On the replicability of sports and exercise science research: Assessing the prevalence of selection bias and studies with underpowered designs by a z-curve analysis

Cristian Mesquida[12] Jennifer Murphy[2] Joe Warne[2] Daniël Lakens[1]

Affiliation

[1] Human-Technology Interaction Group, Eindhoven University of Technology, Eindhoven, The Netherlands

[2] School of Biological, Health and Sports Sciences, Technological University Dublin, Tallaght, Dublin, Ireland

ORCIDs

Cristian Mesquida – 0000-0002-1542-8355

Jennifer Murphy – 0000-0001-8624-3828

Joe Warne – 0000-0002-4359-8132

Daniël Lakens – 0000-0002-0247-239X

Correspondence

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

Abstract

The Sports Science Replication Project has raised concerns about the replicability of published research. Low replication rates may result from an excess of significant findings caused by selection bias, where studies reporting significant findings are more likely to be published, inflating the proportion of significant findings in the literature, while the statistical power remains low. To date, no study has systematically assessed both selection bias and average statistical power in the same set of studies in the field. One method to assess selection bias and average statistical power is the z-curve method. In this study, we manually extracted 350 independent p-values corresponding to the hypothesis tested in 350 studies published across 10 applied sports and exercise science journals. After exclusions, a z-curve analysis was performed on 269 independent p-values. The estimate of the Observed Discovery Rate (68%) is larger than the upper bound of the 95% confidence intervals (CI) of the Expected Discovery Rate of [5; 34%] indicating strong publication bias in the literature. The average statistical power is 11% 95% CI [5; 34%], and only 29% of studies are estimated to have been designed with high power ($\geq$ 80%). The Expected Replication Rate was 50% 95% CI [37; 61%], indicating that only 50% of direct replications with the same sample size should be expected to replicate. Selection bias, combined with low average statistical power, is likely to result in a body of literature characterized by inflated effect sizes, a high proportion of type I and type II errors, and therefore low replicability. Addressing these issues requires a collective effort to build a more informative and reliable knowledge base.

Key points

The Observed Discovery Rate (68%) far exceeds the Expected Discovery Rate 95% [5; 34%], indicating significant selection bias in sports and exercise science research.

The average statistical power across studies is only 11% 95% CI [5; 34%], with just 29% of studies designed with high power, suggesting many findings may be unreliable.

The average power of studies supporting the hypothesis is 50% implying that fewer than half of published significant findings are likely to replicate under the same conditions and sample size, undermining the replicability of the published literature.

Introduction

To appropriately evaluate how well scientific claims are empirically supported, it is essential that all observed results are reported in the published scientific literature. Problematically, there are clear signs of selective reporting, and statistically significant results are much more likely to be published than non-significant results [1,2]. Meta-scientific research has observed that between 73% and 81% of published studies in sports and exercise science journals report a statistically significant effect that supports the hypothesis that the researchers set out to test [3–5]. Is this estimate too high, and a possible sign of bias in the literature, or is it in line with what should be expected in an unbiased literature? The answer to this question depends on two unknown properties: the statistical power of the studies (henceforth, power), and the proportion of hypotheses that test true effects in the published literature [1,6]. Power is the probability of observing a statistically significant effect if there is a true effect, and in turn depends on the sample size, the effect size, the statistical test that is performed, and the alpha level (α). Although the power of studies is unknown, the average power of studies can be meta-analytically estimated under specific assumptions. Although the proportion of tested hypotheses that examine true effects is also unknown, we can examine how high the proportion of hypotheses testing true effects would need to be to achieve the observed rate of significant effects in the literature, given an estimate of the average statistical power. By comparing the proportion of studies supporting their tested hypotheses with the average power of those studies, we offer insights about the plausibility that selection bias is one of the contributing factors of low replicability in sports and exercise science.

The percentage of significant findings in a set of studies, also referred to as the Observed Discovery Rate (ODR), can be computed as follows:

$$ODR = \alpha \times (1 - t) + (1 - \beta) \times t \tag{1}$$

where α is the alpha level, t is the proportion of true hypotheses, β is the type II error and $1 - \beta$ is the power of a test [1]. The Observed Discovery Rate in sports and exercise science of 73% to 81% can therefore result from a range of combinations of the power of tests and proportions of true hypotheses. For example, an Observed Discovery Rate of 73% can be achieved when 100% of the hypotheses tested are true, and the average power of studies is 73%, or when 73% of the hypotheses tested are true, and the average power of studies is close to 100%, or any combination in between these two extremes. Figure 1 visualizes the relationship between power and the proportion of true hypotheses, and as illustrated, the lower the proportion of true hypotheses is, the higher the power of the tests must be.

The assumption that sports and exercise scientists almost exclusively examine true hypotheses seems unrealistic, and this assumption is not in line with empirical evidence from other fields [7,8], or with the results of the sports science replication project [9]. A field that only studies true hypotheses is arguably not sufficiently pushing the boundaries of our current understanding. Less is known about how reasonable the assumption is that studies in sports and exercise science achieve an average power of 73%. If this is not the case, then the high percentage of significant findings could only be explained by bias towards significant effects in the published literature. Given the small sample sizes reported in the field [4,10], it seems doubtful that the average power could be as high as 73%. However, high power could be achieved if studies investigate large effects, use within-subject designs or repeated measures, or include appropriate covariates to account for a greater proportion of the error variance. The aim of the current study is to provide the first systematic assessment of the average power of the published sports and exercise science literature to inform future discussions about the presence of selection bias in the field, and its potential effect on the replicability of sports and exercise science.

It is important to distinguish between two broad categories of bias: "publication bias" [11,12] and "p-hacking" [13]. Publication bias occurs when the studies in the scientific literature are systematically unrepresentative of the studies that are performed. It is often caused by the tendency of editors, reviewers, and researchers to prefer studies that support the hypothesis tested over those that fail to support it (e.g., significance bias). In the context of NHST, a study reporting a significant p-value would provide support in favor of the hypothesis tested, and thus would be more likely to get published than a study that failed to support the same hypothesis.

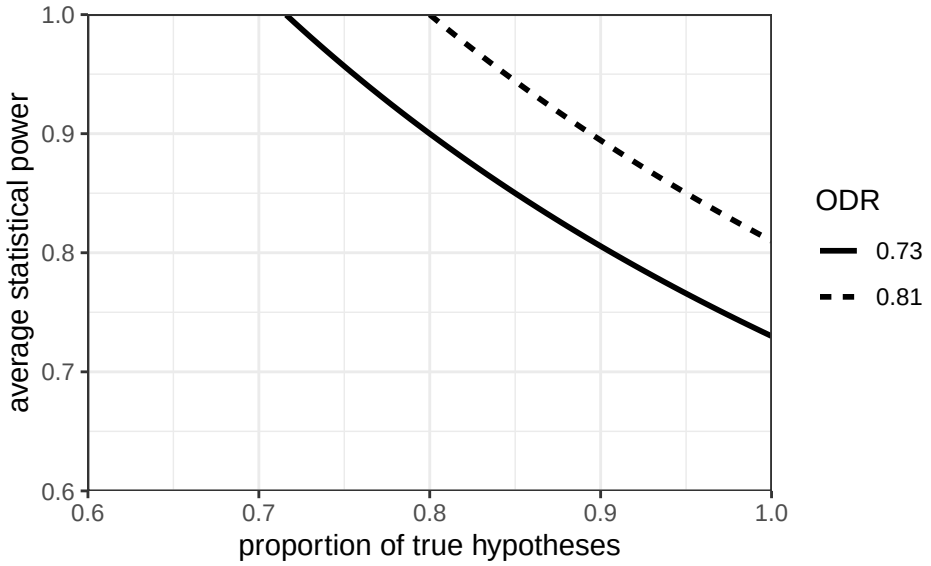

Fig. 1 Combinations of the proportion of true hypotheses (x-axis) and power (y-axis) required to produce 73% or 81% statistically significant findings assuming an alpha level of $\alpha = 0.05$ and no bias. Notably, achieving 73% significant findings requires the average power to be at least 73% if all tested hypotheses are true.

P-hacking can be defined as a set of problematic practices that opportunistically exploit flexibility in data collection and analysis to render non-significant findings significant. Surveys among scientists suggest that *p*-hacking is widespread across disciplines [14], and therefore, there is no reason to believe that sport and exercise science is an exception. In a research environment shaped by publication pressures and incentives that reward significant findings, researchers might resort to *p*-hacking to render their non-significant effect significant and thus increase their chances of publication. A main consequence of publication bias and *p*-hacking is that the percentage of significant findings in the literature is higher than the average power of studies, contributing to an excess of significant findings that reflect type I errors. An excess of significant findings has been reported in sports therapy and rehabilitation [15] and in the *Journal of Sports Sciences* [4]. Whether these findings can be generalized to the field of sport and exercise science has yet to be established and is the primary focus of the current study.

Z-curve

One method that models the presence of selection bias and identifies underpowered designs from the distribution of *p*-values is the *z*-curve method [16]. Briefly, the *z*-curve method transforms reported *p*-values into absolute *z*-scores and compares the observed and expected distribution of *z*-scores. If these two distributions are sufficiently similar, there is no indication of bias, whereas large differences between the observed and expected distribution suggest the presence of bias. The *z*-curve method estimates 4 quantities that provide insights into the replicability of a literature, under specific assumptions [6]. First, *z*-curve assumes that observed *z*-scores are obtained from multiple sampling distributions with different means, allowing for heterogeneity in power estimates. This makes *z*-curve a better choice for our study as opposed to the related *p*-curve analysis, since high heterogeneity can be expected when studies are selected from different sub-disciplines such as sports performance, exercise physiology, biomechanics, and sports psychology.

Second, *z*-curve assumes that all *p*-values are independent. This assumption is met if, as in our study, only one *p*-value per study is included in the *z*-curve analysis. Finally, *z*-curve assumes that all studies used the same criterion for statistical significance ($\alpha = 0.05$). Thus, if a study corrected for multiple comparisons or used a more conservative criterion (e.g., $\alpha = 0.01$), bias is modeled based on an α of 0.05 instead of the actual, more conservative criterion. This is a conservative assumption, as *z*-curve will overestimate replicability when the assumed α is higher than the actual α. The 4 quantities computed in a *z*-curve analysis are described

in Table 1.

Table 1 Description of the 4 quantities estimated by z-curve.

Quantity	Description
Observed Discovery Rate (ODR)	The ODR is the rate of studies reporting a significant p-value that would support the hypothesis tested. However, this rate is not necessarily an accurate reflection of true effects because this rate also includes type I errors. For instance, imagine that researchers test 100 hypotheses of which 70 correspond to true effects and 30 correspond to null effects. Furthermore, assume that researchers design their studies with 80% power and set α to 0.05. Using Equation 1, the ODR would be 58% ($0.05 \times (1 - 0.7) + (1 - 0.2) \times 0.7 = 58$). Out of these 58 significant effects, 56 would correspond to true effects (70 x 0.8) and ~ 2 would correspond to type I errors (30 x 0.05). Furthermore, publication bias inflates the ODR. Following with the previous example, if 15 out of the 30 null effects were not published due to publication bias, the ODR would be 68% ((56 + 2) / 85 = 68). Publication bias and p-hacking increase the proportion of type I errors, thereby inflating the ODR.
Expected Discovery Rate (EDR)	The EDR is an estimate of the average power of all studies included in the z-curve analysis. For example, if we have 3 studies designed with 10%, 50% and 90% power, the EDR would be 50% ((10 + 50 + 90)/3). The EDR can be compared to the ODR to determine the presence of selection bias. If the point estimate of the ODR is larger than the upper bound of the 95% CI of the EDR, we can statistically reject the hypothesis that there is no selection bias [16].
Expected Replication Rate (ERR)	The ERR is an estimate of the average power of studies reporting a significant p-value. This estimate reflects the probability of observing the same significant effect if the study were to be replicated using the same sample size and following the same procedures. Using the prior example, if the study with a 90% power was the only one that yielded a significant finding, then the ERR would be 90%. Thus, the higher the ERR, the more likely it is that the studies that reported a significant effect would replicate.

Quantity	Description
Maximum False Discovery Risk (MFDR)	The MFDR is an estimate of the maximum percentage of significant findings that are type I errors. Importantly, the MFDR does not aim to estimate the actual type I error rate among significant p-values. Rather, it provides an estimate of the worst-case scenario with the highest possible proportion of type I errors. If a literature has a low MFDR, readers can be assured that most significant findings are true effects. The MFDR is estimated using the Expected Discovery Rate and α, and it is computed as: MFDR = ((1 - EDR) / EDR) × (α / (1 - α)). For example, with an EDR of 50% and α = 0.05, the MFDR is 0.053, which is close to the nominal α set to 0.05.

Simulations

The distribution of p-values (and therefore the distribution of z-scores) in the published literature is determined by four factors, namely, the proportion of studies that investigate true and null effects, the power of studies that investigate true effects, publication bias, and p-hacking. To help readers understand which findings to expect from a z-curve analysis, we first simulate 300 p-values to represent six distinct scenarios, and then conduct the corresponding z-curve analysis. The rationale behind these scenarios is to provide a diverse set of conditions, illustrating how the distribution of z-scores is affected by power and selection bias. We simulated 300 p-values because this closely represents the number of p-values reported in our study (i.e., n of 269 after exclusions). For all six scenarios, the p-values were generated using an unpaired one-tailed t-test and α of 0.05. The code for the simulations and the z-curve analyses is available at https://doi.org/10.17605/OSF.IO/SFBVA.

In the first scenario (Figure 2a; "80% power"), p-values were simulated assuming a true effect size (Cohen's d_s) of 0.3 and a total sample size (n) of 278, yielding approximately 80%. In the second scenario (Figure 2b; "Null effect"), p-values were simulated under a true effect size of 0; consequently the proportion of significant findings corresponds to the nominal significance level ($\alpha = 0$). All significant findings in this scenario therefore represent type I errors, as no true effect exists. In the third scenario (Figure 2c; "Publication bias and 20% power"), p-values were simulated assuming a true effect size of 0.3 and n of 30, yielding approximately 20% power. In addition, publication bias was introduced such that 40% of the non-significant findings remained unpublished. In the fourth scenario (Figure 2d; "Publication bias and null effect"), p-values were simulated assuming a true effect size of 0 and n of 30. Additionally, publication bias is introduced such that 90% of the non-significant findings remained unpublished. In the fifth scenario (Figure 2e; "Mild optional stopping"), p-values were simulated assuming a true effect size of 0 and a mild optional stopping strategy in which researchers conducted up to five hypothesis tests. Specifically, researchers repeatedly performed a hypothesis test after adding new participants until either a maximum sample size ($n = 50$) was reached or a significant p-value is observed, without correcting α for multiple comparisons. In the final scenario (Figure 2f; "Mixed"), 300 p-values were simulated as follows: 100 based on a true effect size of 0.3 and n of 278, 100 based on a true effect size of 0.3 and n of 100, and 100 based on a true effect size of 0.3 and n of 26. Subsequently, 100 non-significant p-values were randomly replaced with p-values obtained through severe optional stopping. The dresulting distributions of z-scores for each scenario are shown in Figure 2, and the corresponding results of each z-curve are reported in Table 2.

Table 2 Output estimates [95% CI] for the z-curves conducted under six scenarios.

Scenario	ODR	ERR	EDR	Soric FDR
1	82 [77; 86]	80 [70; 88]	78 [64; 90]	2 [1; 3]

Table 2 Output estimates [95% CI] for the z-curves conducted under six scenarios.

Scenario	ODR	ERR	EDR	Soric FDR
2	4 [2; 7]	3 [2; 6]	5 [5; 10]	100 [47; 100]
3	30 [25; 36]	10 [2; 21]	7 [5; 22]	65 [19; 100]
4	33 [27; 38]	10 [2; 22]	5 [5; 12]	91 [40; 100]
5	14 [10; 19]	3 [2; 6]	5 [5; 10]	100 [47; 100]
6	62 [57; 68]	35 [23; 46]	9 [5; 20]	51 [21; 100]

1 = 80% power; 2 = Null effect, 3 = Publication bias and 20% power; 4 = Publication bias and null effect; 5 = Mild optional stopping; 6 = Mixed

The results of these simulations illustrate several key points. In the absence of bias, when studies are designed with 80% power, the average power of studies matches the Observed Discovery Rate, whose estimate lies within the 95% CI of the Expected Discovery Rate. The Maximum False Discovery Risk is close to α because studies are designed with high power. When studies investigate a null effect, the Observed Discovery Rate corresponds to α, and the Maximum False Discovery Risk is 1, as all significant findings are type I errors. In the presence of publication bias and studies with 20% power, the Observed Discovery Rate becomes inflated and exceeds the upper limit of the Expected Discovery Rate 95% CI. This discrepancy indicates bias in the set of studies. Furthermore, when the Expected Discovery Rate does not exclude 5%, it suggests that all observed effects may be type I errors. Under a true effect size of zero combined with publication bias, the Observed Discovery Rate becomes substantially larger than α. In the fifth scenario, where optional stopping was simulated under a true effect size of zero, the Observed Discovery is again inflated beyond α. The fact that the Observed Discovery Rate is higher than the upper bound of the Expected Discovery Rate 95% CI further signals bias. It is important to note, however, that although the z-curve method can be used to assess publication bias, it is not developed to identify p-hacking, and the z-curve method might not be able to distinguish between publication bias and p-hacking. In this scenario, the Expected Discovery Rate is 5% which corresponds to the expected type I error under the null distribution. When the Expected Discovery Rate does not exclude 5%, it suggests that all effects might be, in fact, null effects. Additionally, the Maximum False Discovery Risk is 1, in line with the fact that all significant findings are type I errors. Finally, in the mixed scenario, the Observed Discovery Rate exceeds the upper bound of the Expected Discovery Rate 95% CI, indicating the presence of bias, and underscoring the limitation that the z-curve method cannot differentiate between publication bias and p-hacking.

To sum up, the z-curve method can be used to distinguish between an unbiased and biased published literature by comparing the Observed Discovery Rate, the Expected Discovery Rate and the Expected Replication Rate. In the absence of bias and high average power, the Observed Discovery Rate should lie inside the 95% CI of the Expected Discovery Rate and the higher the power, the more z-scores should be larger than 3 (i.e., $p < 0.001$). In such case, the published literature is characterized by studies investigating true effects with high-power designs and therefore it should be expected to be highly replicable in direct replications. In contrast, when the Observed Discovery Rate is larger than the upper bound of the 95% CI of the Expected Discovery Rate, the published literature is biased and researchers have reasons to doubt the likelihood that effects will replicate. Put more simply, the blue solid line shows the expected distribution of z-scores in all simulations. When there is a noticeable deficit of observed z-scores below this line—particularly to the left of the red line at $z = 1.96$, which marks the threshold for statistical significance—the model would indicate evidence of bias.

To date, no study in sports and exercise science has jointly examined the proportion of statistically significant findings and the average power within the same sample of studies in sports and exercise science, a condition that is necessary to assess whether there is an excess of significant results. Moreover, although the recently reported replication rate of the Sports Science Replication Project [9] provides empirical evidence of low replicability in the field, skeptical sport and exercise scientists may argue that these findings are not representative due to limited number of replications conducted. Alternatively, replication failures may be attributed to deviations from the original studies, underpowered replciation designs, unaccounted

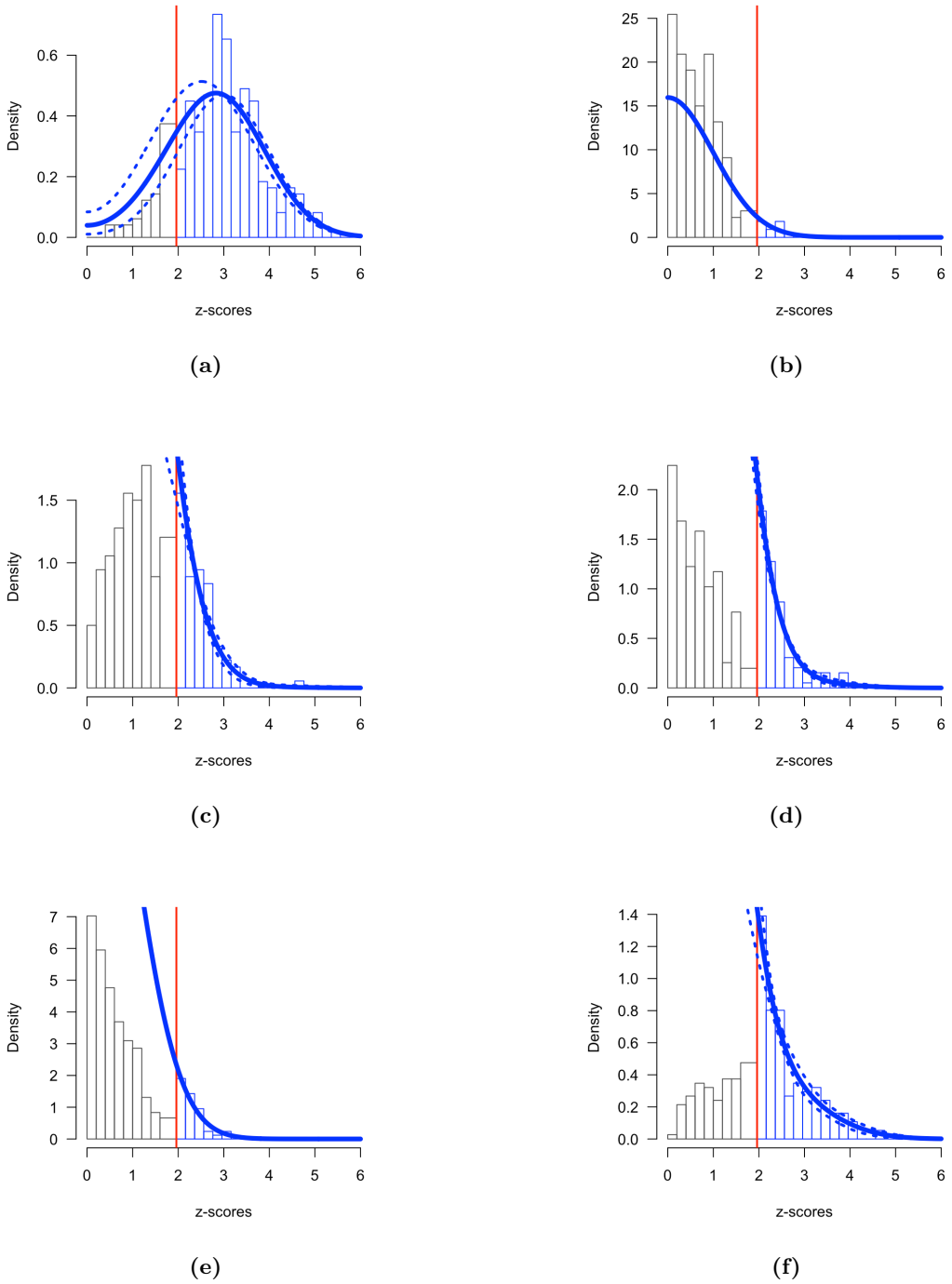

Fig. 2 Distribution of 300 z-scores over the interval 0-6 across six scenarios: (a) ~80% power; (b) no true effect; (c) publication bias and ~20% power; (d) no true effect with publication bias; (e) mild optional stopping with no true effect; (f) mixture of studies designed with high and low power with publication bias and optional stopping. The red line marks the z-score of 1.96, the critical value for statistical significance ($\alpha = 0.05$). The blue line shows the expected density distribution of z-scores, with dotted lines representing its 95% CI.

experimental factors, or even a bias toward non-replication in replication laboratories.

The present study aims to contribute to the Sports Science Replication Project by providing empirical support for the hypothesis that low replication rates are at least partly driven by selection bias and low power. Specifically, we assess the presence of selection bias and estimate average power in a sample of 269 studies published across ten applied sports and exercise science journals using a z-curve analysis of primary statistical results.

Methods

This is a retrospective observational study. The preregistration of this study can be found at https://doi.org/10.17605/OSF.IO/SFBVA.

Study sample size

A sample of 350 studies was used for the purpose of this study. As stated in the preregistration, this sample size was based on a precision analysis conducted for a previous study [17], which aimed to estimate the prevalence, reporting practices and reproducibility of a priori power analyses in sports and exercise science journals. Specifically, the precision analysis was conducted to estimate the number of studies required to detect an expected proportion of studies reporting an a priori power analysis (https://doi.org/10.17605/OSF.IO/MQBR2). Assuming an expected proportion of 30% with a margin of error of 5%, the analysis indicated a required sample of 323 studies, which was subsequently rounded up to 350 studies. For convenience, the same set of 350 studies was used in both the previous study [17] and the present study (https://doi.org/10.17605/OSF.IO/SFBVA.).

Journal and study selection protocol

The 350 studies were sampled from 10 journals ranked in quartile 1 according to www.scimagojr.com (as of 13th September 2022). The list of journals, along with the number of studies sampled from each, is depicted in Figure 3. All studies were published between 2024 and 2018. We started at a given issue and worked backwards. The study selection protocol was based on the Proposal of a Selection Protocol for Replication of Studies in Sports and Exercise Science [18]. First, only applied sport and exercise science studies (studying changes in human performance in response to physical activity, exercise, and sport) in the sub-disciplines of physiology, sports performance, physical activity, injury prevention, and psychology were considered. Second, only confirmatory studies that tested a hypothesis with an experimental (randomized controlled trials) or quasi-experimental design (non-randomized controlled trials) were included. Third, studies had to use an F-test (i.e., ANOVA) or t-test as an inferential test to evaluate the hypothesis; studies that employed correlations, mixed models or Bayesian statistics were excluded. We followed the protocol described in Murphy et al. [18] to select studies, with two deviations from the original protocol. First, whereas the original protocol only selected studies that reported a statistically significant main effect, we considered both statistically significant and non-significant effects. This is because the z-curve analysis uses both significant and non-significant p-values. Second, we also considered interaction effects, which were not considered in the original selection protocol.

Inter-rater reliability

Prior to collecting any data, and in anticipation of difficulties to select the statistical result central to the tested hypothesis, we developed a coding strategy over a four-step process. First, the four authors (CM, JM, DL and JW) developed and discussed the coding form created by CM. Ambiguities in the coding form were discussed, and amendments were made. Second, three raters (CM, JM and JW) independently coded a randomly selected subset of 28 studies from the sample pool of studies (350) as a pilot study. Subsequently, raters' responses were compared, and any disagreements were used to improve the clarity of the coding form. Third, the same three raters (CM, JM and JW) independently coded a second random subset of 19 studies. Inter-rater agreement was assessed by calculating a pooled Fleiss' Kappa estimate for each coding category across the 47 studies coded in the first two rounds of coding. The mean inter-rater agreement for the categorical responses was 0.61, indicating substantial agreement. The second round of coding was also

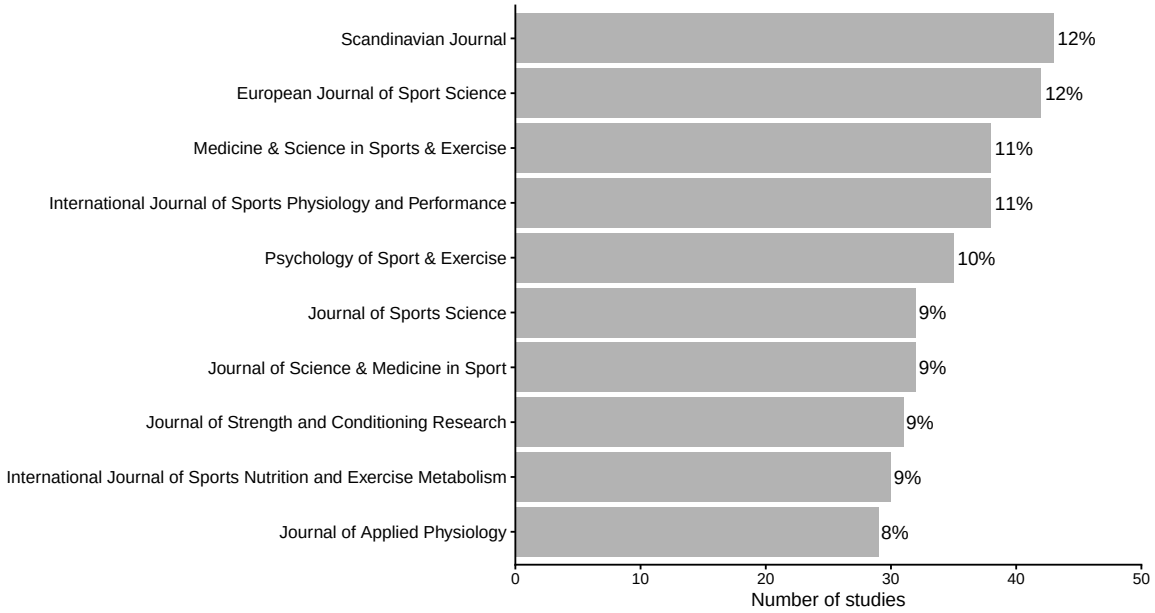

Fig. 3 List of journals from which the 350 independent p-values were extracted, along with the number of studies sampled from each journal and the corresponding proportion.

used to discuss disagreements. The remaining 303 studies were double-coded whereby both JM and JW each coded ~176 studies and CM coded the full sample of studies (350). Inter-rater agreement was assessed by calculating a pooled Cohen's Kappa estimate. The inter-rater agreement for the final round of coding was 0.85, indicating almost perfect agreement. Inter-rater agreements across variables and the coding form can be found at https://doi.org/10.17605/OSF.IO/SFBVA. After termination of data collection, any discrepancies in coding decisions were resolved through discussion between the two pairs of raters and can be found at https://doi.org/10.17605/OSF.IO/SFBVA. DL provided guidance when discrepancies arose and agreement between two raters could not be reached.

Procedures and data extraction

The *z*-curve analysis is an example of a *p*-value meta-analysis and is based on the manually coded *p*-values from the 350 studies. Only one *p*-value per study was extracted, which corresponded to the main statistical test for the central hypothesis of each study. Because hypotheses statements often include vague language and the primary dependent variable is not always operationalised clearly, we used a coding strategy that consisted of several steps to select the key dependent variable. First, the selected dependent variable would be the one for which researchers controlled for both type I and type II error rates. Specifically, in addition to controlling for type I error, researchers conducted an a priori power analysis to control for type II error. Thus, the key dependent variable should be listed in both the a priori power analysis and hypothesis statements. However, on some occasions, the dependent variable stated in the a priori power analysis would not match the dependent variable stated in the hypothesis. In these cases, we would select the dependent variable stated in the hypothesis if it is clearly identifiable. Often, the statistical result central to the hypothesis tested was difficult to identify due to the lack of a priori power analysis and the vagueness of the hypothesis tested. This included hypothesis statements that predicted the effect of one or several interventions on more than one dependent variable or a dependent variable that was measured in multiple ways. In those cases, we selected a dependent variable linked to the central hypothesis test and listed in: 1) the sentence describing the aim of the study; 2) the abstract; 3) the title; 4) or the results, in this order of priority. We selected the dependent variable that best matched the language the authors use to imply the focus of the study, in cases where there were still several dependent variables listed. For each study the following pieces of information were extracted the a priori power analysis statement, the hypothesis statement, whether the

hypothesis predicted the presence or absence of an effect, the type of effect (i.e., a mean difference, a main effect or interaction effect), the statistical result including the degrees of freedom, the test statistic, the effect sizes and its CI, and the p-value. A disclosure table containing all extracted information used to justify the coding decisions regarding the selected key statistical result for each selected study can be found at https://doi.org/10.17605/OSF.IO/SFBVA.

Recomputing p-values

The z-curve method requires exact p-values (e.g., $p = 0.002$) as input. If the corresponding p-value was reported relatively (e.g., $p < 0.05$), we attempted to recompute the p-value when sufficient information was available (i.e., degrees of freedom and F-ratio or t-statistic). P-values were recomputed in Microsoft Excel using the functions T.DIST.2T or F.DIST.RT for t-tests and F-tests, respectively. These functions require both the test statistic and degrees of freedom. In case where a t-statistic or F-ratio from a one-way ANOVA with two levels was reported but the degrees of freedom were not reported, the degrees of freedom were determined using the sample size per group and study design reported in the original study. When the exact p-value and the corresponding statistic were not reported, but an effect size was available, we attempted to convert effect sizes into p-values for study designs involving a t-test and one-way ANOVAs with two levels. Formulas used to recompute p-values from effect sizes can be found in the supplementary information at https://doi.org/10.17605/OSF.IO/SFBVA. We did not attempt to compute other ANOVA effect sizes (i.e., ω^2, ω_p^2) because they require information that is seldom reported in articles, such as mean-square (MS) and sum-of-squares (SS) errors.

Study exclusions

P-values reported as $p < 0.05$ or $p > 0.05$, which could not be recomputed in their exact form, were excluded. There is no optimal decision in how to deal with p-values in studies where results are underreported, and exact p-values cannot be recomputed, which stresses the importance of fully reporting the results of statistical tests. Second, p-values extracted from studies that tested the hypothesis of no effect or equivalence using a classic hypothesis test were not included. Similarly, p-values obtained from studies that tested a directional hypothesis but obtained a significant result in the other direction were not included because they can also distort the results of the z-curve. Finally, studies that used a mixed design but did not directly compare two interventions were excluded, such as claims that one intervention is superior to a control condition after observing a pre-post significant difference in the intervention group, while the corresponding pre-post difference in the control group is not significant. Performing two paired t-tests is statistically invalid because it does not test the hypothesis that researchers set out to test (i.e., one intervention is superior or inferior to the other), which would require a direct comparison between the two groups [19].

Out of the 350 independent p-values extracted, 81 (23%) were excluded. Among those 81 p-values, 46 (57%) could not be recomputed into an exact p-value, 23 (28%) studies tested a hypothesis of no difference without using an equivalence test, 6 (7%) studies reported a significant p-value in the opposite direction as predicted, for 5 (6%) studies the key statistical result was unclear, and 1 (1%) used a within-subject comparison instead of an interaction effect, meaning the result of the test for the hypothesis was not reported. As a result, a total of 269 p-values were converted into z-scores to fit the z-curve model.

Study deviations

In the preregistration, it was stated that studies reporting absolute p-values (e.g., $p > 0.05$) that could not be recomputed into their exact form would not be included. However, p-values reported as $p < 0.001$ or $p < 0.005$ were coded as $p = 0.0001$ and $p = 0.0005$, respectively, and included in the z-curve analysis. This decision represented a deviation from the preregistration. We made this conservative decision because it is common (and defensible) to report results with such small p-values using the 'smaller than' notation, and this reporting strategy is more likely to be observed for studies investigating true effects with high power. Excluding such studies would bias our inclusion criteria towards lower-powered studies, while deviating from our preregistration leads to the inclusion of studies with higher power.

Statistical analysis and software

All simulations and z-curve analyses were performed in R (R version 4.4.2 (2024-10-31); [20]) using the *zcurve 2.0* package [21]. R packages used to produce this manuscript include *readxl* [22], *dplyr* [23], *ggplot2* [24], *knitr* [25] and *purrr* [26]. The manuscript was written in *Quarto* [27]. Data and analysis scripts related to this study are publicly available on the Open Science Framework and can be found at https://doi.org/10.17605/OSF.IO/SFBVA.

Results

Out of all 269 included p-values, 13 were imputed as follows: 11 were reported as $p < 0.001$, 1 as $p < 0.003$ and 1 (1/350) as $p < 0.005$, which were coded as $p = 0.0001$, $p = 0.0003$ and $p = 0.0005$, respectively, and were included in the z-curve model. In addition to the sensitivity analysis excluding these p-values, further sensitivity analyses were conducted at the reviewer's request. These included: (2) a sensitivity analysis in which 13 p-values reported as $p < 0.05$ that could not be recomputed and were not included in the primary z-curve were imputed as 0.25; (3) a sensitivity analysis in which 13 p-values reported as $p > 0.05$ that could not be recomputed and were not included in the primary z-curve were imputed as 0.5; and (4) a sensitivity analysis combining both sets of imputed p-values. All four sensitivity analyses returned results consistent with those of the primary z-curve. The results of the sensitivity analyses can be found at https://doi.org/10.17605/OSF.IO/SFBVA.

The results of the z-curve analysis are shown in Figure 4. The Observed Discovery Rate was 68% (95% CI [62; 74]), indicating that 68% of sampled studies supported the hypothesis tested. The Expected Discovery Rate was 11% (95% [5; 31]), indicating an average power of 11% for studies reporting both significant and non-significant results. The Expected Replication Rate was 37% (95% CI [37; 61]) indicating that studies reporting significant results have an average power of 50%. This suggests that if we were going to conduct direct replications with the sample size of the original studies reporting significant findings, only 50% of these studies would be expected to yield another significant effect.

Selection 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 (68%) is larger than the upper bound of the 95% CI of the Expected Discovery Rate ([5; 31]), suggesting that we can statistically reject the null hypothesis of no selection bias. The point estimate of the Maximum False Discovery Risk was 42% (95% CI [12; 100]), indicating that, in a worst-case scenario, an estimated 42% of the significant effects could be type I errors.

The point estimate of the File-Drawer Ratio was 8 (95% CI [2; 19]), suggesting that for every published significant result, z-curve predicts 8 unpublished studies with non-significant results. Finally, a visual inspection of Figure 4 also indicates that 79 out of 269 (29%) had z-scores greater than 2.8, which indicates the presence of studies investigating true effects with high-power designs ($\geq$ 80%).

Discussion

The first aim of this meta-study was to estimate the average power of studies published across ten journals using a z-curve analysis. The Expected Discovery Rate—reflecting the average power of studies reporting a significant and non-significant effect—was 11% (95% CI [5; 31]), which is substantially lower than the commonly recommended minimum power of 80%. Despite the low average power, 79 out of the 269 (`high_power`%) studies included in the z-curve analysis had z-scores greater than 2.8, suggesting that a subset of studies tested true effects with high-power designs ($\geq$ 80%). In other words, while average power across the literature is low, approximately one quarter of studies seem to have been designed with adequate power, likely due to the investigation of large effects, large sample sizes, or both. Conversely, many studies had extremely low power, in some cases approaching the lower limit of 5%, which corresponds to the nominal type I error rate and reflects the expected probability of a significant result if the true effect size is zero. Low power is not unique to sports and exercise science but a recurrent issue across disciplines [28–30]. Studies designed with low power yielding non-significant effects have low informational value because such findings

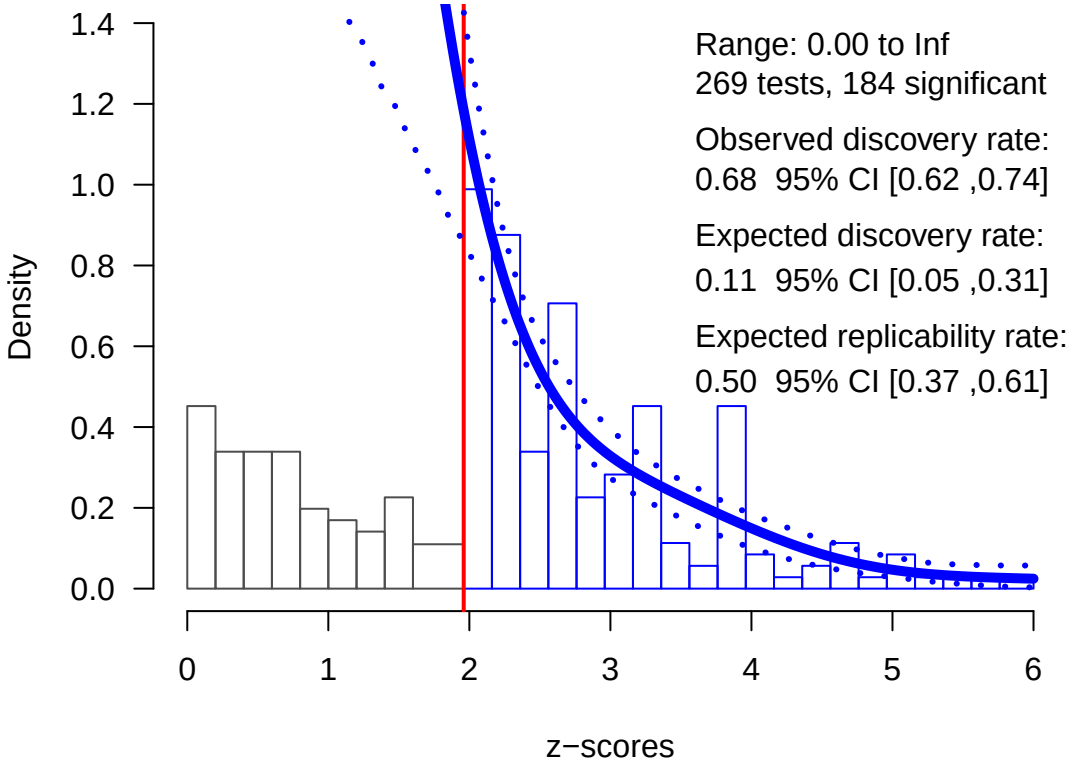

Fig. 4 Distribution of 269 z-scores over the interval 0-6. The vertical red line refers to a z-score of 1.96, the critical value for statistical significance ($\alpha = 0.05$). The dark blue line is the density distribution for the inputted p-values (represented in the histogram as z-scores). The dotted 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.

have a high probability of being a type II error. Moreover, most studies in the field lack sufficient sample sizes to perform an equivalence test with adequate statistical power, preventing researchers from statistically assessing the absence of meaningful effects. This is because the narrower the equivalence bounds, or the smaller the effect sizes one tries to reject, the larger the required sample size [31]. Underpowered study designs also increase uncertainty around effect size estimates, as reflected in the width of the CI. For example, a study with a small sample reporting a 95% CI for a standardised effect size ranging from 0.10 to 0.90 provides little information about the true effect. In contrast, a study with a larger sample that reports a 95% CI ranging from 0.5 to 0.6 provides a far more precise and informative estimate for the scientific literature.

Another consequence of studies with underpowered designs in a literature that selects for significance is a high false discovery risk [28,32]. Researchers typically set α to 0.05 to limit the long-term probability of making a type I error. However, setting α to 0.05 does not ensure that the literature will contain at most 5% of type I errors if there is selection bias in the literature. If researchers select a statistically significant study from a literature that suffers from selection bias and low power, the probability that this study is a type I error will be much higher than 5%. This is indeed what our analysis reveals: an average power of 11% results in a Maximum False Discovery Risk of 42% (95% CI [12; 100]). The point estimate of 42% indicates a high risk of type I errors results where nearly every other significant result is a false positive. However, the range around this estimate is wide. The upper limit reaches 100%, suggesting that all significant results are false positives. While this is unlikely, the results suggest a high risk that significant results are false positives or true effects with negligible effect sizes. Researchers should be aware of the probability that findings in the literature can have a high average probability of being a type I error. To reduce the risk of type I error, it is essential to fully report all results regardless of statistical significance, control the type I error rate for each inferential claim, and design studies with high power.

The second aim of this meta-study was to assess the presence of selection bias. Although our Observed Discovery Rate of 68% is not as high as the 81% estimate previously reported by Büttner et al. [3], there is a discrepancy of 37% between the Observed Discovery Rate (68%) and the upper bound of the Expected Discovery Rate 95% CI [5; 31]. This gap indicates strong evidence of selection bias. The Expected Replication Rate was 50% indicating that only half of the studies that reported a significant effect would replicate. Our observed Expected Replication Rate of 50% (95% CI [37; 61]) is in line with the actually observed 56% replication rate (based on statistical significance of the replication studies) observed in the sports science replication project [9]. Taken together, these findings provide reasonable evidence of inflated type I error rates in the sports and exercise science literature.

It is important to emphasize that the Expected Replication Rate is not a complement of the type I error rate. A 50% Expected Replication Rate does not indicate that 50% of the replications fail because the original findings were type I errors. Rather, the Expected Replication Rate represents the probability of obtaining a significant result in a direct replication using the original sample size. Replication failues may therefore arise either because the original study was a type I error or because the replication study lacks the power to detect the true effect. Although the precise contributions of type I and type II errors to the Expected Replication Rate cannot be disentangled in our sample, insight can be gained by comparing the Expected Replication Failure Rate (1 – Expected Replication Rate) with the Maximum False Discovery Risk to interpret replication failures. The Maximum False Discovery Risk (42%) is close to the Expected Replication Failure Rate (50%) suggesting that in the worst-case scenario almost half of the potential replication failures could be due to type I errors in original studies. While it would be desirable to quantify the exact proportion of false positive in the literature, our study is a stark reminder that without high-powered study designs and an unbiased literature, distinguishing between true and false findings becomes increasingly difficult. Consequently, researchers should remain cautious and consider the possibility that published studies may not replicate.

The concept of the file drawer was introduced by Rosenthal [12] and refers to unpublished studies that produced non-significant results. If studies had 80% power, there would be only one non-significant study in the file drawer for every four published significant studies (File-Drawer Ratio = 1:4 or 0.25:1). However, if studies have only 20% power, there would be four non-significant studies for every published significant study (File-Drawer Ratio = 4:1). Thus, the impact of file-drawer bias increases as study power decreases. The file-drawer problem has important implications for accessing an unbiased literature. Underpowered studies can be combined in a meta-analysis to estimate the true effect size accurately, but this assumes that all effect

sizes are published. If some effect sizes remain unreported, meta-analytic estimates of the true effect size may be biased.

Given the presence of selection bias and an average power of 11% in our sampled studies, statistically significant results in the literature are only possible with inflated effect sizes, where the true, unbiased effect size is actually much smaller—and possibly even zero—. It is therefore not surprising that a common finding among replication projects is that unbiased replication studies with larger sample sizes produce much smaller effect sizes [9,33,34]. For instance, the sport science replication project found that 88% of the original effect sizes were severely inflated in comparison to the replication effect sizes, with a median percentage decrease of 75% [9]. The goal of any empirical science should be the accumulation of reliable knowledge that researchers can build upon to develop new theories, formulate hypotheses, design experiments or conduct meta-analyses [35]. However, our findings in combination with those reported by the sports science replication project suggest that many studies published in our field are upwardly biased, hindering the notion of cumulative science.

To improve the informational value of studies published in the sports and exercise literature, the field should adopt several complementary practices to prevent selection bias and underpowered designs. First, Registered Reports are an effective safeguard against publication bias and p-hacking [36,37]. In this format, the study protocol—including hypotheses, methods, and statistical analyses—is peer reviewed before data collection, and journals offer in-principle acceptance, meaning the study will be published regardless of whether the hypotheses are supported, provided the approved protocol is followed. Despite their utility, only three journals in sport and exercise science currently accept Registered Reports [38–40]. Researchers should also design their studies with high power by conducting rigorous a priori power analyses. Unfortunately, only 41% of studies in our sample performed an a priori power analysis to justify the sample size, and of those, many were poorly conducted [17]. Combined with the systematic use of small samples [10,17,41], this is a serious concern for the field. Journals should require valid sample size justifications [42], researchers should ensure power analyses are conducted correctly, and collaborative research should be considered when individual data collection is challenging. Researchers should avoid overgeneralizing results from underpowered studies [43]. Beyond these practices, the field should increasingly adopt rigorous preregistration of hypothesis-testing studies when researchers choose not to use Registered Reports[14], higher standard for open data and code [44], transparent reporting of exploratory research [45], and collaborate with statisticians to ensure adequate study design and statistical analyses [46]. Together, these measures can substantially increase the reproducibility, replicability and therefore informational value of research in sports and exercise science.

We need to highlight three limitations of our study. First, all bias-detection methods rely on simplified models of selection bias and therefore rest on assumptions that will inevitably be violated to some extent. Although z-curve provides one of the most informative tools currently available for drawing inferences about selection bias in sports and exercise science, the true state of the field cannot be known in the absence of complete transparency. Consequently, z-curve results should be interpreted as our best available indication of selection bias, while recognizing that no statistical method can perfectly quantify the degree of bias in a scientific literature. Second, even though we followed a coding scheme, the raters often had to make subjective decisions when selecting the key statistical result. These difficulties arose because hypotheses were often vaguely stated, mainly as a result of two issues: 1) the effect of interest was often not clearly stated, and 2) the primary outcome was often operationalized using additional measures of the same construct, or measured in multiple ways [47]. These two issues, either in isolation or in combination, result in a multiplicity of hypothesis tests, which makes it difficult to link the tested hypothesis to the statistical result. Third, we included only studies that tested their hypotheses with t-tests or ANOVAs and thus excluded studies that used other types of statistical tests, such as mixed models or Bayesian analyses. We do not know if our results generalize to other designs or analyses. Fourth, 81 (23%) out of the 350 independent p-values were excluded due to poor reporting practices or misuse of hypothesis tests (e.g., testing a hypothesis of no difference with a classic null hypothesis test). This means our findings do not generalize to studies that fail to fully report statistical results.

Conclusion

Overall, our findings indicate substantial selection bias in sports and exercise science. The estimated average power of the sampled studies was 11% (95% CI [5; 31]), and only about a quarter of the studies appear to have been designed with adequate power. The combination of selection bias and low average power is likely to contribute to a literature characterized by inflated effect sizes, a high proportion of type I and II errors, and, consequently, low replicability. Consistent with this interpretation, the z-curve analysis estimates that about half of the published significant findings would fail to replicate in direct replications using the same sample size. Taken together, these results should be a cause for concern for researchers in the discipline. To improve the informativeness and reliability of the evidence base, sport and exercise science must make a collective effort to prioritize high-powered study designs, transparent research practices, and a publication culture that values methodological rigor over statistical significance.

Availability of data and material

Raw data, code and other supplementary materials are available at https://doi.org/10.17605/OSF.IO/D7WYC

Funding

Cristian Mesquida was supported by the Amodo Science Award. Jennifer Murphy was supported by the Irish Research Council's Government of Ireland Postgraduate Scholarship Programme [GOIPG/2020/1155].

Disclosure statement

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

Funding

Cristian Mesquida was supported by the Ammodo Science Award 2023 for Social Sciences. Jennifer Murphy was a recipient of the Irish Research Council's Government of Ireland Postgraduate Scholarship Programme (project ID GOIPG/2020/1155).

CRediT statement

Cristian Mesquida: conceptualization (lead); investigation (lead); methodology (equal); data collection (equal); data curation (lead); formal analysis (lead); writing - original draft preparation (lead); Jennifer Murphy: conceptualization (equal); data collection (equal); Joe Warne: conceptualization (equal); methodology (equal); data collection (equal); supervision (equal); writing and editing (equal); Daniël Lakens: conceptualization (equal); methodology (equal); data curation (equal); supervision (equal); writing and editing (equal).

Acknowledgements

The authors would like to thank Eline Ensinck who contributed to the data curation of this manuscript.

References

1. Scheel AM, Schijen MRMJ, Lakens D. An Excess of Positive Results: Comparing the Standard Psychology Literature With Registered Reports. Advances in Methods and Practices in Psychological Science. 2021;4:1–12. https://doi.org/10.1177/25152459211007467

2. 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. The American Statistician. 1995;49:108–12. https://doi.org/10.2307/2684823

3. 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. British Journal of Sports Medicine. 2020;54:1365–71. https://doi.org/10.1136/bjsports-2019-101863

4. Mesquida C, Murphy J, Lakens D, Warne J. Publication bias, statistical power and reporting practices in the Journal of Sports Sciences: Potential barriers to replicability. Journal of Sports Sciences. 2023;41:1507–17. https://doi.org/10.1080/02640414.2023.2269357

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. Communications in Kinesiology. 2021;1:1–17. https://doi.org/10.51224/cik.v1i3.43

6. Brunner J, Schimmack U. Estimating Population Mean Power Under Conditions of Heterogeneity and Selection for Significance. Meta-Psychology. 2020;4. https://doi.org/10.15626/MP.2018.874

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

8. Wilson BM, Wixted JT. The Prior Odds of Testing a True Effect in Cognitive and Social Psychology. Advances in Methods and Practices in Psychological Science. SAGE Publications Inc; 2018;1:186–97. https://doi.org/10.1177/2515245918767122

9. Murphy J, Caldwell AR, Mesquida C, Ladell AJM, Encarnación-Martínez A, Tual A, et al. Estimating the Replicability of Sports and Exercise Science Research. Sports Medicine. 2025; https://doi.org/10.1007/s40279-025-02201-w

10. 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. Journal of Sports Sciences. Routledge; 2020;38:1933–5. https://doi.org/10.1080/02640414.2020.1776002

11. Mahoney MJ. Publication prejudices: An experimental study of confirmatory bias in the peer review system. Cognitive Therapy and Research. 1977;1:161–75. https://doi.org/10.1007/BF01173636

12. Rosenthal R. The file drawer problem and tolerance for null results. Psychological Bulletin. 1979;83:638–41. https://doi.org/10.1037/0033-2909.86.3.638

13. Stefan AM, Schönbrodt FD. Big little lies: A compendium and simulation of p-hacking strategies. Royal Society Open Science. Royal Society; 2023;10:220346. https://doi.org/10.1098/rsos.220346

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

15. Borg DN, Barnett AG, Caldwell AR, White NM, Stewart IB. The bias for statistical significance in sport and exercise medicine. Journal of Science and Medicine in Sport. 2023;26:164–8. https://doi.org/10.1016/j.jsams.2023.03.002

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

17. Mesquida C, Murphy J, Warne J, Lakens D. The prevalence, reporting practices, and methodological quality of a priori power analyses in sports and exercise science research. SportRxiv; 2025. https://doi.org/10.51224/SRXIV.575

18. Murphy J, Mesquida C, Caldwell AR, Earp BD, Warne JP. Proposal of a Selection Protocol for Replication of Studies in Sports and Exercise Science. Sports Medicine. 2022; https://doi.org/10.1007/s40279-022-01749-1

19. Bland JM, Altman DG. Comparisons against baseline within randomised groups are often used and can be highly misleading. Trials. 2011;12:264. https://doi.org/10.1186/1745-6215-12-264

20. R Core Team. R: A language and environment for statistical computing [Internet]. Vienna, Austria: R Foundation for Statistical Computing; 2024. https://www.R-project.org/

21. Bartoš F, Schimmack U. Zcurve: An r package for fitting z-curves [Internet]. 2020. https://CRAN.R-project.org/package=zcurve

22. Wickham H, Bryan J. Readxl: Read excel files [Internet]. 2025. https://CRAN.R-project.org/package=readxl

23. Wickham H, François R, Henry L, Müller K, Vaughan D. Dplyr: A grammar of data manipulation [Internet]. 2023. https://CRAN.R-project.org/package=dplyr

24. Wickham H. ggplot2: Elegant graphics for data analysis [Internet]. Springer-Verlag New York; 2016. https://ggplot2.tidyverse.org

25. Xie Y. Knitr: A comprehensive tool for reproducible research in R. In: Stodden V, Leisch F, Peng RD, editors. Implementing reproducible computational research. Chapman; Hall/CRC; 2014.

26. Wickham H, Henry L. Purrr: Functional programming tools [Internet]. 2025. https://CRAN.R-project.org/package=purrr

27. Allaire J, Dervieux C. Quarto: R interface to 'quarto' markdown publishing system [Internet]. 2025. https://CRAN.R-project.org/package=quarto

28. 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. Nature Reviews Neuroscience. 2013;14:365–76. https://doi.org/10.1038/nrn3475

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

30. Quintana DS. Most oxytocin administration studies are statistically underpowered to reliably detect (or reject) a wide range of effect sizes. Comprehensive Psychoneuroendocrinology. 2020;4:100014. https://doi.org/10.1016/j.cpnec.2020.100014

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

32. Colquhoun D. An investigation of the false discovery rate and the misinterpretation of p-values. Royal Society Open Science. Royal Society; 2014;1:140216. https://doi.org/10.1098/rsos.140216

33. 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.org/10.7554/eLife.71601

34. Open Science Collaboration. Estimating the reproducibility of psychological science. Science. 2015;349:aac4716. https://doi.org/10.1126/science.aac4716

35. Curran PJ. The seemingly quixotic pursuit of a cumulative psychological science: Introduction to the special issue. Psychological Methods. US: American Psychological Association; 2009;14:77–80. https://doi.org/10.1037/a0015972

36. Chambers CD, Tzavella L. The past, present and future of Registered Reports. Nature Human Behaviour. Nature Publishing Group; 2021;1–14. https://doi.org/10.1038/s41562-021-01193-7

37. Nosek BA, Lakens D. Registered reports: A method to increase the credibility of published results. Social Psychology. 2014;45. https://doi.org/10.1027/1864-9335/a000192

38. Abt G, Boreham C, Davison G, Jackson R, Wallace E, Williams AM. Registered reports in the Journal of Sports Sciences. Routledge; 2021;39:1789–90. https://doi.org/10.1080/02640414.2021.1950974

39. Impellizzeri FM, McCall A, Meyer T. Registered reports coming soon: Our contribution to better science in football research. Science and Medicine in Football. 2019;3:87–8. https://doi.org/10.1080/24733938.2019.1603659

40. Rasmussen P, Tipton MJ, Stewart A, Bailey DM. Advancing physiology through transparency: Celebrating our first registered report. Experimental Physiology [Internet]. 2025;110:351–4. https://doi.org/10.1113/EP091963

41. Mesquida C, Murphy J, Lakens D, Warne J. Replication concerns in sports and exercise science: A narrative review of selected methodological issues in the field. Royal Society Open Science. Royal Society; 2022;9:220946. https://doi.org/10.1098/rsos.220946

42. Lakens D. Sample Size Justification. Collabra: Psychology. 2022;8:33267. https://doi.org/10.1525/collabra.33267

43. Impellizzeri FM, Murphy J, Mesquida C, Warne J, Hecksteden A, Batomen B, et al. Introducing a new "preliminary report" submission category for small-sample intervention studies: Rationale and instructions. Science and Medicine in Football [Internet]. 2025;0:1–11. https://doi.org/10.1080/24733938.2025.2580319

44. Borg DN, Bon JJ, Sainani KL, Baguley BJ, Tierney NJ, Drovandi C. Comment on: 'Moving sport and exercise science forward: A call for the adoption of more transparent research practices'. Sports Medicine [Internet]. 2020;50:1551–3. https://doi.org/10.1007/s40279-020-01298-5

45. Ditroilo M, Mesquida C, Abt G, Lakens D. Exploratory Research in Sport and Exercise Science. https://doi.org/10.51224/SRXIV.457

46. Sainani KL, Borg DN, Caldwell AR, Butson ML, Tenan MS, Vickers AJ, et al. Call to increase statistical collaboration in sports science, sport and exercise medicine and sports physiotherapy. British Journal of Sports Medicine [Internet]. 2021;55:118–22. https://doi.org/10.1136/bjsports-2020-102607

47. 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. Frontiers in Psychology. 2016;7:1832. https://doi.org/10.3389/fpsyg.2016.01832