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

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

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

ORCIDs

Cristian Mesquida – 0000-0002-1542-8355
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

Data availability

The preregistration, data, and analysis scripts related to this study are publicly available on the Open Science Framework and can be found at https://osf.io/d7wyc/.

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: data conceptualization (equal); data collection (equal); Joe Warne: conceptualization (equal); methodology (equal); supervision (equal); data collection (equal); writing review and editing (equal); Daniël Lakens: conceptualization (equal); methodology (equal); data curation (equal); supervision (equal); writing review and editing (equal).

Acknowledgements

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

Abstract

The sports science replication project has raised concerns about the replicability of published research. Low replication rates can have several causes. One possible cause is an excess of significant results caused by publication bias, where selection for significance inflates the proportion of significant findings in the literature, while the statistical power to detect effects is substantially lower. To date, no study has systematically assessed the average statistical power of research in the field. One method to assess publication 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 (0.68) is larger than the upper bound of the 95% confidence intervals (CI) of the Expected Discovery Rate of [0.05; 0.33] indicating strong publication bias in the literature. The average statistical power is 11% [0.05; 0.33], and only 29% of studies are estimated to have been designed with high power. The Expected Replication Rate was 0.49 95% CI [0.36; 0.61], indicating that only 49% of direct replications with the same sample size should be expected to replicate. Publication 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.

1. 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 nonsignificant results (Scheel et al., 2021; Sterling et al., 1995). 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 (Büttner et al., 2020; Mesquida et al., 2023; Twomey et al., 2021). 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 (Brunner & Schimmack, 2020; Scheel et al., 2021). 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. 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 publication 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 (Scheel et al., 2021): $ODR = \alpha \times (1 - t) + (1 - \beta) \times t$, 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. 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.

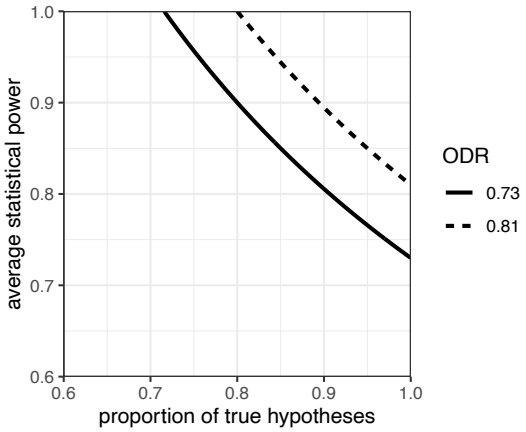

Figure 1. Combinations of the proportion of true hypotheses (x-axis) and power (y-axis) required to produce 73% or 81% statistically significant findings assuming α = 5% and no bias. Notably, achieving 73% significant findings requires the average power to be at least 73% if all tested hypotheses are true.

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 (Szucs & Ioannidis, 2017; Wilson & Wixted, 2018), or with the results of sports science replication project (Murphy et al., 2024). 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 (Abt et al., 2020; Mesquida et al., 2022, 2023), it seems doubtful that the average power could as high as 73%. However, high power could be achieved if studies investigate large effects, or employ within-subject designs. 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 publication 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" (Mahoney, 1977; Rosenthal, 1979) and "*p*-hacking" (Bakker et al., 2012; Stefan & Schönbrodt, 2023). 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 failed to support it (e.g., significance bias). In the context of null hypothesis significance testing, a study reporting a *p*-value below α 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. *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 (see Lakens et al., 2024 for an overview), 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 1 errors. An excess of significant findings has been reported in sports therapy and rehabilitation (Borg et al., 2023) and in the *Journal of Sports Sciences* (Mesquida et al., 2023). 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 publication bias and identifies underpowered designs from the distribution of *p*-values is the z-curve method (Bartoš & Schimmack, 2020). 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 parameters that provide insights into the replicability of a literature, under specific assumptions (Brunner & Schimmack, 2020). 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 subdisciplines 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 a for multiple comparisons or used a more conservative criterion (e.g., $\alpha = 0.01$), bias is modelled based on an alpha level of 0.05 instead of the actual more conservative criterion. This is a conservative assumption, as z-curve will overestimate replicability when the assumed alpha level is higher than the actual alpha level. The 4 parameters computed in a z-curve analysis are described in **Table 1**.

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

Parameter	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 also 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 three studies with 20%, 50% and 90% power each, the EDR would be 50% ((10 + 50 + 90)/3). The EDR can be compared to the ODR to determine the presence of publication 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 publication bias (Bartoš & Schimmack, 2022).

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.
Maximum False Discovery Risk (MFDR)	Th 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 rate type I errors 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 a, and it is computed as MFDR = $\frac{1-EDR}{EDR}$ x $\frac{\alpha}{1-\alpha}$. 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.

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 five distinct scenarios, and then conduct the corresponding z-curve analysis. The rationale behind these scenarios is to provide readers with a diverse set of conditions, illustrating how the distribution of z-scores is influenced by power and selection bias. We simulate 300 *p*-values because this closely represents the number of *p*-values we will report in our study (i.e., $N = 269$ after exclusions). In the first scenario ("80% power"), *p*-values are simulated based on a true effect size (Cohen's d_s) of 0.3 and a total sample size (N) of 278, which yields a power of ~ 80%. In the second scenario ("Null effect"), *p*-values are simulated based on a true effect size of 0, and therefore the number of significant findings corresponds to α. That is, ~ 5% of studies report a significant effect, but all these findings are type I errors, as there is no real effect to be found. In the third scenario ("Publication bias and 18% power"), *p*-values are simulated based on a true effect size of 0.3 and $N = 30$, which yields a power of ~ 20%. Additionally, publication bias is introduced, such that 40% of the non-significant findings remain unpublished. In the fourth scenario ("Mild optional stopping"), *p*-values are simulated based on a true effect size of 0 and a mild optional stopping strategy where researchers perform a maximum of 5 hypothesis tests. When researchers engage in optional stopping, they repeatedly perform a hypothesis test after adding new participants, until either the maximum sample size that researchers are willing to recruit is achieved (in the scenario $N = 50$), or a significant *p*-value is observed, without correcting the alpha level for multiple comparisons. In the last scenario ("Mixed"), the *p*-values are simulated using a number of possible outcomes as follows; 300 *p*-values are first simulated, of which 100 are based on a true effect size of 0.3 and N of 278, 100 are based on a true effect size of 0.3 and $N = 100$, and 100 are based on a true effect size of 0.3 and N of 26. 100 of the non-significant *p*-values have been then randomly replaced by *p*-values obtained through severe optional stopping. The distributions of z-scores under each scenario are presented in Figure 2 and the corresponding results of each z-curve are presented in **Table 2**.

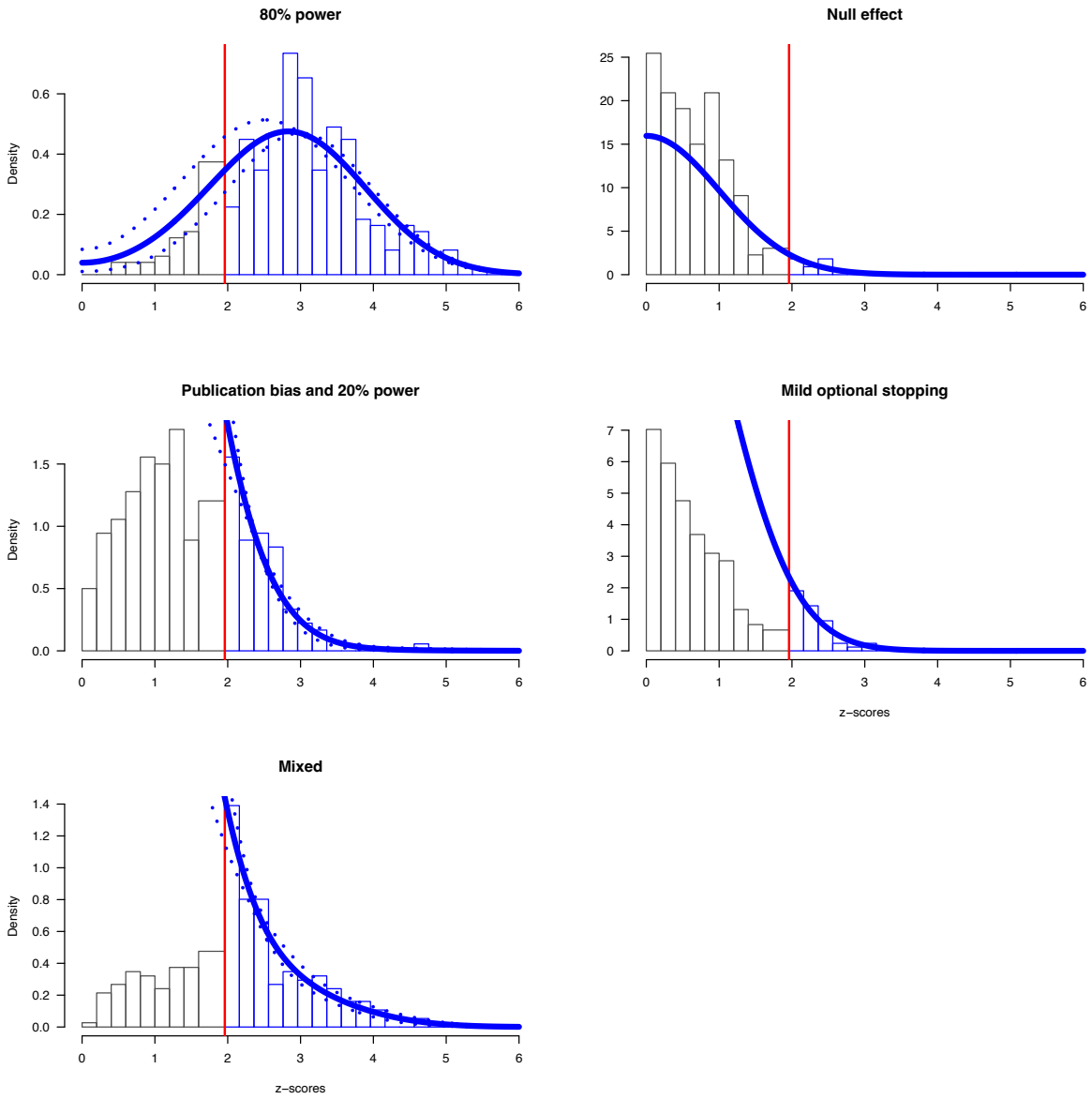

Figure 2. Distribution of 300 z-scores over the interval 0-6. For all five scenarios we simulated 300 *p*-values using an unpaired one-tailed *t*-test and set α to 0.05. The vertical red line refers to a z-score of 1.96, the critical value for statistical significance when using an a of 0.05 in a two-sided test. The solid blue line is the expected density distribution for the observed *p*-values (represented in the histogram as z-scores). The dotted lines represent the 95% CI for the density distribution. The code for the simulations and the z-curve analyses can be found at https://osf.io/d7wyc/.

Table 2. Parameter estimates [95% CI] of the z-curves conducted under five scenarios.

	Observed Discovery Rate	Expected Discovery Rate	Expected Replication Rate	Maximum False Discovery Rate
80% power	0.82 [0.77; 0.86]	0.78 [0.64; 0.91]	0.8 [0.70; 0.88]	0.02 [0.01 0.03]
null effect	0.06 [0.03; 0.09]	0.05 [0.05; 0.10]	0.03 [0.03; 0.06]	1 [0.47; 1]

publication bias and 20% power	0.30 [0.25.; 0.36]	0.08 [0.05; 0.22]	0.10 [0.03; 0.20]	0.65 [0.19; 1]
mild optional stopping	0.14 [0.10; 0.19]	0.05 [0.05; 0.10]	0.03 [0.03; 0.06]	1 [0.47; 1]
mixed	0.62 [0.57; 0.68]	0.09 [0.05; 0.20]	0.35 [0.23; 0.46]	0.5 [0.21; 1]

The results of these simulations illustrate that 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 the alpha level 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, because all significant findings are indeed type I errors. In the presence of publication bias and studies with 20% power we can see that publication bias inflates the Observed Discovery Rate and yields an estimate that is larger than the upper limit of the Expected Discovery Rate 95% CI. This indicates there is 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. In the fourth situation, where optional stopping was simulated under the null distribution, we can see that Observed Discovery is inflated beyond the alpha level. Note that the Observed Discovery Rate is higher than the upper bound of the Expected Discovery Rate 95% CI, again indicating bias. It is important to note 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. Furthermore, 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 is higher than the upper bound of the Expected Discovery Rate 95% CI, once again indicating the presence of bias, but also highlighting how z-curve analysis is not able to distinguish 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. On the contrary, if 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 even more simply, the blue solid line shows the expected distribution of *p*-values in all simulations. In the cases where there is a questionable absence of observed *p*-values below this line, in particular to the left of the red line of z =1.96 (representing non-significant effects), the model would indicate evidence of bias.

To date, there is no study that has complemented the percentage of significant studies with the average power in the same sample of studies in sports and exercise science, which is required to interpret whether there is an excess of significant findings. Furthermore, although the recently reported replication rate of the sports science replication project (Murphy et al., 2024) should provide empirical evidence of the low replicability of the field, skeptical sport and exercise scientists may argue that such low replication rates are not representative of the field due to the small number of replications conducted. Alternatively, they might attribute failures to replicate original studies to deviations from the original studies, replication studies with underpowered designs, unaccounted experimental factors, or even to a bias towards non-replication by replication labs. The study aims to contribute to the sports science replication project by providing empirical support for the idea that low replication rates are at least in part caused by publication bias. Specifically, we assess the presence of publication bias and average power in a sample of 350 studies published across 10 quartile 1 applied sports and exercise science journals using a z-curve analysis of primary statistical results.

2. Methods

This is a retrospective observational study. The preregistration of this study can be found at https://osf.io/d7wyc/.

2.1. 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 to estimate an expected proportion of 30% of studies reporting an a priori power analysis (https://osf.io/mqbr2/). To estimate a proportion of 30% with a margin of error of 5%, the precision analysis returned a sample of 323 studies, which was rounded up to 350 studies. These 350 studies were also used to conduct the z-curve analysis.

2.2. 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 can be found at https://osf.io/d7wyc/. The study selection protocol was based on the *Proposal of a Selection Protocol for Replication of Studies in Sports and Exercise Science* (Murphy et al., 2022). Only applied sport and exercise science studies (studying changes in human performance in response to physical activity, exercise, and sport) in the subdisciplines of physiology, sports performance, physical activity, injury prevention and psychology were considered. Moreover, only confirmatory studies that tested a hypothesis with an experimental or quasi-experimental design were included. We followed the protocol described in Murphy et al., (2022) to select studies, but we deviated on two points from the original protocol. First, while the *Proposal of a Selection Protocol for Replication of Studies in Sports and Exercise* only selected studies that reported a statistically significant main effect, we considered both statistically significant and nonsignificant effects. Second, we also considered interaction effects as opposed to the original selection protocol.

2.3. 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 coding strategy over a 4-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 coding form. Third, the same three raters (CM, JM and JW) independently coded a second random subset of 19 studies. Interrater 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 interrater agreement for the categorical responses was 0.61, indicating substantial agreement. The second round of coding was also 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). Interrater agreement was assessed by calculating a pooled Cohen's Kappa estimate. The interrater agreement for the final round of coding was 0.84, indicating almost perfect agreement. Interrater agreements across variables and the coding form can be found at https://osf.io/d7wyc/. 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://osf.io/d7wyc/. DL provided guidance when discrepancies arose and agreement between two raters could not be reached.

2.4. 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 operationalized 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 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 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) 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 confidence interval (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://osf.io/d7wyc/.

2.5. Recomputing *p*-values

The z-curve method requires exact *p*-values (e.g., $p = 0.002$) as the input parameter. 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 was not reported, 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://osf.io/d7wyc/. 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.

2.6. 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 that 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 (Bland & Altman, 2011). Out of the 350 independent p-values extracted, 57% (46/81) could not be recomputed into an exact *p*-value, 28% (23/81) studies tested the hypothesis of no difference, 7% (6/81) studies reported a significant *p*-value in the opposite direction as predicted, for 6% (5/81) of studies key statistical result was unclear and 1% (1/81) used a within-subject comparison instead of an interaction. Therefore, a total of 81 p-values were excluded, and a total of 269 p-values were converted into z-scores to fit the z-curve model.

2.7. 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. A sensitivity analysis excluding these *p*-values can be found at [https://osf.io/d7wyc/]. The results of the sensitivity analysis are similar to the results of primary z-curve analysis.

3. Results

Out of all 269 included *p*-values, 3% (11/350) were reported as $p < 0.001$, 0.35% (1/350) as $p < 0.003$ and 0.35% (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. As a sensitivity analyses, z-curve analysis excluding the p-values coded as $p = 0.0001$ and $p = 0.0005$ can be found at [https://osf.io/d7wyc/]. The z-curve analysis was performed with the z-curve 2.0 package in R (R Core Team, 2019). The results of the z-curve analysis are shown in **Figure 3**. The Observed Discovery Rate was 0.68 95% CI [0.62; 0.74] indicating that 68% of sampled studies supported the hypothesis tested. The Expected Discovery Rate was 0.11 [0.05; 0.33] indicating an average power of 11% for studies reporting both significant and non-significant results. The Expected Replication Rate was 0.49 95% CI [0.36; 0.61] indicating that studies reporting significant results have an average power of 49%. This suggests that if we were going to conduct direct replications with the sample size of the original studies reporting significant results, only 49% of these studies would be expected to 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 (0.68) is larger than the upper bound of the 95% CI of the Expected Discovery Rate of [0.05; 0.33] suggesting that we can statistically reject the null hypothesis that there is no publication bias. The point estimate of the Maximum False Discovery Risk was 0.43 95% CI [0.11; 1] indicating that, in a worst-case scenario, an estimated 43% of the significant effects could be type I errors. Finally, a visual inspection of Figure 3 also indicates that there is a high number of studies (79 out of 269) with z-scores greater than 2.8 that indicates the presence of studies investigating true effects with high-power designs ($\geq$ 80%).

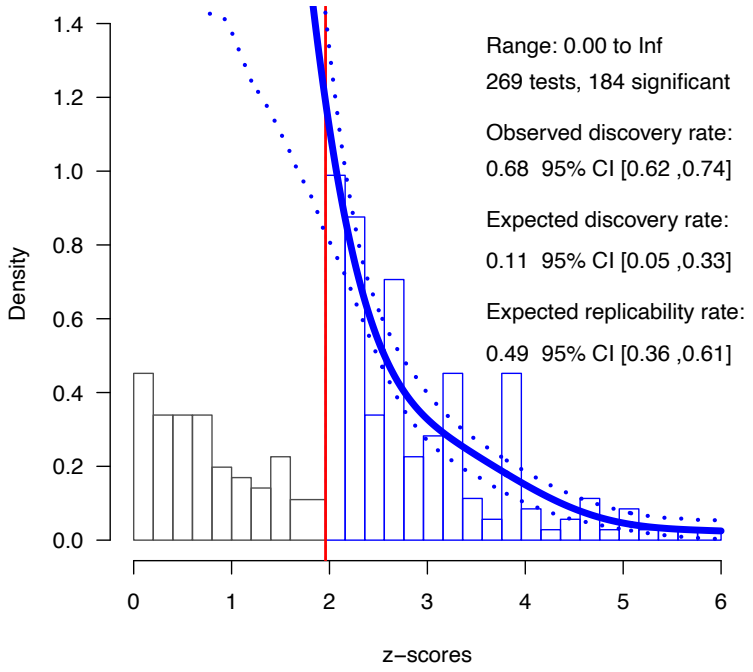

Figure 3. Distribution of 269 z-scores over the interval 0-6. The vertical red line refers to a z-score of 1.96, the critical value for statistical significance when using a two-tailed *p*-value of 0.05. The dark blue line is the density distribution for the inputted *p*-values (represented in the histogram 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.

4. Discussion

The first aim of this meta-study was to estimate the average power of studies published across ten journals by conducting a z-curve analysis. The Expected Discovery Rate – the average power of studies reporting a significant and non-significant effect– was only 11% 95% CI [0.05; 0.33] which is much lower than the minimum recommended level of power of 80%. Despite the low average power, 79 out of the 269 (29%) studies included in the z-curve analysis had z-scores greater than 2.8, suggesting that some studies tested true effects with high-power designs (≥ 80%). In other words, while average power is low, approximately one quarter of studies seem to have been designed with adequate power, likely due to examining large effects, using large sample sizes, or both. Conversely, many studies had extremely low power, in some cases approaching the lower limit of 5%–the type I error rate, which is 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 (Button et al., 2013; Maxwell, 2004; Quintana, 2020). Studies designed with low power yielding non-significant effects have low informational value because such findings have a high probability of being a type II error. A more widespread adoption of equivalence tests (Lakens, 2017) to statistically test for the absence of meaningful effects could highlight the difficulty of interpreting null findings, particularly when sample sizes are small. Furthermore, studies with underpowered designs to detect the effect of interest increase the uncertainty around the true effect size, as reflected in the width of the CI. For instance, a study conducted with a small sample that reports a 95% CI for a

standardized effect size ranging from 0.10 to 0.90 offers little clarity 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 more precise estimate to the scientific literature.

Another consequence of studies with underpowered designs in a literature that selects for significance is a high false discovery risk (Button et al., 2013; Colquhoun, 2014). Researchers set α to 0.05 with the goal of limiting the long-term probability of making a type I error. However, setting α to 0.05 does not ensure that a literature will contain at most 5% of type I errors if there is publication bias in the literature. If researchers select a statistically significant study from a literature that suffers from high publication 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 0.43 95% CI [0.11; 1]. In other words, up to 43% (95% CI [11% to 100%]) of the significant findings in the literature could be a type I error. Researchers should be aware of the probability that findings in the literature can have a high average probability of being a type 1 error, and should aim to reduce this probability by fully reporting all findings regardless of significance, actively designing studies to control for bias, utilizing registered reports, and aiming to design well-powered studies.

The second aim of this meta-study was to assess the presence of publication bias. Although our Observed Discovery Rate of 68% is not as high as the 81% estimate previously reported (Büttner et al., 2020; Twomey et al., 2021), there is a discrepancy of 35% between the Observed Discovery Rate (68%) and the upper bound of the Expected Discovery Rate 95% CI (0.33). This suggests strong publication bias. As explained above, strong publication bias increases the false discovery risk. The Expected Replication Rate was 49% which estimates that only half of the studies that reported a significant effect will replicate. Our observed Expected Replication Rate of 49% 95% CI [0.36; 0.61] 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 (Murphy et al., 2024). Therefore, taken together these findings present reasonable evidence of inflated type I error rates in our literature body. It is important to point out that the Expected Replication Rate is not a complement of the type I error rate. That is, a 50% expected replication rate does not indicate that 50% of the replications would fail because the original findings were type I errors. Recall that the Expected Replication Rate is defined as the probability of obtaining a significant result using the original sample size in a replication study. Thus, a replication study could fail because the original study was a type I error, but also because its study design lacks the power to detect the true effect. While the exact contributions of type I and type II errors to the Expected Replication Rate remain unknown in our sample of studies, we can compare the Expected Replication Failure Rate (1 – Expected Replication Rate) with the Maximum False Discovery Risk to interpret replication failures. The Maximum False Discovery Risk (0.43) is close to the Expected Replication Failure Rate (0.51) suggesting that in the worst-case scenario almost half of the potential replication failures could be due to type I errors in original studies. Although it is desirable to be able to determine how many type I errors are published 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. The best we can do is to urge researchers to consider the possibility that published studies might not replicate, even though the exact probability remains unknown.

Given the presence of publication 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 (Errington et al., 2021; Murphy et al., 2024; Open Science Collaboration, 2015). For instance, the sport science replication project found that 88% of the replication effect sizes were severely inflated in comparison to the original effect sizes, with a median percentage decrease of 75% (Murphy et al., 2024). 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 (Curran, 2009). 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 there are two practices that should be widely adopted to prevent publication bias and underpowered studies. First, an effective safeguard against publication bias and *p*-hacking is the adoption of Registered Reports as a publication format (Chambers & Tzavella, 2021; Nosek & Lakens, 2014). Registered Reports prevent publication bias by peer-reviewing the study protocol containing the hypothesis, methods, and statistical analysis before data collection, and offer in-principle acceptance – regardless of whether the hypothesis are supported or not – as long as the protocol is followed. Despite their utility as remedy against publication bias and *p*-hacking only two journals in sport and exercise science accept Registered Reports (Abt et al., 2021; Impellizzeri et al., 2019). Second, researchers should design their studies with high power to detect the effect of interest by conducting an a priori power analysis, and ideally even design a study that has a high probability to detect the absence of a meaningful effect in an equivalence test to guarantee an informative result both when the hypothesis is true, as when it is false. Regrettably, 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 (manuscript in preparation). These findings, along with the systematic use of small samples (Mesquida et al., 2022; Murphy et al., 2024) should be a real concern to all sport and exercise scientists. Journals should require valid sample size justifications (Lakens, 2022), researchers should make sure their power analyses are conducted correctly, and whenever it is difficult to collect sufficient data individually, researchers in our field should consider collaborative research projects. Where feasibility and resource constraints limit sample sizes, researchers should avoid making overly generalizable claims based on studies with underpowered designs.

We need to highlight three limitations of our study. First, 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 (Wicherts et al., 2016). 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. Second, 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. Third, 81 out of the 350 independent *p*-values, which represents 23% of the sampled studies, 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.

5. Conclusion

Overall, our findings indicate that there is substantial publication bias in sports and exercise science. The average power of the sampled studies is 11%, and just one quarter of the studies seem to have been designed with high power. The presence of publication bias in combination with 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 therefore a low replicability rate. The z-curve analysis estimates that about half of the published significant findings might not replicate in a direct replication. The recent replication project in sports science (Murphy et al., 2022) observed a replication rate of 56% (based on statistical significance of the replication studies), despite the fact that these replication studies had a larger sample size than original studies. Together, these findings should be a cause of concern for all researchers in the discipline. Sport and exercise science should make a collective effort to build a more informative and reliable knowledge base.

6. 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

Abt, G., Boreham, C., Davison, G., Jackson, R., Wallace, E., & Williams, A. M. (2021). *Registered reports in the Journal of Sports Sciences*. *39*(16), 1789–1790. https://doi.org/10.1080/02640414.2021.1950974

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

Bartoš, F., & Schimmack, U. (2020). Z-Curve 2.0: Estimating Replication Rates and Discovery Rates. In *PsyArXiv*. https://doi.org/10.31234/osf.io/urgtn

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

Borg, D. N., Barnett, A. G., Caldwell, A. R., White, N. M., & Stewart, I. B. (2023). The bias for statistical significance in sport and exercise medicine. *Journal of Science and Medicine in Sport*, *26*(3), 164–168. https://doi.org/10.1016/j.jsams.2023.03.002

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

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

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

Chambers, C. D., & Tzavella, L. (2021). The past, present and future of Registered Reports. *Nature Human Behaviour*, 1–14. https://doi.org/10.1038/s41562-021-01193-7

Colquhoun, D. (2014). An investigation of the false discovery rate and the misinterpretation of p-values. *Royal Society Open Science*, *1*(3), 140216. https://doi.org/10.1098/rsos.140216

Curran, P. J. (2009). The seemingly quixotic pursuit of a cumulative psychological science: Introduction to the special issue. *Psychological Methods*, *14*(2), 77–80. https://doi.org/10.1037/a0015972

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

Impellizzeri, F. M., McCall, A., & Meyer, T. (2019). Registered reports coming soon: Our contribution to better science in football research. *Science and Medicine in Football*, *3*(2), 87–88. https://doi.org/10.1080/24733938.2019.1603659

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

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

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

Mesquida, C., Murphy, J., Lakens, D., & Warne, J. (2022). *Replication concerns in sports science: A narrative review of selected methodological issues in the field*. SportRxiv. https://doi.org/10.51224/SRXIV.127

Mesquida, C., Murphy, J., Lakens, D., & Warne, J. (2023). Publication bias, statistical power and reporting practices in the Journal of Sports Sciences: Potential barriers to replicability. *Journal of Sports Sciences*, *41*(16), 1507–1517. https://doi.org/10.1080/02640414.2023.2269357

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 (Auckland, N.Z.)*. https://doi.org/10.1007/s40279-022-01749-1

Murphy, J., Warne, J., Mesquida, C., & Caldwell, A. R. (2024). *Sports Science Replication Centre*. OSF. https://doi.org/10.17605/OSF.IO/3VUFG

Nosek, B. A., & Lakens, D. (2014). Registered reports: A method to increase the credibility of published results. *Social Psychology*, *45*(3). https://doi.org/10.1027/1864-9335/a000192

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

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

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

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

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

Sterling, T. D., Rosenbaum, W. L., & Weinkam, J. J. (1995). Publication Decisions Revisited: The Effect of the Outcome of Statistical Tests on the Decision to Publish and Vice Versa. *The American Statistician*, *49*(1), 108–112. https://doi.org/10.2307/2684823

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

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