/journal.pone.0005738

83. Cumming G. Understanding The New Statistics: Effect Sizes, Confidence Intervals, and Meta-Analysis. Routledge; 2013.

84. Greenland S, Senn SJ, Rothman KJ, Carlin JB, Poole C, Goodman SN, et al. Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. Eur J Epidemiol. 2016; 31(4):337-50. https://doi:10.1007/s10654-016-0149-3

85. Motulsky HJ. Common misconceptions about data analysis and statistics. Pharmacol Res Perspect. 2014; 387(11):1017-23. https://doi:10.1002/prp2.93

86. Sullivan GM, Feinn R. Using Effect Size—or Why the P Value Is Not Enough. J Grad Med Educ. 2012; 4(3):279-82. https://doi:10.4300/JGME-D-12-00156.1

87. Lakens D, Hilgard J, Staaks J. On the reproducibility of meta-analyses: Six practical recommendations. BMC Psychol. 2016; 4(1):24. https://doi:10.1186/s40359-016-0126-3

88. Lakens D. Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. Front Psychol. 2013; 4:863. https://doi:10.3389/fpsyg.2013.00863

89. DeBruine L. Within-subject t-test forensics. 2021. Available from: https://github.com/debruine/within/

90. Faul F, Erdfelder E, Buchner A, Lang A-G. Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. Behav Res Methods. 2009; 41(4):1149-60. https://doi:10.3758/BRM.41.4.1149

91. Chavalarias D, Wallach JD, Li AHT, Ioannidis JPA. Evolution of Reporting P Values in the Biomedical Literature, 1990-2015. JAMA. 2016; 315(11):1141-8. https://doi:10.1001/jama.2016.1952

92. Szucs D, Ioannidis JPA. Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. PLoS Biol. 2017; 19(3): e3001151. https://doi:10.1371/journal.pbio.2000797

93. Wilkinson MD, Dumontier M, Aalbersberg IjJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2016; 3:160018. https://doi:10.1038/sdata.2016.18

94. Munafò MR, Nosek BA, Bishop DVM, Button KS, Chambers CD, Percie du Sert N, et al. A manifesto for reproducible science. Nat Hum Behav. 2017; 1:0021. https://doi:10.1038/s41562-016-0021

95. Smaldino PE, McElreath R. The natural selection of bad science. R Soc Open Sci. 2016; 3(9):160384. https://doi:10.1098/rsos.160384

96. Nosek BA, Alter G, Banks GC, Borsboom D, Bowman SD, Breckler SJ, et al. Promoting an open research culture: Author guidelines for journals could help to promote transparency, openness, and reproducibility. Science. 2015; 348(6242):1422-1425. https://doi:10.1126/science.aab2374

97. Caldwell AR, Vigotsky AD, Tenan MS, Radel R, Mellor DT, Kreutzer A, et al. Moving Sport and Exercise Science Forward: A Call for the Adoption of More Transparent Research Practices. Sports Med. 2020; 50(3):449-459. https://doi:10.1007/s40279-019-01227-1

98. 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:10.3389/fpsyg.2019.00813

99. Goldcare B, Drysdale H, Powell-Smith A, Dale A, Milosevic I, Slade E, et al. The COMPare Trials Project. 2016. http://www.COMPare-trials.org

100. Rasmussen N, Lee K, Bero L. Association of trial registration with the results and conclusions of published trials of new oncology drugs. Trials. 2009; 10:116. https:doi:10.1186/1745-6215-10-116

101. Nosek BA, Lakens D. Registered reports: A method to increase the credibility of published results. Soc Psychol. 2014; 45(3):137-141. https://doi:10.1027/1864-9335/a000192

102. Allen C, Mehler DMA. Open science challenges, benefits and tips in early career and beyond. PLoS Biol. 2019; 7(12):e3000587. https://doi:10.1371/journal.pbio.3000246

103. Impellizzeri FM, McCall A, Meyer T. Registered reports coming soon: our contribution to better science in football research. Sci Med Footb. 2019; 3(2):87-88. https://doi:10.1080/24733938.2019.1603659

104. 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):1789-1790. https://doi:10.1080/02640414.2021.1950974

105. Field SM, Hoekstra R, Bringmann L, van Ravenzwaaij D. When and Why to Replicate: As Easy as 1, 2, 3? Collabra Psychol. 2019; 5(1): 46. https://doi:10.1525/collabra.218

106. Isager PM, van Aert RCM, Bahník Š, Brandt MJ, DeSoto KA, Giner-Sorolla R, et al. Deciding what to replicate: A decision model for replication study selection under resource and knowledge constraints. Psychol Methods. 2021. https://doi:10.1037/met0000438

107. Coles NA, Tiokhin L, Scheel AM, Isager PM, Lakens D. The costs and benefits of replication studies. Behav Brain Sci. 2018; 41:e124. https://doi:10.1017/S0140525X18000596

108. Nosek BA, Hardwicke TE, Moshontz H, Allard A, Corker KS, Dreber A, et al. Replicability, Robustness, and Reproducibility in Psychological Science. Annu Rev Psychol. 2022; 73(1):719-748. https://doi:10.1146/annurev-psych-020821-114157

109. Nosek BA, Errington TM. What is replication? PLoS Biol. 2020; 18(3):e3000691. https://doi:10.1371/journal.pbio.3000691

110. Brunner J, Schimmack U. How replicable is psychology? A comparison of four methods of estimating replicability on the basis of test statistics in original studies. 2016. Available from: http://www.utstat.utoronto.ca/~brunner/zcurve2016/HowReplicable.pdf