Publication bias, statistical power and reporting practices in the Journal of Sports Sciences: Potential barriers to replicability

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

[1]Centre of Applied Science for Health, Technological University Dublin, Tallaght, Dublin, Ireland
[2]Human-Technology Interaction Group, Eindhoven University of Technology, Eindhoven, The Netherlands

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). Publication bias, statistical power and reporting practices in the Journal of Sports Sciences: Potential barriers to replicability. *SportRχiv*

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

Abstract

When designing studies researchers often assume that findings can be replicated, and are not false positive results. However, in literature that suffer from underpowered designs and publication bias, the replicability of findings can be hindered. A previous study by Abt et al., (2020) reported a median sample size of 19 and the scarce usage of pre-study power analyses in studies published in the *Journal of Sports Sciences*. We meta-analyzed 89 studies from the same journal to assess the presence and extent of publication bias, as well as the average statistical power, by conducting a z-curve analysis. In a larger sample of 179 studies, we also examined a) the usage, reporting practices, and reproducibility of pre-study power analyses; and b) the prevalence of reporting practices of *t*-statistic or *F*-ratio, degrees of freedom, exact *p*-values, effect sizes and confidence intervals. Our results indicate that there was some indication of publication bias and the average observed power was low (53% for significant and non-significant findings and 61% for only significant findings). Finally, the usage and reporting practices of pre-study power analyses as well as statistical results including test statistics, effect sizes and confidence intervals were suboptimal.

Keywords

replicability, publication bias, statistical power, reporting practices, reproducibility

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

Data availability statement All data sets generated and the R code used to analyse data are available on Open Science Framework at https://osf.io/e3rab/.

Funding details Cristian Mesquida is funded by Technological University Dublin (project ID PTUD2002). Jennifer Murphy is a recipient of the Irish Research Council's Government of Ireland Postgraduate Scholarship Programme (project ID GOIPG/2020/1155).

1. Introduction

Replicability refers to testing an effect observed in a prior finding using the same study design and data analysis but collecting new data (Nosek et al., 2022). When a study finding can be replicated, researchers can therefore be more confident the original finding is not a false negative. Replication projects across several scientific disciplines such as psychology (Open Science Collaboration, 2015), the social sciences (Camerer et al., 2018) and, more recently, cancer biology (Errington et al., 2021) have attempted to replicate original studies. A common outcome of these replication projects was that original effects were often difficult to replicate even when larger sample sizes are collected, and if detected, effect sizes were smaller than in the original report (i.e., overestimated effect sizes). These results have sparked renewed interest in research practices that hinder the replicability of prior findings (Button et al., 2013; Carter & McCullough, 2014; Errington et al., 2021; Francis, 2012; Simmons et al., 2011; Wicherts et al., 2016). Three issues that are known to lower the replicability of published findings are studies with underpowered designs, *p*-hacking, and a scientific literature that suffers from publication bias (Bakker et al., 2016; Button et al., 2013; Fraley & Vazire, 2014; Francis, 2012; Franco et al., 2014; Stefan & Schönbrodt, 2022).

Statistical power is the probability of rejecting the null hypothesis when it is false (i.e., the probability of finding a significant effect when there is one to be found) and depends on the effect size of interest, the sample size, the statistical test and the Type I error rate (Cohen, 1962; Maxwell et al., 2017). For example, studies investigating small and medium effects with small samples are likely to be underpowered, and therefore they have a higher probability of yielding a false negative result. Interestingly, Abt et al., (2020) reported that the *Journal of Sports Sciences* published studies with a median sample size of 19 participants. Depending on the design and the effect size, a study using a sample size of 19 participants may not have sufficient power, particularly when effects are relatively small and between participant designs are used (Maxwell et al., 2017). For example, a within-participant design with a sample size of 20 participants and where the effect of interest, d_z, is 0.5, would have 56% power for a two-sided test with an alpha of 5%. A between-participant design with a sample size of 10 in each condition and an effect of interest, d_s, of 0.5 would have a power of 19% for a two-sided test with an alpha of 5%. These two studies would require a total sample size of 44 and 172, respectively, to detect a Cohen's d_z of 0.5 with a statistical power of 90%. Consequently, it is important to examine the designs of the studies published in the *Journal of Sports Sciences* are sufficiently powered for effects of interest despite the small sample sizes previously reported (Abt et al., 2020).

Publication bias occurs when studies with statistically significant findings have a higher chance of being published than statistically non-significant findings. This phenomenon includes editors and reviewers selectively publishing studies with significant findings (i.e., review bias; Mahoney, 1977) and researchers deciding not to submit studies with non-significant results (i.e., the file-drawer problem; Rosenthal, 1979). This is especially problematic when studies have underpowered designs because such studies suffer from large sampling error which leads to substantial uncertainty about the true effect size (Cumming, 2013). Furthermore, when a study with a between-subject design investigates a true Cohen's d_s effect size = 0.5 and there are only 20 subjects per condition, it is not possible to get a $p < 0.05$ unless the true effect size is overestimated (Cumming, 2013), as the minimal detectable effect size with an alpha of 0.05 is $d_s = 0.64$ (Lakens, 2022). Publication bias increases the false positive report probability (Wacholder et al., 2004), or the probability that a published significant finding is actually a Type I error. Furthermore, publication bias based on statistical significance and in the presence of studies with small sample sizes leads to overestimated effect size estimates (Anderson et al., 2017; Bartoš & Schimmack, 2022). Despite the relevance of publication bias to the non-replication of studies and cumulative research (Carter & McCullough, 2014; Francis, 2012; Franco et al., 2014), it has been overlooked in the field of sports and exercise science. The presence of publication bias and studies with underpowered designs in a body of literature can be examined using a z-curve analysis (Bartoš & Schimmack, 2022; Brunner & Schimmack, 2020; see also Simonsohn et al., 2014a, 2014b for *p*-curve). The z-curve method converts significant and non-significant *p*-values reported in a literature into z-scores, and uses the distribution of z-scores to determine the presence of publication bias. It also estimates the average statistical power of the studies conducted and provides an estimate of their replicability.

To ensure studies are adequately powered to observe the effect size of interest in studies in which researchers aim to perform a hypothesis test, one should conduct a pre-study power analysis (Lakens, 2022). However, despite the importance of providing an adequate sample size justification, Abt et al., (2020) reported that only 10% of articles (12 out of 120) published in the *Journal of Sports Sciences* included a pre-study power analysis. The lack of pre-study power analysis may indicate that researchers rely on intuition, rules of thumb, or prior practices (a.k.a., heuristics) to determine study sample sizes, such as “20 subjects per condition” or otherwise simply using the same sample sizes typically reported in their field of research (Anderson et al., 2017; Bakker et al., 2016; Lakens, 2022). Alternatively, it may also indicate that some researchers determine the sample size based on the questionable research practices of optional stopping (see John et al., 2012 and Wicherts et al., 2016) which ultimately increase the chances of committing a Type I error (Simmons et al., 2011; Stefan & Schönbrodt, 2022). Furthermore, Abt et al., (2020) also reported that all studies (12 out of 12) that included a pre-study power analysis failed to disclose information on the statistical test to be conducted to detect the chosen effect size. Although this prevents other researchers from evaluating the adequacy of the power analysis, as well as making it impossible to assess the reproducibility of these pre-study power analysis, no study has examined the reporting practices including the magnitude of the effect size of interest, the statistical test and the intended power which are required to enable the reproducibility of pre-study power analyses at the very least.

Given that the presence of publication bias and studies with underpowered designs are a threat to the replicability of original findings, one response to the presence of these issues is the replication of original studies with well-powered designs (e.g., Open Science Collaboration, 2015). To facilitate the replicability of original studies, studies should provide a complete description of statistical results. Several current practices in terms of Null Hypothesis Significance Testing require the use of the original effect size for assessing the replicability of original studies (Camerer et al., 2018; Errington, Mathur, et al., 2021; Open Science Collaboration, 2015; Simonsohn, 2015). Furthermore, effect sizes from published studies can be used to conduct pre-study power analysis for sample size planning in follow-up studies and to draw meta-analytic conclusions by comparing effect sizes across studies (i.e., in a meta-analysis). Finally, the reporting of effect size estimates allows researchers to discuss the magnitude or practical significance of the studied effect (Kelley & Preacher, 2012; see also Götz et al., 2022 and Primbs et al., 2022). However, the reporting of only the effect size estimate might not be sufficient. The American Psychological

Association's (APA) recommendations for best reporting practices include the effect size, confidence intervals (CI), and exact *p*-value (see Appelbaum et al., 2018). Studies with underpowered designs increase the uncertainty around the effect size estimate which is reflected in the width of the CI for the effect size estimate (Asendorpf et al., 2013). However, to what extent these recommended best practices are implemented in sport science journals are unknown.

Our first aim in this study was to assess the presence of publication bias and studies with underpowered designs in a set of studies published in the *Journal of Sport Sciences*. The rationale of selecting the *Journal of Sports Sciences* was the use of small samples (n = 19) and the scarce use of pre-study power analysis in studies published in this journal (Abt et al., 2020). The second aim was to examine the usage, reporting practices and reproducibility of pre-study power analysis. Thirdly, we sought to investigate the prevalence of reporting practices of *t*-statistics or *F*-ratios, degrees of freedom, exact *p*-values, and effect sizes and their CI.

2. Methods

The materials including the study selection protocol, dataset generated, disclosure table and R code for the z-curve analysis are available at https://osf.io/e3rab/. This study was exploratory with an observational and retrospective design.

2.1. Selection protocol

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* (Murphy et al., 2022). Hence, only applied sports and exercise science studies in the subdisciplines of biomechanics, injury prevention, nutrition, physical activity, physiology, psychology and sports performance 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 tested a hypothesis and contained an inference test such as a *t*-test and *F*-test. Studies that test a hypothesis are especially sensitive to publication bias, compared to studies that only report descriptive statistics or effect size estimates, as both authors and scientific journals value significant results more than non-significant results (Greenwald, 1975). The 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. A total of 523 studies were screened of which 349 were excluded for not meeting the above criteria. 89 studies met the above criteria and were included in the z-curve analysis (**Figure 1**).

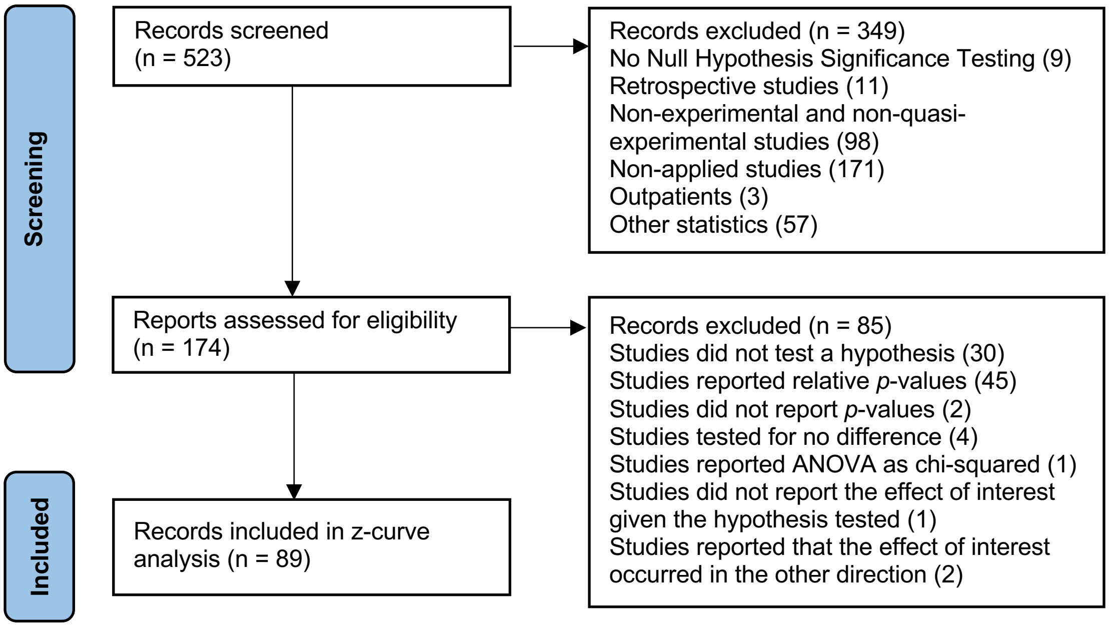
Screening
Included
Records screened
(n = 523)
Records excluded (n = 349)
No Null Hypothesis Significance Testing (9)
Retrospective studies (11)
Non-experimental and non-quasi-experimental studies (98)
Non-applied studies (171)
Outpatients (3)
Other statistics (57)
Reports assessed for eligibility
(n = 174)
Records excluded (n = 85)
Studies did not test a hypothesis (30)
Studies reported relative p-values (45)
Studies did not report p-values (2)
Studies tested for no difference (4)
Studies reported ANOVA as chi-squared (1)
Studies did not report the effect of interest given the hypothesis tested (1)
Studies reported that the effect of interest occurred in the other direction (2)
Records included in z-curve analysis (n = 89)

Figure 1. PRISMA flow diagram for inclusion of studies in z-curve analysis

2.2. Extracting *p*-values

After study selection, only one *p*-value per independent experiment was extracted in order to meet the independence criteria (Bartoš & Schimmack, 2022). The extracted *p*-value corresponded to the first or primary dependent variable stated in the hypothesis. In cases where there were multiple hypotheses, the first or primary hypothesis was considered. If the selected hypothesis 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) using the functions *T.DIST.2T* or *F.DIST.RT* for *t*-tests and *F*-tests in Microsoft Excel for Mac (Version 16.45). *P*-values were discarded under 5 circumstances; a) when the *p*-value was reported relatively (e.g., $p < 0.05$) and it could not be recomputed due to lack of sufficient information; b) when studies tested an hypothesis for non-significance; c) the described statistical test in the methods did not match the statistical test reported in the results section of the study; d) the study did not report the effect of interest given the hypothesis stated in the introduction; and e) the study expected to find a significant difference in one direction but observed an effect in the other direction; the inclusion of this category of significant *p*-values in z-curve would be problematic because it could create bias in favor of statistical significance. A disclosure table containing all extracted information for the z-curve analysis can be found at https://osf.io/e3rab/. A total of 174 studies were screened of which 85 did not meet the above criteria. Thus, 89 studies were included in the z-curve analysis. A secondary z-curve performed on 119 *p*-values obtained from studies that aimed to test a hypothesis (n = 89) and studies that were considered to be descriptive because no hypothesis was tested (n = 30) can be found in supplemental material at https://osf.io/e3rab/ .

2.2. Publication bias and statistical power

Z-curve is based on the idea that the average power of a set of studies can be derived from the distribution of z-scores (Bartoš & Schimmack, 2022; Brunner & Schimmack, 2020). 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 statistical power. 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. If there is no true effect, the Expected Replication Rate equals the Type 1 error rate and if there is a true effect, it equals the average power estimate. 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 examined by comparing the Observed Discovery Rare to the Expected Discovery Rate. If the point estimate of the Observed Discovery Rate lies within the 95% CI of the Expected Discovery Rate, there is no evidence of publication bias. The z-curve method also provides other estimates of publication bias such as the file-drawer ratio which is the ratio between the Expected Discovery Rate and the Observed Discovery Rate and is expressed as the number of unpublished studies that are predicted to exist for every published study. However, one should note the file-drawer ratio is simply a transformation of the Expected Discovery Rate.

2.3 Pre-study power analysis and their reporting practices

To investigate the frequency of usage of pre-study power analysis and their reporting practices, the sample of studies was expanded to include those studies that did not meet the criteria for the z-curve analysis (see **Figure 1**). Thus, a total sample of 174 studies was used for the second aim of this study. Two strategies were used to detect the use of pre-study power analyses. First, a visual inspection was performed. The author C.M. searched for any mention of a pre-study power analysis or implicit suggestions of power reported within the methods section (i.e., Participants and Statistical analysis) of an article. If the first strategy was unsuccessful, the article was then downloaded as a PDF and a search was conducted by using keywords "power", "sample", "size" and "participants" or "subjects". In case the study reported the use of a pre-study power analysis, the following information was retrieved when available: type of power analysis (i.e., pre-study or post-study), software, statistical test, variable of interest, magnitude of the effect size and its type (e.g., Cohen's *d*, Hedge's *g,* Cohen's *f)*, effect size justification (i.e., previous study, pilot study, Cohen's *d* benchmarks, smallest effect size of interest (SESOI) and meta-analysis), alpha level, intended power, and the sample size required to achieve the intended power. Once this information was retrieved, each category was scored dichotomously as either one or zero (1 = *present,* 0 = *not present*). The use of a post-study power analysis or implicit suggestions of its use were also coded, but no information regarding the reporting practices of such analysis was retrieved. This is because post-study power analyses are considered bad practice (Christogiannis et al., 2022; Yuan & Maxwell, 2005). Moreover, the author C.M. coded whether each one of the sampled 174 studies that tested a hypothesis included a pre-study power analysis because studies that have the goal to test a hypothesis (compared to studies that have a descriptive or estimation goal) should be designed to explicitly control the Type 2 error rate by collecting sufficient data (Lakens & Evers, 2014). We also attempted to reproduce the sample size obtained from pre-study power analyses that reported effect size magnitude and type, statistical test and intended statistical power using the original statistical software. For the studies that included this information, all studies used G*Power. We therefore attempted to reproduce the sample size calculations using G*Power (version 3.1.9.6).

2.4. Reporting practices of statistical results

To investigate the reporting practices of statistical results, the same sample of studies as described above was used (n = 174). To select the statistical result, the same procedures applied to extract the *p*-value for the z-curve analysis were followed. Thus, the statistical result selected was chosen in relation to the first or primary study hypothesis/aim as well as the first or primary dependent variable stated within. The following statistics were retrieved from results section of an article when available: mean ± standard deviation (SD) or mean ± standard error of mean (SEM), *t*- or *F*-statistic, degrees of freedom, *p*-value, standardized effect sizes (e.g., eta squared (η^2) and Cohen's *d* family) and its CI. For the purpose of this study, only standardized effect sizes were considered because such effect sizes allow researchers to conduct pre-study power analyses for follow-up studies. For studies in which the study hypothesis was linked to a factorial analysis, we only considered the effect size (e.g., partial eta squared (η_p^2), eta squared (η^2)) for the omnibus effect of interest (i.e., main or interaction effect). For instance, if a study using a one-way between-subject ANOVA with 4 levels only reported pairwise effect size but not the omnibus effect, the pairwise effect size was not considered. A pairwise effect size was only considered if the omnibus effect of interest was a main effect and with only two levels. This is because a main effect with only one degree of freedom would be equivalent to a statistical test of mean differences (e.g., one-sample and two-sample *t*-test), and therefore the correct effect size to report would be part of Cohen's *d* family. Once the above information was retrieved, each category was scored dichotomously as either one or zero (1 = *present,* 0 = *not present*).

2.5. Statistical analysis

The R package *zcurve 2.0* was used to conduct the z-curve analysis (Bartoš & Schimmack, 2022). Descriptive statistics in the form of count and frequency (%) were used to evaluate the prevalence of both pre-study and post-study power analyses, and reporting practices for both power analysis and statistical results. Two two-tailed Welch's *t*-tests were performed to determine whether a) studies that performed a pre-study power analysis had different sample sizes compared to studies without a pre-study power analysis, and b) amongst studies that tested a hypothesis, studies that performed a pre-study power analysis had different sample sizes compared to those that did not perform a pre-study power analysis. Hedges' g_s effect size and its 95% CI was calculated to present the magnitude of the difference using the R package *deffectsize* (Delacre et al., 2021; see also https://effectsize.shinyapps.io/deffsize/). Alpha level was set to $\alpha < 0.05$. Statistical tests were conducted using R (Version 4.1.2; R Core Team, 2021). To reproduce the pre-study power analyses reported in the set of studies, we used G*Power (Version 3.1.9.6).

3. Results

A total of 89 independent *p*-values (including 65 significant and 24 non-significant *p*-values) were converted into z-scores to fit the z-curve model. The Expected Discovery Rate was 0.53 [0.13; 0.71] indicating an average power of 53% for studies reporting both significant and non-significant results (see **Figure 2)**. The Expected Replication Rate was 0.61 95% CI [0.42; 0.75] 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 same statistical power, effect size and sample size) of the studies reporting significant findings, only 60% of these studies would yield another significant effect. Publication bias can be examined by comparing the Observed Discovery Rate (the percentage of significant results in the set of studies) to the expected discovery rate (the proportion of the area under the curve on the right side of the significance criterion). The point estimate of the Observed Discovery Rate (0.73) lies outside the 95% CI of the Expected Discovery Rate of 0.53 [0.13; 0.71] suggesting that we can statistically reject the null hypothesis that there is no publication bias. This conclusion is also supported by a visual inspection of the obtained results, which suggest there is a potential indication of publication bias (see **Figure 2**); there is a steep drop from the frequency of just statistically significant values (i.e., z > 1.96) relative to the frequency of 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% (i.e., z = 1.96) than below this threshold.

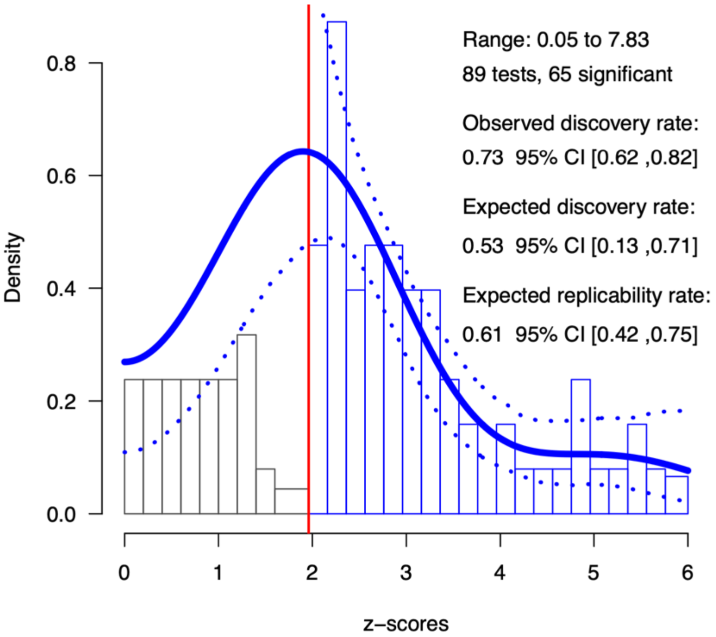

Figure 2. Distribution of z-scores over [0-6] interval. The vertical red line refers to a z-score of 1.96, the critical value for statistical significance when using a two-tailed alpha of 0.05. The dark blue line is the density distribution for the inputted *p*-values (represented in the histogram as 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.

Out of 174 sampled studies, only 46 (26%) included a pre-study power analysis and 10 studies (6%) reported a post-study power analysis. The result of the two-tailed Welch *t*-test indicated that there was no statistically significant difference in sample sizes between studies that performed a pre-study power analysis (median = 24) and studies which did not (median = 19) (*t* (131) = -0.94, *p* = 0.35, 95% CI for the mean difference [-68; 24], Hedge's g_s effect size corrected for bias = –0.12, 95% CI [–0.36; 0.13]. Out of 174 studies, 129 (74%) tested a hypothesis. Of those, only 39 studies (30%) included a pre-study power analysis and 8 studies (6%) included a post-power analysis**.** Amongst studies that tested a hypothesis, the result of the two-tailed Welch *t*-test indicated that there was no statistically significant difference in sample sizes between studies that performed a pre-study power analysis (median = 21) and studies which did not (median = 19) (t (106) = 0.47, p = 0.63, 95% CI for the mean difference [-5; 9], Hedge's g_s effect size corrected for bias = –0.08, 95% CI [–0.43; 0.26]. **Table 1** presents the frequency of usage of reporting practices in studies with pre-study power calculations. Results indicate that most studies did not report all components required to allow a full assessment of the pre-study analysis. The minimum components required to computationally reproduce a pre-study power analysis are the statistical test, the magnitude and type of effect size and the intended power, which, with the exception of the latter, were often nonreported. Thus, only 9 out of 46 (20%) studies that reported a pre-study analysis could be computationally reproduced. We could fully reproduce the sample size reported in 8 out of 9 pre-study power analysis. The pre-study power analysis that could not be fully reproduced reported a sample size of 61, whereas our analysis yielded a sample size of 58.

Table 1. Reporting frequencies of pre-study power analysis (n = 46)

Component reported	**Frequency** (%)
Software	27 (59)
Statistical test	10 (22)
Dependent variable	26 (57)
Effect size magnitude	30 (65)
Effect size type	25 (54)
Effect size justification	31 (67)
Alpha level	43 (93)
Intended power	45 (98)
Required sample size	41 (89)
All components	5 (11)

The types of justification for the effect size estimate used to conduct the pre-study power analyses are presented in **Table 2.** The most used justifications to select the effect size of interest were based on a previous study, followed by Cohen's d benchmark and a pilot study. The use of the two justifications considered best practice including a meta-analytic effect size and SESOI was almost non-existent.

Table 2. Justifications of the selected effect size used in the pre-study power analysis (n = 46)

Justification presented	**Frequency (%)**
Previous study	28 (61)
Pilot study	4 (9)
Meta-analysis	1 (2)
Cohen's *d* benchmark	7 (15)
SESOI	0 (0)
No justification	6 (13)

SESOI = smallest effect size of interest

The reporting practices of inferential tests are presented in **Table 3.** The most reported components were mean ± SD or mean ± SEM for both inferential tests. Other components such as test statistics and degrees of freedom were usually nonreported, although the frequency of reporting is lower for *t*-tests. Contrarily, effect sizes were reported more often for F-tests than for t-tests. CI for effect sizes were not reported in studies using F-tests, whereas in studies using t-tests CI were seldom reported.

Table 3. Frequency of reporting practices for both *F*-tests and *t*-tests

	Frequency (%)	
Component	***F*-tests** (n = 122)	***t*-test** (n = 52)
Mean ± SD / mean ± SEM	85 (70%)	40 (77%)
Test statistic	59 (48%)	10 (19%)
Degrees of freedom	46 (38%)	5 (10%)
Effect size	54 (44%)	41 (79%)
CI for effect size	0 (0%)	4 (8%)
Exact *p*-value	73 (60%)	30 (58%)
Relative *p*-value	37 (30%)	22 (42%)
No *p*-value	12 (10%)	0 (0%)

SD = standard deviation; SEM = standard error of mean; CI = confidence interval

4. Discussion

The first aim of this study was to investigate the presence of publication bias and studies with underpowered designs in a set of studies published in the *Journal of Sport Sciences*. The statistical power estimates observed in our sample of studies are not as low as in other disciplines such as psychology and neuroscience (Bakker et al., 2012; Button et al., 2013; Stanley et al., 2018; Szucs & Ioannidis, 2017). For instance, Stanley et al., (2018) reported an average power of 36% in studies included in a sample of 200 meta-analyses. The observed 73% of studies reporting a significant finding is in agreement with Twomey et al., (2021) who similarly observed that approximately 70% of the studies published in three flagship sports science journals reported significant findings.

The percentage of non-significant results is slightly higher than in many other disciplines (Fanelli, 2010; Scheel et al., 2021). For instance, Scheel et al., (2021) compared the number of significant findings reported in a sample of registered reports with a sample of standard studies in psychology and they found that 96% of significant findings in standard studies but only 44% in registered reports. The extent of publication bias in sports and exercise science is unknown. However, one estimate can be derived from investigating the difference between the percentage of significant findings and the statistical power. Assuming an average power of 61%, only about 60% of the studies investigated in our sample would detect the investigated effect as statistically significant. Yet, if we consider our study sample, we find that 73% of studies report statistically significant findings, which is at least 12 percentage points more than we should expect suggesting the presence of a biased literature. However, it is theoretically possible that the estimate of 73% significant results emerges when all studies that are performed are submitted for publication and published, or in other words, when there is no publication bias. To explain the 73% of significant results (Positive Result Rate (PRR)), we must assume some combination of statistical power and proportion of true hypotheses that researchers test (Scheel et al., 2021). The percentage of observed significant results can be computed as $\text{PRR} = \alpha \times (1 - t) + (1 - \beta) \times t$, where α is the Type 1 error rate, t is the proportion of true hypotheses and $1 - \beta$ is the power of a test (Scheel et al., 2021). Assuming no publication bias, and fixing the alpha level to 0.05, a PRR = 0.73 can be achieved with, for example, a statistical power of 96% when 75% of the hypotheses that are tested are true hypotheses. However, we observed relatively low power estimates in the sampled studies (i.e., 53% for both significant and non-significant studies, and 61% for significant studies). If we assume the upper bound (75%) of the 95% CI (0.42, 0.75) for significant findings as the true power estimate, researchers would need to test almost exclusively true hypotheses (> 95%) to observe a 73% of significant findings. Yet, these estimates of power and the proportion of true hypotheses seem overly optimistic and might not be supported by empirical evidence (Szucs & Ioannidis, 2017; Wilson & Wixted, 2018). Altogether, our results indicate the presence of some publication bias and studies with underpowered designs, which are likely to increase the number of false positives in a body literature (Ioannidis, 2005) and produce overestimated effect sizes (Bakker & Wicherts, 2011; Button et al., 2013; Kvarven et al., 2020).

The second aim was to examine the frequency of reported pre-study power analysis and their reporting practices. The low prevalence of studies with pre-study power analysis is concerning because researchers should aim to perform studies that yield informative results when they test hypotheses (as was the goal in 129 out of the 174 studies we examined). A pre-study power analysis is one important way to design studies that have a high probability to yield informative results. First, a study with an underpowered design that reports a non-significant effect is barely informative because it lacked power to find a significant effect if there was one to be found. This makes it especially difficult to publish null-findings, which contributes to publication bias. Second, studies with high-power designs yield more precise effect size estimates and reduce the uncertainty around CI. Therefore, the adoption of pre-study power analysis is one way to move the field forward (for other approaches to sample size justifications that do not rely on power analysis, see Lakens, 2022). Surprisingly, there was no significant difference in sample size between studies which included a pre-study power analysis and studies which did not include it. It is possible that this is a coincidence, but it also raises the possibility that power analyses were performed following the 'sample size samba' where researchers choose an 'expected' effect size for their power analysis that yields the sample size they wanted to collect to begin with (Schulz & Grimes, 2005). Furthermore, the similar sample sizes observed (n = 21 and n = 19 for studies with and without a pre-study power analysis that tested a hypothesis, respectively) might indicate that the effect size estimates included in the pre-study power analyses are overestimated and if all things equal, the sample size required to achieve the intended power will be smaller (Anderson et al., 2017).

We found that some studies included a post-study or 'retrospective' power analysis. This form of power analysis uses the observed effect size, the alpha level and the actual sample size to evaluate power of the study after it has been completed. However, this is not a good practice because treating the observed effect size as the true effect size in a power analysis is simply a transformation of the observed *p*-value (Hoenig & Heisey, 2001; Yuan & Maxwell, 2005; see Christogiannis et al., 2022 for a non-technical explanation). For a *t*-test, whenever the *p*-value = 0.05, post-study power will always be 50%, regardless of the combination of sample size and study effect size (Yuan & Maxwell, 2005). If a non-significant *p*-value is observed, retrospective power will always be low, regardless of the true (always unknown) power of the study (Yuan & Maxwell, 2005). These reasons render post-study power analyses uninformative, and it is better to interpret non-significant results with equivalence tests.

When pre-study power analyses were reported, the reporting practices were often suboptimal. Effect size type and magnitude, the statistical test and intended statistical power are key components to ensure reproducibility of pre-study power analysis because otherwise any attempt to reproduce such analysis would require a large amount of guesswork. For instance, omitting the statistical test used is problematic because often studies perform multiple statistical tests and thus researchers might not be able to evaluate which statistical test the power analysis was conducted for. Furthermore, power is impacted by the study design and the statistical test used (Maxwell et al., 2017). For example, within-subject statistical tests such as a paired *t*-test and a one-way within-subject ANOVA will achieve higher power in comparison to their between-subject counterparts (Maxwell et al., 2017). The omission of the dependent variable would not be problematic if studies tested only one single hypothesis that predicted the effect of a treatment or intervention on one dependent variable. However, this is far from reality because studies often test multitude of hypotheses, and a multitude of dependent variables are measured. The non-reporting of the magnitude of the effect size of interest prevents other researchers and reviewers from reproducing and evaluating the pre-study power analysis. Finally, reporting the type of effect size is important because there are several effect sizes within the same family (Goulet-Pelletier & Cousineau, 2018; Lakens, 2013; Morris & DeShon, 2002). For example, considering the simple case of a one-sample design, Cohen's *d* can be computed as d_z, d_{rm}, and d_{av} (see Lakens, 2013). Researchers should include a detailed description and justification of the steps followed to conduct the pre-study power analysis that allows other researchers and reviewers to reproduce its content and ultimately evaluate the validity of the analysis.

The process of planning the study sample size based on an effect size estimate is not as straightforward as it might seem (Bakker et al., 2016; Collins & Watt, 2021). Researchers are faced with the dilemma of justifying the effect size estimate they are interested in. This is a critical step because the magnitude of the effect size determines the sample size given an intended power. However, despite its importance in a pre-study power analysis, there is empirical data suggesting researchers have difficulties in justifying the effect size estimate for a pre-study power analysis (Bakker et al., 2016; Collins & Watt, 2021). When the effect size estimate is obtained from a previous underpowered study, it is likely that the original effect size estimate is overestimated (Bakker et al., 2012; Button et al., 2013; Simmons et al., 2011). Similarly, pilot studies are also likely to provide overestimated effect sizes (Albers & Lakens, 2018). This is problematic because the use of overestimated effect sizes for pre-study power analyses will result in studies with underpowered designs unless adjusting methods are used (see Anderson et al., 2017). The use of fixed effect sizes based on Cohen's benchmarks may not match well with the typical effect size observed in another research area because Cohen's benchmarks were derived from effects observed in behavioural science (Cohen, 1988). For instance, Swinton et al., (2022) conducted a Bayesian hierarchical meta-analysis to identify specific effect size benchmarks in strength and conditioning interventions and reported that the benchmarks for small, medium and large effect sizes were 0.12, 0.43 and 0.78, respectively. A better practice would be to obtain the effect size of interest based on a meta-analysis which can provide more accurate effect size estimates than single studies. However, to further compound the problem, some caution is needed as the quality of a meta-analysis is related to the quality of individual studies (Kvarven et al., 2020). Best practice would be to power a study based on the *smallest effect size of interest* (SESOI; see Anvari & Lakens, 2021; Lakens, 2022). Thus, instead of conducting a pre-study power analysis based on the effect size estimate that the researcher expects to observe, researchers should rely on the *smallest effect* that they consider theoretically or practically meaningful. However, none of the studies sampled did so. Researchers might benefit from consulting a statistician if they find it challenging to determine the required sample size for a future study, and researchers in sports and exercise science might want to start a discussion about which effect sizes are deemed large enough to matter, so that future studies can be designed to detect the presence *or absence* of the smallest effect size of interest.

The third aim was to investigate the reporting practices of inferential tests. Overall, reporting practices of statistical results were suboptimal and journals and researchers should adopt the journal article reporting standards recommended by APA (Appelbaum et al., 2018). Following APA standards, results of inferential tests should be reported in the following order: the *F*-ratio or *t*-statistic and degrees of freedom (in parentheses) followed by the exact *p*-value (e.g., $F(1,35) = 5.45$, $p = 0.001$ or $t(85) = 2.86$, $p = 0.025$). This would be beneficial for a few reasons. First, the reporting of the *F*-ratio or *t*-statistic and degrees of freedom allow to recompute the *p*-value reported and therefore verify the reported *p*-value. This and data sharing is of importance when there is evidence that one in eight papers contained errors in the reported *p*-value that may have affected the statistical conclusion of the study (Nuijten et al., 2016; see also Artner et al., 2021 for a summary of studies on this topic). From an epistemological point of view, reproducibility should be assessed before replicability because it makes little sense to try to replicate a prior finding if the results supporting the finding are numerically incorrect. Second, both the *F*-ratio and *t*-statistic can be used to compute the effect size estimate (see Lakens, 2013). For instance, the reporting of the *F*-ratio and degrees of freedom allows computation of eta partial squared (η_p^2; e.g., $F(1,35) = 5.45$, $\eta_p^2 = 5.45 \times 1/(5.45 \times 1 + 35)$). Third, it would facilitate machine readability and data usability enabling the analysis of large sets of data containing *p*-values. Methods such as *p*-curve and z-curve that can be used to address meta-scientific questions require the input of exact *p*-values, which are not always reported. Therefore, researchers should fully report the statistical results of inferential tests with the goal of facilitating computational reproducibility and allow other researchers to assess the veracity of published results.

The omission of (standardized) effect size estimates and their CI is concerning for a few reasons. First, effect size estimates allow researchers to make a judgement on the practical significance of the magnitude of the studied effect (Asendorpf et al., 2013; Kelley & Preacher, 2012; Schäfer & Schwarz, 2019). Second, effect size estimates can be used to conduct pre-study power analysis for follow-up studies (Cohen, 1988; Lakens, 2022; Schäfer & Schwarz, 2019). Third, (standardized) effect size estimates permit direct comparison across similar studies that collected dependent variables on different raw scales, and can be used in meta-analysis to draw meta-analytic conclusions. Fourth, when researchers report effect sizes estimates, researchers should acknowledge and quantify the uncertainty in these estimates. CIs provide information of how accurately a true effect size was estimated (Asendorpf et al., 2013; Kelley & Preacher, 2012). This is especially of interest if studies have small sample sizes because such studies suffer from large sampling error which leads to substantial uncertainty around the true effect size. For instance, imagine a researcher that conducted a study with a two-cell design where there are 10 participants per condition, and reported a significant Cohen's d_S of 0.5 omitting its 95% CI [0.05; 1.05]. Although the observed effect size and *p*-value were reported, the uncertainty around the estimate makes clear that the test was not very informative about the true effect size. Therefore, researchers should follow the journal article reporting standards recommended by APA (Appelbaum et al., 2018) and report both effect sizes estimates and their CI.

Our investigation has a few limitations that should be addressed herein. Firstly, our selection is a pilot 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 and exercise science, and should be considered a pilot study for a more comprehensive examination in the future. Furthermore, the small sample of studies included (n = 89) increased the uncertainty around the parameter estimates (Brunner & Schimmack, 2020). Secondly, the z-curve analysis included only studies that tested a hypothesis but the distinction between the former and descriptive studies was sometimes ambiguous. This could be resolved if authors stated explicitly whether the study was intended to be hypothesis-testing or hypothesis-generating in the methods section. Thirdly, 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 Wicherts et al., 2016 for researchers' degrees of freedom). Fourthly, although two secondary authors undertook some random verification of the data selected (D.L. verified some coded data for z-curve analysis and J.W. verified some coded data for the reporting practices and reproducibility of the pre-study power analysis), 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. Finally, the leading author acknowledges that this study should have been preregistered despite its exploratory nature.

Overall, our results suggest that there are substantial barriers that would hinder both computational reproducibility and replicability. First, the point estimate of the Observed Discovery Rate (0.73) lies outside the 95% CI of the Expected Discovery Rate [0.13; 0.71] suggesting the presence of publication bias. Second, the two power estimates indicate that the sampled studies had, on average, inadequately powered designs (as a Type 2 error rate of 40% should be considered too high). Third, the low usage of pre-study power analyses as well as the use of effect size estimates obtained from previous studies or pilot studies is problematic given the small samples observed in the field of sport and exercise science (Abt et al., 2020) and the issues with overestimated effect sizes as a result (Albers & Lakens, 2018; Anderson et al., 2017). Fourth, the reporting practices of pre-study power analyses and inferential tests were often suboptimal preventing researchers from assessing the validity of the results. Therefore, it seems there is substantial opportunity to improve researchers' behaviours through the adoption of Open Science practices such as sample size planning based on a pre-study power analysis and full reporting of statistical results, if the scientific community is to improve these factors in the future.

5. References

Abt, G., Boreham, C., Davison, G., Jackson, R., Nevill, A., Wallace, E., & Williams, M. (2020). Power, precision, and sample size estimation in sport and exercise science research. *Journal of Sports Sciences*, *38*(17), 1933–1935. https://doi.org/10.1080/02640414.2020.1776002

Albers, C., & Lakens, D. (2018). When power analyses based on pilot data are biased: Inaccurate effect size estimators and follow-up bias. *Journal of Experimental Social Psychology*, *74*, 187–195. https://doi.org/10.1016/j.jesp.2017.09.004

Anderson, S. F., Kelley, K., & Maxwell, S. E. (2017). Sample-Size Planning for More Accurate Statistical Power: A Method Adjusting Sample Effect Sizes for Publication Bias and Uncertainty. *Psychological Science*, *28*(11), 1547–1562. https://doi.org/10.1177/0956797617723724

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

Appelbaum, M., Cooper, H., Kline, R. B., Mayo-Wilson, E., Nezu, A. M., & Rao, S. M. (2018). Journal article reporting standards for quantitative research in psychology: The APA Publications and Communications Board task force report. *American Psychologist*, *73*(1), 3. https://doi.org/10.1037/amp0000191

Artner, R., Verliefde, T., Steegen, S., Gomes, S., Traets, F., Tuerlinckx, F., & Vanpaemel, W. (2021). The reproducibility of statistical results in psychological research: An investigation using unpublished raw data. *Psychological Methods*, *26*(5), 527–546. https://doi.org/10.1037/met0000365

Asendorpf, J. B., Conner, M., Fruyt, F. D., Houwer, J. D., Denissen, J. J. A., Fiedler, K., Fiedler, S., Funder, D. C., Kliegl, R., Nosek, B. A., Perugini, M., Roberts, B. W., Schmitt, M., Aken, M. A. G. van, Weber, H., & Wicherts, J. M. (2013). Recommendations for Increasing Replicability in Psychology. *European Journal of Personality*, *27*(2), 108–119. https://doi.org/10.1002/per.1919

Bakker, M., Hartgerink, C. H. J., Wicherts, J. M., & van der Maas, H. L. J. (2016). Researchers' Intuitions About Power in Psychological Research. *Psychological Science*, *27*(8), 1069–1977. https://doi.org/10.1177/0956797616647519

Bakker, M., van Dijk, A., & Wicherts, J. M. (2012). The Rules of the Game Called Psychological Science. *Perspectives on Psychological Science*, *7*(6), 543–554. https://doi.org/10.1177/1745691612459060

Bakker, M., & Wicherts, J. M. (2011). The (mis)reporting of statistical results in psychology journals. *Behavior Research Methods*, *43*(3), 666–678. https://doi.org/10.3758/s13428-011-0089-5

Bartoš, F., & Schimmack, U. (2022). Z-curve 2.0: Estimating Replication Rates and Discovery Rates. *Meta-Psychology*, *6*. https://doi.org/10.15626/MP.2021.2720

Brunner, J., & Schimmack, U. (2020). Estimating Population Mean Power Under Conditions of Heterogeneity and Selection for Significance. *Meta-Psychology*. https://doi.org/10.15626/MP.2018.874

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*(5), 365–376. https://doi.org/10.1038/nrn3475

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Nave, G., Nosek, B. A., Pfeiffer, T., Altmejd, A., Buttrick, N., Chan, T., Chen, Y., Forsell, E., Gampa, A., Heikensten, E., Hummer, L., Imai, T., … Wu, H. (2018). Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour*, *2*(9), 637–644. https://doi.org/10.1038/s41562-018-0399-z

Carter, E. C., & McCullough, M. E. (2014). Publication bias and the limited strength model of self-control: Has the evidence for ego depletion been overestimated? *Frontiers in Psychology*, *5*. https://doi.org/10.3389/fpsyg.2014.00823

Christogiannis, C., Nikolakopoulos, S., Pandis, N., & Mavridis, D. (2022). The self-fulfilling prophecy of post-hoc power calculations. *American Journal of Orthodontics and Dentofacial Orthopedics: Official Publication of the American Association of Orthodontists, Its Constituent Societies, and the American Board of Orthodontics*, *161*(2), 315–317. https://doi.org/10.1016/j.ajodo.2021.10.008

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal and Social Psychology*, *65*, 145–153. https://doi.org/10.1037/h0045186

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

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

Cumming, G. (2013). *Understanding The New Statistics: Effect Sizes, Confidence Intervals, and Meta-Analysis*. Routledge.

Delacre, M., Lakens, D., Ley, C., Liu, L., & Leys, C. (2021). *Why Hedges' g*s based on the non-pooled standard deviation should be reported with Welch's t-test*. PsyArXiv. https://doi.org/10.31234/osf.io/tu6mp

Errington, T. M., Denis, A., Perfito, N., Iorns, E., & Nosek, B. A. (2021). Challenges for assessing replicability in preclinical cancer biology. *ELife*, *10*, e67995. https://doi.org/10.7554/eLife.67995

Errington, T. M., Mathur, M., Soderberg, C. K., Denis, A., Perfito, N., Iorns, E., & Nosek, B. A. (2021). Investigating the replicability of preclinical cancer biology. *ELife*, *10*, e71601. https://doi.org/10.7554/eLife.71601

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

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

Francis, G. (2012). Publication bias and the failure of replication in experimental psychology. *Psychonomic Bulletin & Review*, *19*(6), 975–991. https://doi.org/10.3758/s13423-012-0322-y

Franco, A., Malhotra, N., & Simonovits, G. (2014). Publication bias in the social sciences: Unlocking the file drawer. *Science*, *345*(6203), 1502–1505. https://doi.org/10.1126/science.1255484

Götz, F. M., Gosling, S. D., & Rentfrow, P. J. (2022). Small Effects: The Indispensable Foundation for a Cumulative Psychological Science. *Perspectives on Psychological Science*, *17*(1), 205–215. https://doi.org/10.1177/1745691620984483

Goulet-Pelletier, J.-C., & Cousineau, D. (2018). A review of effect sizes and their confidence intervals, Part I: The Cohen's d family. *The Quantitative Methods for Psychology*, *14*(4), 242–265. https://doi.org/10.20982/tqmp.14.4.p242

Greenwald, A. G. (1975). Consequences of prejudice against the null hypothesis. *Psychological Bulletin*, *82*, 1–20. https://doi.org/10.1037/h0076157

Hoenig, J. M., & Heisey, D. M. (2001). The Abuse of Power: The Pervasive Fallacy of Power Calculations for Data Analysis. *The American Statistician*, *55*(1), 19–24. https://doi.org/10.1198/000313001300339897

Ioannidis, J. P. A. (2005). Why Most Published Research Findings Are False. *PLOS Medicine*, *2*(8), e124. https://doi.org/10.1371/journal.pmed.0020124

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the Prevalence of Questionable Research Practices With Incentives for Truth Telling. *Psychological Science*, *23*(5), 524–532. https://doi.org/10.1177/0956797611430953

Kelley, K., & Preacher, K. J. (2012). On effect size. *Psychological Methods*, *17*, 137–152. https://doi.org/10.1037/a0028086

Kvarven, A., Strømland, E., & Johannesson, M. (2020). Comparing meta-analyses and preregistered multiple-laboratory replication projects. *Nature Human Behaviour*, *4*(4), 423–434. https://doi.org/10.1038/s41562-019-0787-z

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. *Frontiers in Psychology*, *4*, 863. https://doi.org/10.3389/fpsyg.2013.00863

Lakens, D. (2022). Sample Size Justification. *Collabra: Psychology*, *8*(1), 33267. https://doi.org/10.1525/collabra.33267

Lakens, D., & Evers, E. R. K. (2014). Sailing From the Seas of Chaos Into the Corridor of Stability: Practical Recommendations to Increase the Informational Value of Studies. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, *9*(3), 278–292. https://doi.org/10.1177/1745691614528520

Mahoney, M. J. (1977). Publication prejudices: An experimental study of confirmatory bias in the peer review system. *Cognitive Therapy and Research*, *1*, 161–175. https://doi.org/10.1007/BF01173636

Maxwell, S. E., Delaney, H. D., & Kelley, K. (2017). *Designing experiments and analyzing data: A model comparison perspective* (3rd ed). Routledge.

Morris, S. B., & DeShon, R. P. (2002). Combining effect size estimates in meta-analysis with repeated measures and independent-groups designs. *Psychological Methods*, *7*(1), 105–125. https://doi.org/10.1037/1082-989x.7.1.105

Murphy, J., Mesquida, C., Caldwell, A. R., Earp, B. D., & Warne, J. P. (2022). Proposal of a Selection Protocol for Replication of Studies in Sports and Exercise Science. *Sports Medicine*. https://doi.org/10.1007/s40279-022-01749-1

Nosek, B. A., Hardwicke, T. E., Moshontz, H., Allard, A., Corker, K. S., Dreber, A., Fidler, F., Hilgard, J., Struhl, M. K., Nuijten, M. B., Rohrer, J. M., Romero, F., Scheel, A. M., Scherer, L. D., Schönbrodt, F. D., &

Vazire, S. (2022). Replicability, Robustness, and Reproducibility in Psychological Science. *Annual Review of Psychology*, *73*(1), 719–748. https://doi.org/10.1146/annurev-psych-020821-114157

Nuijten, M. B., Hartgerink, C. H. J., van Assen, M. A. L. M., Epskamp, S., & Wicherts, J. M. (2016). The prevalence of statistical reporting errors in psychology (1985–2013). *Behavior Research Methods*, *48*, 1205–1225. https://doi.org/10.3758/s13428-015-0664-2

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716. https://doi.org/10.1126/science.aac4716

Primbs, M. A., Pennington, C. R., Lakens, D., Silan, M. A. A., Lieck, D. S. N., Forscher, P. S., Buchanan, E. M., & Westwood, S. J. (2022). Are Small Effects the Indispensable Foundation for a Cumulative Psychological Science? A Reply to Götz et al. (2022). *Perspectives on Psychological Science*, 17456916221100420. https://doi.org/10.1177/17456916221100420

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, *83*(3), 638–641. https://doi.org/10.1037/0033-2909.86.3.638

Schäfer, T., & Schwarz, M. A. (2019). The Meaningfulness of Effect Sizes in Psychological Research: Differences Between Sub-Disciplines and the Impact of Potential Biases. *Frontiers in Psychology*, *10*, 813. https://doi.org/10.3389/fpsyg.2019.00813

Scheel, A. M., Schijen, M. R. M. J., & Lakens, D. (2021). An Excess of Positive Results: Comparing the Standard Psychology Literature With Registered Reports. *Advances in Methods and Practices in Psychological Science*, *4*(2), 1–12. https://doi.org/10.1177/25152459211007467

Schulz, K. F., & Grimes, D. A. (2005). Sample size calculations in randomised trials: Mandatory and mystical. *The Lancet*, *365*(9467), 1348–1353. https://doi.org/10.1016/S0140-6736(05)61034-3

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632

Simonsohn, U. (2015). Small Telescopes: Detectability and the Evaluation of Replication Results. *Psychological Science*, *26*(5), 559–569. https://doi.org/10.1177/0956797614567341

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014a). P-curve: A key to the file-drawer. *Journal of Experimental Psychology: General*, *143*(2), 534–547. https://doi.org/10.1037/a0033242

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014b). p-Curve and Effect Size: Correcting for Publication Bias Using Only Significant Results. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, *9*(6), 666–681. https://doi.org/10.1177/1745691614553988

Stanley, T. D., Carter, E. C., & Doucouliagos, H. (2018). What meta-analyses reveal about the replicability of psychological research. *Psychological Bulletin*, *144*(12), 1325–1346. https://doi.org/10.1037/bul0000169

Stefan, A., & Schönbrodt, F. (2022). *Big Little Lies: A Compendium and Simulation of p-Hacking Strategies*. PsyArXiv. https://doi.org/10.31234/osf.io/xy2dk

Swinton, P. A., Burgess, K., Hall, A., Greig, L., Psyllas, J., Aspe, R., Maughan, P., & Murphy, A. (2022). Interpreting magnitude of change in strength and conditioning: Effect size selection, threshold values and Bayesian updating. *Journal of Sports Sciences*, 1–8. https://doi.org/10.1080/02640414.2022.2128548

Szucs, D., & Ioannidis, J. P. A. (2017). Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. *PLoS Biology*, *19*(3), e3001151. https://doi.org/10.1371/journal.pbio.2000797

Twomey, R., Yingling, V., Warne, J., Schneider, C., McCrum, C., Atkins, W., Murphy, J., Medina, C. R., Harlley, S., & Caldwell, A. (2021). The Nature of Our Literature: A Registered Report on the Positive Result Rate and Reporting Practices in Kinesiology. *Communications in Kinesiology*, *1*(3), 1–17. https://doi.org/10.51224/cik.v1i3.43

Wacholder, S., Chanock, S., Garcia-Closas, M., El ghormli, L., & Rothman, N. (2004). Assessing the Probability That a Positive Report is False: An Approach for Molecular Epidemiology Studies. *JNCI: Journal of the National Cancer Institute*, *96*(6), 434–442. https://doi.org/10.1093/jnci/djh075

Wicherts, J. M., Veldkamp, C. L. S., Augusteijn, H. E. M., Bakker, M., van Aert, R. C. M., & van Assen, M. A. L. M. (2016). Degrees of Freedom in Planning, Running, Analyzing, and Reporting Psychological Studies: A Checklist to Avoid p-Hacking. *Frontiers in Psychology*, *7*, 1832. https://doi.org/10.3389/fpsyg.2016.01832

Wilson, B. M., & Wixted, J. T. (2018). The Prior Odds of Testing a True Effect in Cognitive and Social Psychology. *Advances in Methods and Practices in Psychological Science*, *1*(2), 186–197. https://doi.org/10.1177/2515245918767122

Yuan, K.-H., & Maxwell, S. (2005). On the Post Hoc Power in Testing Mean Differences. *Journal of Educational and Behavioral Statistics Summer*, *30*, 141–167. https://doi.org/10.3102/10769986030002141