Assessing the evidential value of mental fatigue and exercise research

Darías Holgado[1,2,] & Cristian Mesquida[3]

[1] Department of Experimental Psychology, Mind, Brain & Behavior Research Center, University of Granada, Spain

[2] Institute of Sport Sciences, University of Lausanne, Quartier, UNIL-Centre, Bâtiment, Synathlon, Lausanne, Switzerland

[3] Centre of Applied Science for Health, Technological University Dublin, Tallaght, Ireland

This is a pre-print version of the article. To cite:

Holgado, D. & Mesquida, C.: Assessing the evidential value of the mental fatigue and exercise research. 2023. doi:

Corresponding author: Darias Holgado. darias.holgado@unil.ch

Abstract

It has been reported repeatedly that mental fatigue can negatively affect exercise performance, but recent findings have questioned the strength of the effect. To further complicate this issue, an overlooked problem might be the presence of publication bias in a body of literature with studies using underpowered designs which is known to inflate false positive report probability and effect size estimates. Altogether, the presence of bias in the literature is likely to reduce the evidential value of the published literature on this topic, although it is unknown to which extent. The purpose of the current work was to assess the evidential value of studies published up to date on the effect of mental fatigue on exercise performance by assessing the presence of publication bias and the observed statistical power achieved by these studies. A traditional meta-analysis revealed a Cohen's d_z effect size of –0.49, 95% CI [–0.63, –0.34], $p < 0.001$. However, when we applied methods for estimating and correcting for publication bias (such as small-study effects), we found that the bias-corrected effect size decreased to –0.17. Furthermore, the median observed statistical power assuming the meta-analytic effect size (i.e., –0.49) as the true effect size was 34% (min = 16%, max = 97%), indicating that on average these studies only had a 34% chance of observing a significant result if the true effect was Cohen's $d_z = –0.49$. If the adjusted effect size (–0.17) was assumed as the true effect, the median statistical power was just 0.09%. We conclude that the evidence for the effect of the mental fatigue effect is a useful case study for illustrating the dangers of small-study effects.

1 Introduction

Mental fatigue has attracted the attention of many sport scientists over the course of the last two decades given the potential negative consequences for exercise performance [1–3]. A key hypothesis in this area is that performing a long demanding mental effort increases perception of mental fatigue and this reduces performance in a subsequent physical exercise task [4]. There is some doubt, however, regarding the strength of the effect of mental fatigue on exercise performance [3,5]. For instance, the only pre-registered study that has attempted to replicate the seminal study by Marcora et al., [6], failed to replicate these findings [5]. After watching a 90-minute documentary or performing a mental fatigue task, 30 participants (in comparison with 16 participants in the original study) completed a time-to-exhaustion task. There was no evidence of reduced performance or increased perceived effort during the cycling task in the mental fatigue condition [5]. Nonetheless, the fact that an original finding cannot be replicated does not mean that it does not exist, since science relies on accumulating evidence [7].

In some cases, replications do not succeed because of inadequate replication methods or because of factors that moderate the results. However, original studies might not be replicated for a few other reasons. First because there is no effect to be found. For instance, despite initial claims about the effect of ego depletion, a multilab preregistered project attempting to replicate the ego-depletion effect revealed that the size of the effect was small with 95% CI encompassing zero (Cohen's d = 0.04, 95% CI [–0.07, 0.15]) [8]. Second, the use of questionable research practices such as *p*-hacking or optional stopping can overestimate the true effect size and lead to a large number of Type 1 errors in the published literature after selecting for statistical significance [9–11]. Third, the presence of publication bias in combination with studies using underpowered designs can also distort the cumulative evidence and consequently body literature [10,12,13]. Publication bias is the research practice of selectively favoring the publication of studies that obtain significant effects. On the other hand, if studies are conducted with underpowered designs, the sampling error can cause large swings in effect size estimates [12,14]. Indeed, studies with underpowered designs will only reach statistical significance if the study happens to yield an overestimated effect size. Indeed, most of the studies in the literature are based on low sample sizes (mean = 15, SD = 9.14; min = 8, max = 63) suggesting that some studies might have underpowered designs to detect a range of hypothetical small and medium effect sizes. For instance, assuming a true Cohen's d effect size of –0.49 and a within-subject design, a study would require a sample size of 35 participants to achieve a statistical power of 80%.

The presence of these biases is problematic for the credibility of research because it reduces the evidential value of published literature leading to overestimated meta-analytical effect sizes [15,16]. For instance, the results of a systematic review and meta-analysis [3] on the effect of mental fatigue on exercise performance seemed to indicate a significant negative effect (d_z = –0.48, with a 95% CI [–0.70, –0.28]), but a bias-sensitive analysis suggested that after adjusting for publication bias, this estimate was significantly smaller (d_z = –0.14, 95% CI [–0.46, 0.16]) [3]. The evidential value of a literature

body is determined by the number of studies examining true and false effects, the power of the studies that examine true effects, the frequency of type I error rates (and how they are inflated by *p*-hacking) and publication bias [17,18]. Given that the likely presence of studies with underpowered designs has been overlooked, there is therefore uncertainty as to whether or not the published literature on this topic has provided reliable estimates of the effect of mental fatigue on exercise performance.

One way to assess the evidential value of a body of literature is by considering the presence of publication bias and studies with underpowered designs. However, meta-analytic effect sizes are often taken at face value without considering the evidential value of the primary studies (i.e., publication bias and the statistical power of primary studies). Therefore, we considered it pertinent to perform further analysis to examine the evidential value of the studies investigating this topic, as has been done in other sport science areas [19]. Furthermore, to date, meta-analyses on the effect of mental fatigue have relied on one [1,20] or two methods [2,3] to assess for publication bias and small-study effects. For instance, Giboin & Wolf [1] and McMorris et al. [20] only used Egger's test and Begg's test, respectively. Likewise, besides Egger 's test, Brown et al. [2] relied on the Fail-Safe method (which is known to be outdated and it should be avoided) and Holgado et al. [3] a 3-Parameter-Selection model. However, simulation studies investigating the accuracy of publication bias tests have shown that factors such as high heterogeneity and studies with small sample sizes can inflate type I error rates, decrease statistical power and overestimate or underestimate the meta-analytic effect size to a higher or smaller degree [21–23]. Indeed, Carter et al. [21] argued that no single meta-analytic method consistently outperformed all the others due to the different assumptions underlying these methods. As a result, researchers have been recommended to rely on several publication bias tests [21,22]. In the present manuscript, we therefore examined the evidential value of primary studies included in previous meta-analyses and recent articles that have been published afterwards by assessing the presence of publication bias using several tests and the observed statistical power for a range of hypothetical effect sizes achieved by these studies. We hypothesized that a) there will be evidence of publication bias and b) most of the published articles will not have adequate power to detect the estimated meta-analytic effect size.

2 Methods

The hypothesis, methodology and analysis plan for this study were pre-registered in Open Science Framework along with the datasets generated and R scripts required to reproduce both the statistical analyses and figures included in this meta-analysis (https://osf.io/5zbyu/).

Literature search

We included studies with within-participants designs from previous meta-analyses investigating the effect of performing a cognitive task before a physical exercise which provided enough information and fulfilled the inclusion criteria [1–3]. Additionally, given that the last available meta-analysis was

published in 2020 [3], studies published afterwards and up to May 2022 were also considered. Thus, we conducted a literature search for new studies through Medline, Scopus and Web of Science in May 2022. We used 4 search terms related to mental fatigue and another 4 terms related to exercise: "*mental fatigue*" OR "*cognitive fatigue*" OR "*mental exertion*" OR "*ego-depletion*" AND "*physical performance*" OR "*exercise*" OR "*muscle fatigue*" OR "*sport*".

Study selection

Studies were selected on the basis of the following inclusion criteria: 1) available in English; 2) within-participant design; 3) participants completed a cognitive task of any duration prior to an exercise; 4) the main outcome was a measure of exercise performance (e.g., time, distance completed, average power/speed or total work done); 5) the study provides necessary descriptive information of the main performance outcome. Studies investigating the effect of mental fatigue on psychomotor or tactical skills were not included.

Data extraction

A table containing data extracted from each study can be found at https://osf.io/5zbyu/. The major two major pieces of information for the current study were study effect size and its associated *p*-value. If participants completed more than two experimental conditions, we only considered the control condition and experimental condition (i.e., mental fatigue) without other factors (e.g., mental fatigue in hypoxia). For each study, the following information was extracted: 1) study design; 2) type of experimental conditions; 3) exercise protocol and type of test; 4) statistical test and level of significance; 5) descriptive statistics (study sample size and mean ± SD) for both the experimental and control condition; and 6) the result of the statistical test (e.g., $t(11) = 7.2$, $p < 0.001$). We contacted authors to request unpublished data under two circumstances. First, when no sufficient statistical information was reported to recompute either the study effect size (i.e., *t*-statistic and sample size) or the *p*-value (i.e., degrees of freedom and *F*-ratio or *t*-statistic). Second, when a study used a factorial design with more than two experimental groups, but not pairwise comparison on mental fatigue condition and control condition was reported. Only one *p*-value per independent study/sample of participant for the main outcome was extracted to meet the independence criteria. The extracted p-value corresponded to the same statistical contrast as the effect size estimate.

Effect size calculation

Because we only included within-participant designs, we decided to use Cohen's d_z as our type of effect size estimate. The advantage of doing so is that d_z scores are computed on the basis of the same information that is used to test for statistical significance in these studies (i.e., a paired-samples *t*-test) and, consequently, the confidence intervals of the effect size are more consistent with the *p*-values reported in the original papers. Second, the computation of d_z does not require the correlation between dependent measures since correlation parameters are seldom reported as part of statistical analysis.

Thus, all study effect sizes were calculated as Cohen's d_z, representing the standardized mean difference between mental fatigue and the control group. Cohen's d_z was calculated directly from the *t*-value and the number of participants using the formula provided by Rosenthal (1991) [24], as follows: $d_z = t / \sqrt{n}$. If a study performed a repeated measures one-way within-participant ANOVA for the effect of condition, the F-ratio was converted into a *t*-statistic as $t = \sqrt{F}$.

P-value recalculation

In the case that the corresponding *p*-value was reported relatively (i.e., $p < 0.05$), the *p*-value was recomputed for z-curve analysis when degrees of freedom and *t*-statistic were reported. In the case where the *t*-test was reported but not the degrees of freedom, degrees of freedom were inferred from the study sample size (N - 1). *P*-values were recomputed in Microsoft Excel for Mac version 16.45 using the functions *T.DIST.2T* or *F.DIST.RT* for *t*-tests and *F*-tests, respectively.

2.6. Statistical Analysis

Meta-analysis

The meta-analysis was performed using the *metafor* R package [25] in R version 3.6.1 (R Core Team, 2019) and relied on a random-effects model to fit the overall effect size to estimate the average reported effect of mental fatigue and to assess heterogeneity in effect sizes. The overall effect size is reported along with 95% confidence and prediction intervals. Heterogeneity across studies was assessed by means of Cochran's *Q* to test whether the true effect size differs between the studies, Thompson's I^2 to assess the proportion of total variability due to between-study heterogeneity and tau-squared (τ^2) as estimate of the variance of the underlying distribution of true effect sizes.

Testing for small-study effects and publication bias

Because previous research has shown that there is no single publication bias method that outperforms all the other methods under each and every assumption tested [21,22,26–28], we used a triangulation approach, also known as sensitivity analysis, where researchers do not rely on only one single publication bias method, but use multiple publication bias methods instead [21,29–31]. To test for publication bias, we relied on two types of methods based either on effect sizes or *p*-values. Effect size methods were Egger's regression test for funnel plot asymmetry (R package *weightr* [32]), a Three-Parameter Selection Model with a one-tailed *p*-value cutpoint of 0.025 (3PSM; R package *weightr* [32]), the Precision-Effect Test - Precision-Effect Estimate with Standard Error (PET-PEESE; JASP [31]), the skewness test (R package *altmeta* [27]) and the limit meta-analysis (R package *metsense* [35]). As a *p*-value method, only z-curve was used (R package *z-curve 2.0* [36]). The z-curve method allows testing for publication bias by considering whether the point estimate of the Observed Discovery Rate lies within the 95% confidence interval (CI) of the Expected Discovery Rate. If the Observed Discovery Rate estimate lies outside the 95% CI of the Expected Discovery Rate is considered evidence of

publication bias [36]. For a detailed description of the limitations and assumptions of these methods, readers are referred to Carter et al. [21], McShane et al. [22], Stanley [26] Bartos and Schimmack [36] and Sladevoka [23]. We report two deviations from the pre-registered analysis. First, the limit-meta was not included in the pre-registered protocol [37]. Limit meta-analysis (LMA) is based on the concept of increasing the precision of the meta-analytic effect size using a random-effects model that accounts for small-study effects [23]. Second, both *p*-curve and *p*-uniform methods were discarded following Carter et al. [21] recommendations. These methods result in the overestimation of the true effect size under moderate-to-large heterogeneity.

Statistical power

Several statistical power estimates were calculated using two different methods based on *p*-values and effect sizes. First, we conducted a z-curve analysis which is based on the concept that the average power of a set of studies can be derived from the distribution of *p*-values (see [36] for technical details). This method converts significant and non-significant *p*-values reported in a literature into two-tailed z-scores, and uses the distribution of z-scores to calculate two estimates of average power using finite mixture modeling: the Expected Discovery Rate (EDR) which is the percentage of studies predicted to be significant based on the average power of published studies and the Expected Replication Rate (ERR) which is the average power of the studies entered, which is also an estimate of the percent of the studies that one would expect to replicate if one performed the studies in exactly the same way as they were done before. *Z*-curve also allows measuring the false discovery risk (i.e., Sõric False Discovery Rate) which is an estimate of the maximum percentage of studies that could be false positive. Second, we used a range of hypothetical effect sizes to estimate statistical power using the R package *metameta* [38]. This package allows researchers to estimate the statistical power of the studies included in the meta-analysis by using a) a range of hypothetical effect sizes, and b) the meta-analytic effect size estimate as the true effect size.

3 Results

A total of 56 effect sizes were selected for eligibility. 12 effects sizes were discarded because either inclusion criteria were not met (n = 8) or descriptive data were not reported and authors did not provide raw data upon request (n = 4)[1]. A total of 44 effect sizes from independent samples were included in the meta-analysis. A disclosure table containing the list of 56 effects selected for eligibility, including those included in meta-analysis and the literature search output can be found at https://osf.io/5zbyu/.

[1]Researchers should ensure that raw data is made publicly available on public repositories, such as the Open Science Framework to facilitate reproducibility and reuse of data. The statement "data will be made available upon request" is outdated and in most cases, it implies that raw data will not be shared [39].

Overall meta-analysis

The results of the random-effects meta-analysis are summarized in **Fig. 1**. Across all studies, the random-effects meta-analysis revealed a statistically significant meta-analytic effect size of $d_z = –0.49$, 95% CI [–0.63, –0.34], $p < 0.001$. The 95% prediction interval for the meta-analytic effect size was [–1.26, 0.28]. The significant Q-statistic ($Q(43) = 115.20$, $p < .0001$) leads us to reject the null hypothesis that all studies share a common effect size. Instead, there is true heterogeneity between studies, suggesting the true effect sizes differ between the studies. The estimated heterogeneity was $\tau^2 = 0.14$. Furthermore, I^2 was 65.97% indicating that about two thirds of total variability is due to between-study heterogeneity.

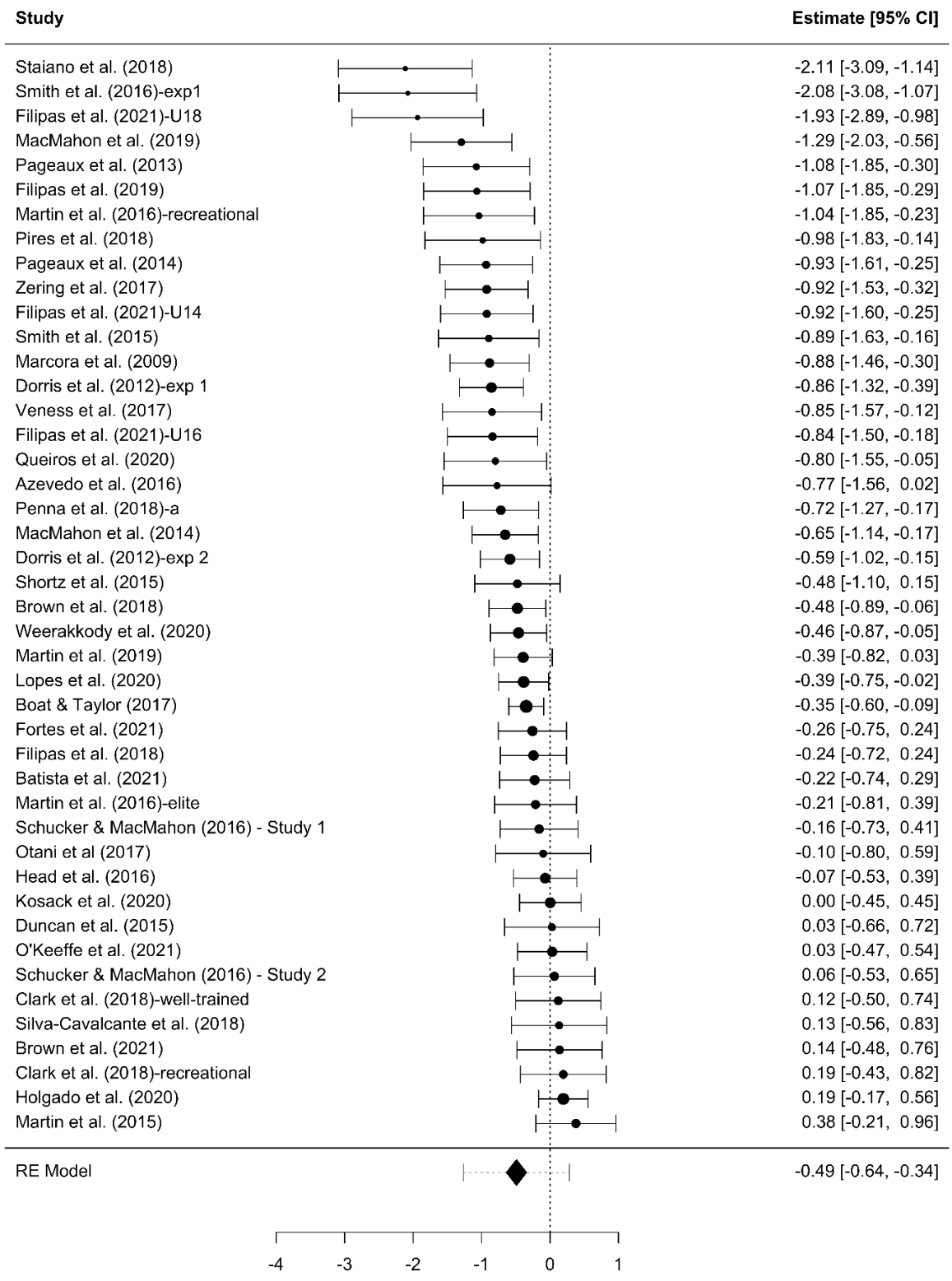

Fig. 1 Forest plot of study Cohen's d_z effect size and meta-analyzed effect size [95% CI].

Small study effect and publication bias

Results of small-study and publication bias tests are summarized in **Table 1**. Egger's regression test for funnel plot asymmetry was b_1 = -3.52, SE = 0.8, z = -4.40, $p < 0.001$, indicating funnel plot asymmetry (**Fig. 2**). Egger's regression test evaluates the null hypothesis of no asymmetry. When Egger's regression test rejects the null hypothesis, it suggests the presence of small-study effects which is usually interpreted as publication bias. In contrast, the skewness test was T_s = 0.46, 97.5% CI [–0.01, 0.92], p = 0.21 and the distribution of study effect sizes was inconclusive. Similarly, the Observed Discovery Rate estimate (0.55) lies within the 95% CI [0.05, 0.66] of the Expected Discovery Rate; this suggests that, even though there is a discrepancy, there is not enough data to statistically reject the hypothesis that there is evidence of publication bias.

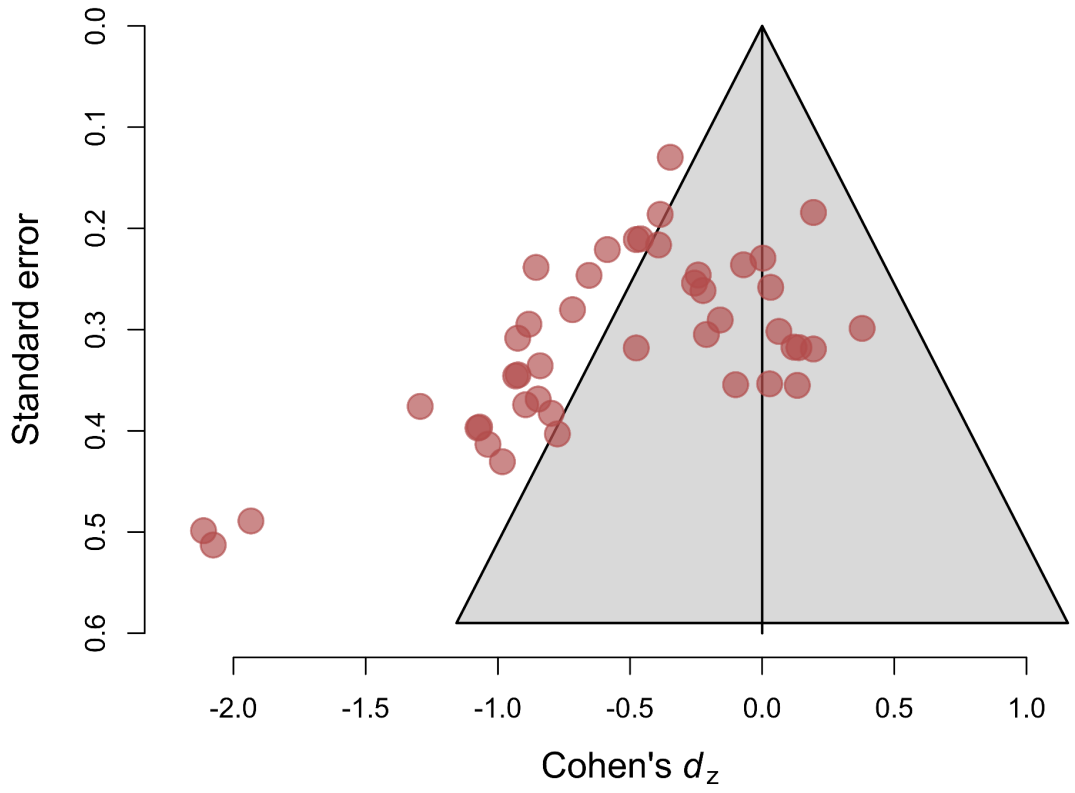

Fig. 2 Funnel plot of study Cohen's d_z effect size versus study standard error.

To obtain an estimate of the likely effect size in the absence of bias, we used several techniques. The fit of the 3PSM was improved significantly assuming the presence of publication bias, $\chi^2(1) = 10.34$, p = .001, and returned an significant adjusted d_z of –0.19, 95% CI [–0.37, –0.007], p = 0.04. PET-PEESE is a two-step procedure whereby only if the null hypothesis is rejected, the second-step PEESE is performed to calculate the adjusted meta-analytic effect size [31]. Only PET procedure was performed because method it returned a non-significant effect ($t(42)$ = 1.21, d_z = 0.24, 95% CI [–0.15, 0.65], p = 0.23) and therefore, the null hypothesis of zero effect could not be rejected based on the 95% CI of the

estimated effect size**.** Likewise, limit meta-analysis (LMA) returned a non-significant adjusted d_z = –0.16, 95% CI [–0.44, 0.1], p = 0.22).

Table 1 Summary results of publication bias tests

Test	Result	Interpretation
Egger's	b_1 = 3.52, SE = 0.8, z = -4.4, p < = 0.001	Small-study effect
skewness	T_s = 0.46, 97.5% CI [–0.01, 0.92], p = 0.21	Inconclusive
z-curve	ODR estimate (0.55) ∈ 95% CI [0.05, 0.66] of the EDR	No publication bias
3PSM	adjusted d_z = –0.19, 95% CI [–0.37, –0.007], p = 0.04	Publication bias
PET	t(42) = 1.21, d_z = 0.24, 95% CI [–0.15, 0.65], p = 0.23	A null effect could not be rejected
LMA	adjusted d_z = –0.16, 95% CI [–0.44, 0.1], p = 0.22	Publication bias

Statistical Power

The *metameta* package [38] was used to calculate statistical power estimates for a range of hypothetical effect sizes. The median statistical power of the studies included in the meta-analysis, using the meta-analytic effect size estimate as the true effect size (d_z = –0.49) was 34% (min = 16%, max = 97%). The statistical power estimates of the studies included in the meta-analysis for a range of hypothetical effect sizes is presented in **Fig. 3**. If we assume the bias-corrected (i.e., the average from 3PSM and LMA) effect size estimate (d_z = –0.17) as a true effect size, the median observed statistical power of the studies would be 0.09% (min = 6%, max = 28%). The z-curve method was also used to estimate average statistical power of all studies included in the meta-analysis (see **Fig. 4**). The Expected Discovery Rate was 0.27 with 95% CI [0.05, 0.66] indicating that the average power of all studies was 27%. The Expected Replication Rate was 0.44 with 95% CI [0.16, 0.70] indicating that the average power of only those studies reporting a statistically significant effect was 44%.

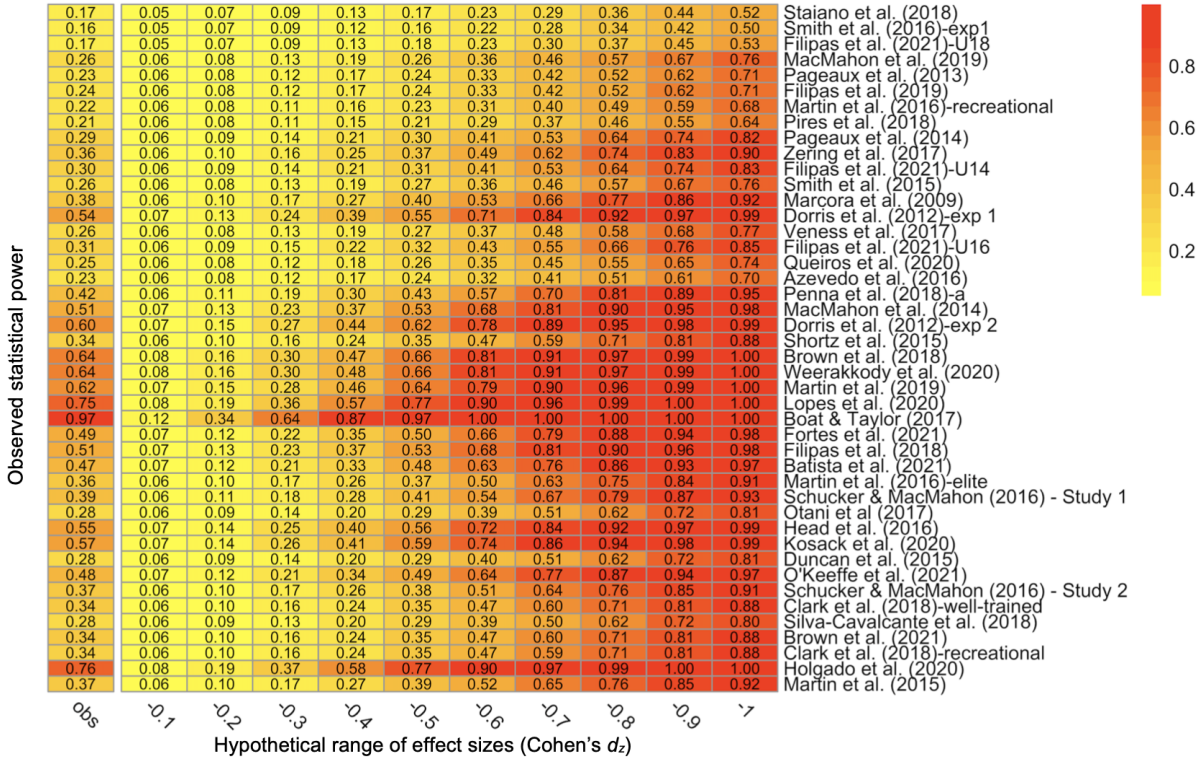

Fig. 3 Observed statistical power estimates for studies included in the meta-analysis assuming a range of hypothetical effect sizes [–0.1, –1]. The leftmost column (*obs)* refers to the observed statistical power assuming the meta-analytic effect size (d_z = –0.49). The rest of the columns represent the observed statistical power of each study given an hypothetical effect size.

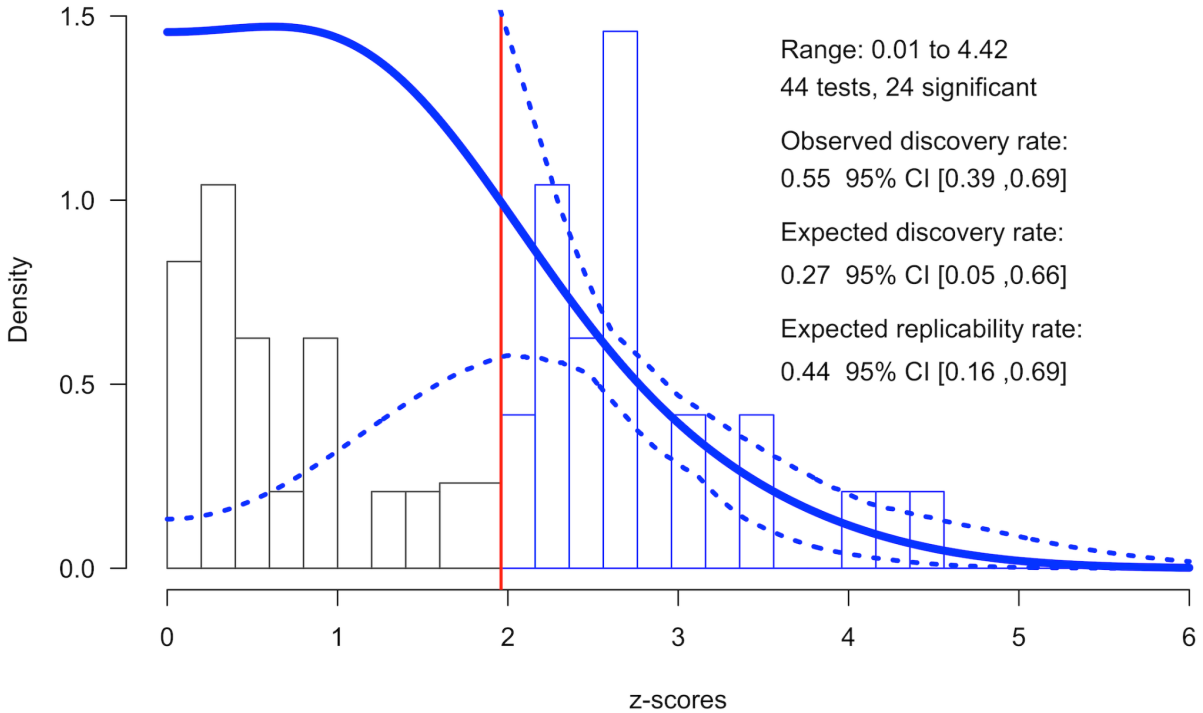

Fig. 4 Distribution of z-scores. The vertical red line refers to a z-score of 1.96, the critical value for statistical significance when using a two-tailed alpha of 0.05. The dark blue line is the density distribution for the inputted *p*-values (represented in the histogram as z-scores). The dotted lines represent the 95% CI for the density distribution. Range represents the minimum and maximum values of z-scores used to fit the z-curve. A total of 44 independent *p*-values (24 significant) were converted into z-scores to fit the z-curve model.

4 Key findings

In the present manuscript, we attempted to examine the evidential value of studies investigating the effect of mental fatigue and we hypothesized that a) there would be evidence of publication bias and b) most of the published articles would not have adequate power to detect the estimated meta-analytic effect size. Although Egger's regression test was significant for small study effect, type I error rates for this test are slightly inflated when the heterogeneity is substantial ($\tau = 4$ or $50\% \leq I^2 \leq 94\%$). Meanwhile, the skewness test and z-curve are known to perform well under heterogeneity. The skewness was inconclusive, and the z-curve could not reject the null hypothesis of no publication bias, respectively. When considering the results from the 3PSM ($d_z = -0.19$, $p = 0.04$) and limit meta-analysis ($d_z = -0.16$, $p = 0.22$), the new meta-analytic effect size after adjusting for publication bias was reduced by at least 0.30 standard deviations. Surprisingly, PET returned a positive effect although based on a rather wide 95% CI [–0.15, 0.65].[2] Due to small samples and high heterogeneity, the 95% CI for the effect size estimates derived from these methods are very wide, ranging from positive to negative estimates and in some instances including 0. Furthermore, another limitation of 3PSM and PET is that they suffer from reduced power when sample sizes are small, heterogeneity is high and there is little publication bias, or there is heavy use of flexibility in data analysis. Similarly, these two methods also tend to underestimate effects when sample sizes are small, heterogeneity is high, or there is heavy use of flexibility in data analysis. On the basis of the results from the 3PSM, and limit-meta-analysis, it seems reasonable to conclude that the negative effect is likely to be much smaller than reported.

Furthermore, the median observed statistical power assuming the meta-analytic effect size ($d_z = -0.49$) as the true effect size was 34% and if the adjusted effect size (–0.17) was assumed as the true effect, the median statistical power was 0.09%. These results are also in line with the results obtained from the z-curve analysis which yielded an observed statistical power of 27% 95% CI [0.05, 0.66] for both significant and non-significant results. Both the shrinkage of the meta-analytic effect size estimate after adjusting for publication bias and the presence of underpowered designs might therefore suggest that the evidential value of the studies included in this meta-analysis is low on average.

4a. The expectancies about mental fatigue should be lower

In this study, we questioned the settled claim that mental fatigue impairs exercise performance. In line with the meta-analysis results (meta-analytic effect size $d_z = -0.49$, 95% CI [–0.63, –0.343]), empirical studies have shown that exerting cognitive effort (regardless of the duration) before exercise might reduce exercise performance [6,40,41]. However, when the overall estimate is adjusted for publication bias, most of the tests provided a lower adjusted estimate of the true effect of mental fatigue on exercise performance, suggesting that the effect might be substantially smaller (see **Table 1**). For instance, the

[2] There are a few studies with very large effects (around $d_z = 2$) that could condition the outcome of this test. However, after removing it, the result is still similar: t(40) = 0.52 dz = 0.10, 95% CI [-0.29, 0.51], p = 0.6

new meta-analytic effect size was reduced by at least 0.30 standard deviations based on the results from the 3PSM (d_z = –0.19, p = 0.04) and limit meta-analysis (d_z = –0.16, p = 0.22). The reasons for publication bias are multiple [42,43], but it varies from editorial predilection for publishing positive findings, researchers' degree of freedom in analyzing the data [44], authors not writing up null results and other causes. The presence of publication bias in a set of published studies is likely to inflate study effect sizes and type I error rate, especially when these studies have underpowered designs. Indeed, the overall negative effect observed in this meta-analysis (Cohen's d_z = –0.49) might be driven by some studies reporting inflated large effects and with high standard error due to study small sample sizes (see **Fig. 2**). Due to the unreliability of the cumulative evidence from experimental studies, there is no certainty of a causal effect, and neither is its absence certain.

4b. Low replicability

The power analysis revealed that even considering d_z = –0.49 as the true effect size, only one study achieved the considered adequate power of 80%, and only two others would be close (see **Fig. 3**). The median power of the literature indicates that if we were to conduct 10 exact replications, only ~3 out 10 studies would find the expected effect. If we assume the adjusted estimate (for any of the publication bias methods), these results would be even more dramatic and all studies would be highly underpowered to detect the adjusted effect. Indeed, a sample size of 274 participants would be required to find a true effect size of –0.17 given an intended power of 80% and a paired t-test. The z-curve analysis adds further support to the above results. The analysis provided that the expected discovery rate (value that is computed from mean power before significance selection [36]) is 0.27 95% CI [0.06, 0.66], which corresponds to the long-run relative frequency of statistically significant results. Therefore, in the future, the sample size should be significantly increased, rather than performing exact replicates. The problem of underpowered studies stems from three main issues.

First, just by the definition of statistical power, if a study has an underpowered design it has a low probability of detecting a significant effect even if there is one to be found (or the null hypothesis is false) [45]. This is reflected in the Observed Discovery Rate for this literature, which was estimated to be 55%– this is, the percentage of articles providing a significant result assuming there is a significant effect to be found. One consideration of this value is that the Observed Discovery Rate does not distinguish between true and false discoveries. Second, studies with low power designs are more likely to produce overestimated effect sizes [46–49]. This will result in a literature filled with exaggerated effect estimates if significant original findings are more likely to be published. As far as we know, only one pre-registered and replication study to date has been conducted and the reported effect size was substantially lower and in the opposite direction of the original study [5,6]. Though replication efforts are still very scarce in the field [50], data from similar disciplines such as psychology show that only half of the original studies were replicated [49,51] and this would be in agreement with the Expected Replicability Rate of 0.44 (see **Fig. 4**). Third, underpowered analyses may lead to a bias in the literature

due to the increased proportion of false positives. [52,53]. Altogether, the presence of studies with underpowered designs hinders the replicability of scientific results and if only studies with significant results were going to replicate, only 44% of them would yield another significant result. Despite these limitations, the results obtained from studies with underpowered designs have usually been taken at face value. In the past, power issues had been overlooked in the evaluation of results and whenever an effect was significant, it was assumed that the study had enough power [54]. The result is that there has been limited incentive to conduct studies with adequate power [55–57]. Meta-analyses may minimize some of the shortcomings of low power studies, but they cannot provide a realistic picture of the literature as a whole from a set of low powered studies [58]. In light of this, the aphorism "Extraordinary claims require extraordinary evidence" may be applicable.

4c. This effect cannot be discarded

Although it might sound cliché, absence of evidence of an effect does not necessarily prove its absence. In line with our results, a major caveat in the literature is effect-size heterogeneity [1–3,59]. Effect-size heterogeneity refers to the variance in true effect sizes underlying the different studies– this is, there is no single true effect size but rather there is a distribution of true effect sizes. Even when the mean distribution of the true effect size is negative, it is likely that some studies yield effect size estimates around zero or even positive. Heterogeneity is not only reflected on the results of *Q-statistic* test and I^2 estimate but also on the 95% prediction interval as its width accounts for the uncertainty of the summary estimate, the estimate of between study standard deviation in the true effect, and the uncertainty in the between study standard deviation estimate itself. Although the prediction interval is below zero [–1.26, 0.28] and thus indicating the effect will be detrimental in most settings, the interval overlaps zero and so in some studies the effect may actually be non-detrimental. Then, as we are unaware of the true effect, the results of future implementations are unclear. This finding is masked when we focus only on the average effect and its confidence interval. However, its width will be also enlarged by bias such as publication bias and studies with underpowered designs, in addition to that caused by genuine effect. Therefore, it is possible that mental fatigue does not affect all types of exercise or that the fitness level of participants mediates its effect. However, the actual presence of these moderators should be interpreted with caution, in a set of studies with low statistical power and publication bias.

Final remarks

Studies conducted so far have not provided reliable evidence that a causal effect exists, but it is also not certain that one does not exist. It is not only this field of research that suffers from publication bias and low statistical power [19,43]. In fact, numerous voices have recently highlighted this problem in sport science literature [19,43,60–63]. However, this should not be used as an excuse to ignore publication bias and low statistical power. Moreover, as we have seen, meta-analyses are not the ultimate tool to solve the problem of low power. Despite the potential for meta-analyses to mitigate some shortcomings of individual studies, results are largely dependent on the quality of the included reports. Moreover,

several publication bias methods should be considered when performing meta-analyses, as each method is built over specific assumptions [21]. An intervention is sometimes considered to be effective or not based solely on its estimated effect size in a meta-analysis, rather than considering the quality of the primary studies and publication bias. Results from a meta-analysis that shows a high selection bias and low replication rate need to be verified independently in experiments with larger samples (ideally in a multi-lab study [64]). Nonetheless, the possible negative effects that mental fatigue could have on human physical performance cannot be ruled out, but the current evidence suggests that perhaps expectations about this effect should be reduced [1–3,59]. Finally, researchers cannot survive as transparent individuals in a system in which the lack of Open Science practices is the norm [65] and we strongly encourage researchers to preregister study protocols, conduct pre-study power calculations for sample size justification and make data and materials publicly available to improve credibility [39,66–68]. Considering the present findings, we encourage caution when making claims or making recommendations on how to counteract mental fatigue's detrimental effects on exercise performance.

Acknowledgement

We thank Daniel S. Quintana, Department of Psychology -University of Oslo and NevSom, Oslo University; James Steele, School of Sport, Health, and Social Sciences -Southampton Solent University; Gerta Rücker, Institute of Medical Biometry and Statistics, Faculty of Medicine and Medical Center - University of Freiburg-; Daniel Sanabria, Mind, Brain & Behavior Research Center - University of Granada, Spain-; Franco M. Impellizzeri, Human Performance Research Centre, Faculty of Health - University of Technology Sydney; Daniel Lakens, Human-Technology Interaction Group, Eindhoven -University of Technology and Miguel A. Vadillo, Department of Basic Psychology - Autonomous University of Madrid-. for their valuable comments that helped improve this manuscript. We also thank all the authors who kindly provided all the raw data or included a link in their manuscript to access it.

Declarations

Funding Cristian Mesquida is funded by Technological University Dublin (project ID PTUD2002) and Darías Holgado is supported by a grant from “Ministerio de Universidades” of Spain and Next Generation Fonds from the European Union.

Conflict of interest Darías Holgado and Cristian Mesquida declare no conflicts of interest.

Ethics not applicable.

Availability of data and material All datasets created are available at https://osf.io/5zbyu/.

Code availability R scripts to analyze data and create figures are available at https://osf.io/5zbyu/.

Author contributions DH and CM equally contributed to the study design, analysis and interpretation of results and writing of the manuscript.

5 References

1. Giboin L-S, Wolff W. 2019 The effect of ego depletion or mental fatigue on subsequent physical endurance performance: A meta-analysis. *Perform. Enhanc. Health* **7**, 100150. (doi:10.1016/j.peh.2019.100150)
2. Brown DMY, Graham JD, Innes KI, Harris S, Flemington A, Bray SR. 2019 Effects of Prior Cognitive Exertion on Physical Performance: A Systematic Review and Meta-analysis. *Sports Med.* (doi:10.1007/s40279-019-01204-8)
3. Holgado D, Sanabria D, Perales JC, Vadillo MA. 2020 Mental fatigue might be not so bad for exercise performance after all: a systematic review and bias-sensitive meta-analysis. *J. Cogn.* **3**, 1–14. (doi:https://doi.org/10.5334/joc.126)
4. Van Cutsem J, Marcora S, De Pauw K, Bailey S, Meeusen R, Roelands B. 2017 The Effects of Mental Fatigue on Physical Performance: A Systematic Review. *Sports Med.* **47**, 1569–1588. (doi:10.1007/s40279-016-0672-0)
5. Holgado D, Troya E, Perales JC, Vadillo M, Sanabria D. 2020 Does mental fatigue impair physical performance? A replication study. *Eur. J. Sport Sci.* (doi:10.1080/17461391.2020.1781265)
6. Marcora S, Staiano W, Manning V. 2009 Mental fatigue impairs physical performance in humans. *J. Appl. Physiol.* **106**, 857–864. (doi:10.1152/japplphysiol.91324.2008)
7. Earp BD, Trafimow D. 2015 Replication, falsification, and the crisis of confidence in social psychology. *Front. Psychol.* **6**, 621. (doi:10.3389/fpsyg.2015.00621)
8. Hagger MS *et al.* 2016 A Multilab Preregistered Replication of the Ego-Depletion Effect. *Perspect. Psychol. Sci. J. Assoc. Psychol. Sci.* **11**, 546–573. (doi:10.1177/1745691616652873)
9. Simmons JP, Nelson LD, Simonsohn U. 2011 False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychol. Sci.* **22**, 1359–1366. (doi:https://doi.org/10.1177/0956797611417632)
10. Bakker M, van Dijk A, Wicherts JM. 2012 The Rules of the Game Called Psychological Science. *Perspect. Psychol. Sci.* **7**, 543–554. (doi:https://doi.org/10.1177/1745691612459060)
11. Stefan A, Schönbrodt F. 2022 Big Little Lies: A Compendium and Simulation of p-Hacking Strategies. (doi:10.31234/osf.io/xy2dk)
12. Button KS, Ioannidis JPA, Mokrysz C, Nosek BA, Flint J, Robinson ESJ, Munafò MR. 2013 Power failure: why small sample size undermines the reliability of neuroscience. *Nat. Rev. Neurosci.* **14**, 365–376. (doi:10.1038/nrn3475)
13. Anderson SF, Kelley K, Maxwell SE. 2017 Sample-Size Planning for More Accurate Statistical Power: A Method Adjusting Sample Effect Sizes for Publication Bias and Uncertainty. *Psychol. Sci.* **28**, 1547–1562. (doi:10.1177/0956797617723724)
14. Cumming G. 2011 *Understanding The New Statistics: Effect Sizes, Confidence Intervals, and Meta-Analysis*. New York: Routledge. (doi:10.4324/9780203807002)
15. Carter EC, Kofler LM, Forster DE, McCullough ME. 2015 A series of meta-analytic tests of the depletion effect: Self-control does not seem to rely on a limited resource. *J. Exp. Psychol. Gen.* **144**, 796–815. (doi:10.1037/xge0000083)
16. Kvarven A, Strømland E, Johannesson M. 2020 Comparing meta-analyses and preregistered multiple-laboratory replication projects. *Nat. Hum. Behav.* **4**, 423–434. (doi:10.1038/s41562-019-0787-z)
17. Lakens D. 2015 What p-hacking really looks like: A comment on Masicampo and LaLande (2012). *Q. J. Exp. Psychol.* **68**, 829–832. (doi:10.1080/17470218.2014.982664)
18. Simmons JP, Simonsohn U. 2017 Power Posing: P-Curving the Evidence. *Psychol. Sci.* **28**, 687–693. (doi:10.1177/0956797616658563)
19. McKay B, Bacelar M, Parma JO, Miller MW, Carter MJ. 2022 The combination of reporting bias and underpowered study designs have substantially exaggerated the motor learning benefits of self-controlled practice and enhanced expectancies: A meta-analysis. (doi:10.31234/osf.io/3nhtc)
20. McMorris T, Barwood M, Hale BJ, Dicks M, Corbett J. 2018 Cognitive fatigue effects on

physical performance: A systematic review and meta-analysis. *Physiol. Behav.* **188**, 103–107. (doi:10.1016/j.physbeh.2018.01.029)
21. Carter EC, Schönbrodt FD, Gervais WM, Hilgard J. 2019 Correcting for bias in psychology: A comparison of meta-analytic methods. *Adv. Methods Pract. Psychol. Sci.* **2**, 115–144. (doi:10.1177/2515245919847196)
22. McShane BB, Böckenholt U, Hansen KT. 2016 Adjusting for Publication Bias in Meta-Analysis: An Evaluation of Selection Methods and Some Cautionary Notes. *Perspect. Psychol. Sci. J. Assoc. Psychol. Sci.* **11**, 730–749. (doi:10.1177/1745691616662243)
23. Sladekova M, Webb LEA, Field AP. 2022 Estimating the change in meta-analytic effect size estimates after the application of publication bias adjustment methods. *Psychol. Methods* (doi:10.1037/met0000470)
24. Cumming G. 2012 *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. New York, NY, US: Routledge/Taylor & Francis Group.
25. Viechtbauer W. 2010 Conducting Meta-Analyses in R with the metafor Package. *J Stat Softw* **36**, 1–48. (doi:10.18637/jss.v036.i03)
26. Stanley TD. 2017 Limitations of PET-PEESE and other meta-analysis methods. *Soc. Psychol. Personal. Sci.* **8**, 581–591. (doi:10.1177/1948550617693062)
27. Hong S, Reed WR. 2021 Using Monte Carlo experiments to select meta-analytic estimators. *Res. Synth. Methods* **12**, 192–215. (doi:10.1002/jrsm.1467)
28. Ciria LF, Román-Caballero R, Vadillo M, Holgado D, Luque-Casado A, Perakakis P, Sanabria D. 2022 A call to rethink the cognitive benefits of physical exercise: An umbrella review of randomized controlled trials. , 2022.02.15.480508. (doi:10.1101/2022.02.15.480508)
29. Coburn KM, Vevea JL. 2015 Publication bias as a function of study characteristics. *Psychol. Methods* **20**, 310–330. (doi:10.1037/met0000046)
30. Kepes S, Banks GC, McDaniel M, Whetzel DL. 2012 Publication Bias in the Organizational Sciences. *Organ. Res. Methods* **15**, 624–662. (doi:10.1177/1094428112452760)
31. Bartoš F, Maier M, Quintana DS, Wagenmakers E-J. 2022 Adjusting for Publication Bias in JASP and R: Selection Models, PET-PEESE, and Robust Bayesian Meta-Analysis. *Adv. Methods Pract. Psychol. Sci.* **5**, 25152459221109260. (doi:10.1177/25152459221109259)
32. Coburn KM, Vevea JL. 2019 weightr: Estimating Weight-Function Models for Publication Bias.
33. Lin L, Chu H. 2018 Quantifying publication bias in meta-analysis. *Biometrics* **74**, 785–794. (doi:10.1111/biom.12817)
34. Lin L, Rosenberger KJ, Shi L, Wang Y, Chu H. 2022 altmeta: Alternative Meta-Analysis Methods.
35. Schwarzer G, Carpenter JR, Rücker G. 2022 metasens: Statistical Methods for Sensitivity Analysis in Meta-Analysis.
36. Bartoš F, Schimmack U. 2022 Z-curve 2.0: Estimating Replication Rates and Discovery Rates. *Meta-Psychol.* **6**. (doi:10.15626/MP.2021.2720)
37. Rücker G, Schwarzer G, Carpenter JR, Binder H, Schumacher M. 2011 Treatment-effect estimates adjusted for small-study effects via a limit meta-analysis. *Biostat. Oxf. Engl.* **12**, 122–142. (doi:10.1093/biostatistics/kxq046)
38. Quintana D. 2022 metameta: A Meta-meta-analysis Package for R.
39. Borg DN, Bon J, Sainani KL, Baguley BJ, Tierney N, Drovandi C. 2020 Sharing Data and Code: A Comment on the Call for the Adoption of More Transparent Research Practices in Sport and Exercise Science. (doi:10.31236/osf.io/ftdgj)
40. Pageaux B, Lepers R, Dietz KC, Marcora SM. 2014 Response inhibition impairs subsequent self-paced endurance performance. *Eur. J. Appl. Physiol.* **114**, 1095–1105. (doi:10.1007/s00421-014-2838-5)
41. Penna EM, Filho E, Wanner SP, Campos BT, Quinan GR, Mendes TT, Smith MR, Prado LS. 2018 Mental Fatigue Impairs Physical Performance in Young Swimmers. *Pediatr. Exerc. Sci.* **30**, 208–215. (doi:10.1123/pes.2017-0128)
42. Thornton A, Lee P. 2000 Publication bias in meta-analysis: its causes and consequences. *J. Clin. Epidemiol.* **53**, 207–216. (doi:10.1016/S0895-4356(99)00161-4)
43. Borg DN, Barnett A, Caldwell AR, White N, Stewart I. 2022 The Bias for Statistical Significance in Sport and Exercise Medicine. (doi:10.31219/osf.io/t7yfc)
44. Wicherts JM, Veldkamp CLS, Augusteijn HEM, Bakker M, van Aert RCM, van Assen MALM.

2016 Degrees of Freedom in Planning, Running, Analyzing, and Reporting Psychological Studies: A Checklist to Avoid p-Hacking. *Front. Psychol.* **7**, 1832. (doi:10.3389/fpsyg.2016.01832)
45. Cohen J. 1962 The statistical power of abnormal-social psychological research: a review. *J. Abnorm. Soc. Psychol.* **65**, 145–153. (doi:https://doi.org/10.1037/h0045186)
46. Button KS, Ioannidis JPA, Mokrysz C, Nosek BA, Flint J, Robinson ESJ, Munafò MR. 2013 Power failure: why small sample size undermines the reliability of neuroscience. *Nat. Rev. Neurosci.* **14**, 365–376. (doi:10.1038/nrn3475)
47. Szucs D, Ioannidis JPA. 2017 Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. *PLoS Biol.* **19**, e3001151. (doi:10.1371/journal.pbio.2000797)
48. Maxwell SE, Lau MY, Howard GS. 2015 Is psychology suffering from a replication crisis? What does 'failure to replicate' really mean? *Am. Psychol.* **70**, 487–498. (doi:10.1037/a0039400)
49. OPEN SCIENCE COLLABORATION. 2015 Estimating the reproducibility of psychological science. *Science* **349**, aac4716. (doi:10.1126/science.aac4716)
50. Murphy J, Mesquida C, Caldwell AR, Earp BD, Warne JP. 2022 Proposal of a Selection Protocol for Replication of Studies in Sports and Exercise Science. *Sports Med.* (doi:10.1007/s40279-022-01749-1)
51. Camerer CF *et al.* 2016 Evaluating replicability of laboratory experiments in economics. *Science* **351**, 1433–1436. (doi:10.1126/science.aaf0918)
52. Maxwell SE, Kelley K, Rausch JR. 2008 Sample size planning for statistical power and accuracy in parameter estimation. *Annu. Rev. Psychol.* **59**, 537–563. (doi:10.1146/annurev.psych.59.103006.093735)
53. Ioannidis JPA. 2005 Why Most Published Research Findings Are False. *PLOS Med.* **2**, e124. (doi:10.1371/journal.pmed.0020124)
54. Brysbaert M. 2019 How many participants do we have to include in properly powered experiments? A tutorial of power analysis with reference tables. *J. Cogn.* **2**, 16. (doi:10.5334/joc.72)
55. Higginson AD, Munafò MR. 2016 Current Incentives for Scientists Lead to Underpowered Studies with Erroneous Conclusions. *PLOS Biol.* **14**, e2000995. (doi:10.1371/journal.pbio.2000995)
56. Smaldino PE, McElreath R. 2016 The natural selection of bad science. *R. Soc. Open Sci.* **3**, 160384. (doi:10.1098/rsos.160384)
57. Maxwell SE. 2004 The Persistence of Underpowered Studies in Psychological Research: Causes, Consequences, and Remedies. *Psychol. Methods* **9**, 147–163. (doi:https://doi.org/10.1037/1082-989X.9.2.147)
58. Turner RM, Bird SM, Higgins JPT. 2013 The Impact of Study Size on Meta-analyses: Examination of Underpowered Studies in Cochrane Reviews. *PLOS ONE* **8**, e59202. (doi:10.1371/journal.pone.0059202)
59. Pageaux B, Lepers R. 2018 Chapter 16 - The effects of mental fatigue on sport-related performance. In *Progress in Brain Research* (eds S Marcora, M Sarkar), pp. 291–315. Elsevier. (doi:10.1016/bs.pbr.2018.10.004)
60. 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. *J. Sports Sci.* **0**, 1–3. (doi:10.1080/02640414.2020.1776002)
61. Caldwell AR *et al.* 2020 Moving Sport and Exercise Science Forward: A Call for the Adoption of More Transparent Research Practices. *Sports Med.* (doi:10.1007/s40279-019-01227-1)
62. Abt G, Jobson S, Morin J-B, Passfield L, Sampaio J, Sunderland C, Twist C. 2022 Raising the bar in sports performance research. *J. Sports Sci.* **40**, 125–129. (doi:10.1080/02640414.2021.2024334)
63. Sainani KL *et al.* 2021 Call to increase statistical collaboration in sports science, sport and exercise medicine and sports physiotherapy. *Br. J. Sports Med.* **55**, 118–122. (doi:10.1136/bjsports-2020-102607)
64. Brown D, Boat R, Graham J, Martin K, Pageaux B, Pfeffer I, Taylor I, Englert C. 2021 A Multi-Lab Pre-Registered Replication Examining the Influence of Mental Fatigue on Endurance Performance: Should We Stay or Should We Go?: North American Society for the Psychology of

Sport and Physical Activity Virtual Conference. pp. 57–57. (doi:10.1123/jsep.2021-0103)
65. Vazire S. 2019 Do We Want to Be Credible or Incredible? *APS Obs.* **33**.
66. Asendorpf JB *et al.* 2013 Recommendations for Increasing Replicability in Psychology. *Eur. J. Personal.* **27**, 108–119. (doi:https://doi.org/10.1002/per.1919)
67. Lakens D. 2022 Sample Size Justification. *Collabra Psychol.* **8**, 33267. (doi:10.1525/collabra.33267)
68. Nosek BA, Ebersole CR, DeHaven AC, Mellor DT. 2018 The preregistration revolution. *Proc. Natl. Acad. Sci.* **115**, 2600–2606. (doi:10.1073/pnas.1708274114)