Preregistration

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

Authorship: Cristian Mesquida, Jennifer Murphy, Daniël Lakens & Joe Warne

1. Introduction

Empirical studies have reported that over 90% of published articles using null-hypothesis significance testing (NHST) in biomedicine and psychology reported statistically significant results (i.e., $p < .05$). If this percentage is true, it would require that the average statistical power to be over 90% and that all hypothesis tested in these studies were true. However, there is empirical evidence that the average statistical power for psychology research is much lower (Bakker et al., 2012; Cohen, 1962; Fraley & Vazire, 2014; T. D. Stanley et al., 2018). For instance, Fraley & Vazire (2014) reported that the average statistical power to detect the typical effect size in the field was about 50%. When the percentage of studies reporting statistically significant results is higher than the average statistical power for the same sample of studies, there is an excess of significant findings suggesting that there are external factors that contribute to falsely increase number of statistically significant studies.

Some factors that contribute to create an excess of significant findings are publication bias and studies with underpowered designs (Bakker et al., 2012; Scheel et al., 2021; Simmons et al., 2011). Publication bias is defined as the publishing behavior that gives studies which find support for their tested hypotheses a higher chance of being published, as opposed to studies that do not find support for the hypothesis of interest. These behaviors include editors and reviewers selectively publishing studies with significant findings (i.e., review bias) and researchers deciding not to submit studies with non-significant results (i.e., file-drawering). Thus, if there is publication bias, one should expect to observe a high percentage of studies reporting statistically significant findings regardless of the probability of the hypothesis tested being true and the average statistical power. On the other hand, it is known that the presence of studies with underpowered designs is also likely to increase the type I error beyond alpha level in the published literature (known as the positive predictive value). As a result, it is likely that the number of studies in the published literature reporting a statistically significant effect despite being a false positive is higher than the typical alpha level. Because both biases might negatively affect the trustworthiness of published findings, it is important to assess the extent to which they are present in published literature.

To date, it has been observed that between 70% and 82% of published articles in sports science journals reported statistically significant results (Büttner et al., 2020; Twomey et al., 2021). However, these estimates do not provide insights into possible reasons for these high percentages if not complemented with statistical power estimates from the same sample of studies. By comparing the percentage of studies which reported statistically significant results against the average statistical power observed in the same sample of studies, it could offer insights about whether there is an excess of significant findings above chance alone. Furthermore, unlike psychology and neurosciences where the presence of studies with underpowered designs has been investigated thoroughly (Bakker et al., 2012; Button et al., 2013; Cohen, 1962; Fraley & Vazire, 2014; T. D. Stanley et al., 2018), a systematic investigation is lacking in sports and exercise science research to date.

One way to analyze the prevalence of publication bias and studies with underpowered designs is by analyzing the distribution of *p*-values. Both *p*-curve and z-curve methods compare the *p*-value distribution from a set of studies of interest to predict biases and the average power of the studies included (see Bartoš & Schimmack, 2022 for z-curve; see Simonsohn et al., 2014a, 2014b for *p*-curve) Both methods compare the observed and expected distribution of *p*-values to determine the prevalence of bias for a set of published studies. Both provide estimates of observed average power of the studies entered into the analysis, and estimate the degree of publication bias. Given that it has been suggested that z-curve analysis outperforms *p*-curve under conditions of effect size heterogeneity (Brunner & Schimmack, 2020), we will use z-curve as we expect a high-degree of heterogeneity in our selected sample of studies. This is because the selected sample of studies will include studies from different subfields of the sport and exercise science such as sports and exercise physiology, sports and exercise performance and sports and exercise psychology as well as different populations.

By investigating the relationship between the latter two, we can examine the presence of small-study bias in a literature body.

2. Research aims (RAs)

The research aims (RA) of this preregistered study are:

RA1: To examine the presence of publication bias by conducting a z-curve analysis; and,

RA2: To assess the average statistical power in a sample of studies by conducting a z-curve analysis.

3. Hypothesis

Given that this is a relatively nascent area of empirical enquiry, our study is intended to be exploratory and no research hypothesis will be tested.

We will rely on the statistical results extracted from a sample of 350 studies published in 10 sports and exercise science journals that were screened for study 1 and 2. The sample size is based on a precision analysis used in preregistration 1 (https://osf.io/mqbr2/) to find a *true* proportion of 30% for studies that include a pre-study power calculation in the field of sport and exercise science. Such analysis returned a sample size of 350 studies (35 studies per journal).

4. Methods

The pre-registration of our study occurs prior to collecting and coding of data. This preregistration is part of a research project consisting of 3 studies, which will be preregistered all at once, where we aim to survey a large sample of studies from sport and exercise science literature to investigate:

a) Prevalence, reporting practices and reproducibility of pre-study power calculations (preregistration 1: https://osf.io/mqbr2/);

b) Reporting practices and reproducibility of *p*-values (preregistration 2: https://osf.io/3juyq/); and

c) Prevalence of publication bias and studies with underpowered designs by using a z-curve analysis (current preregistration).

The data for these 3 preregistrations will be coded over one single data collection. As a result, the same sample of studies will be used in each of the 3 preregistered studies. Once the data collection is over, data collected through Google forms will be downloaded in the form of a unique dataset as an excel spreadsheet. Deviations from this original protocol will be explicitly acknowledged and documented in any final report of this study.

4.1. Study design

This is a retrospective observational study with a cross-sectional design.

4.2. Study sample size

The sample size is based on a precision analysis used in preregistration 1 (https://osf.io/mqbr2/) to find a *true* proportion of 30% for studies that include a pre-study power calculation in the field of sport and exercise science. Thus, a sample of 350 studies published in 10 sports and exercise science journals will be screened for the purposes of this study.

4.3. Data extraction

The major piece of information will be *p*-values for the z-curve. Only one *p*-value per study will be extracted, which will correspond to the effect of interest stated in the primary hypothesis of each study. For scientific results to be interpretable, it is imperative that researchers disclose how ambiguities surrounding the collection and analysis of data have been resolved (Simmons et al., 2011). As recommended by Simonsohn et al., (2014) for *p*-curve users, we will provide a disclosure table to justify the coding decisions regarding the selected key statistical result (**Table 1**).

Table 1. Coding strategy to create a z-curve disclosure table.

Item	**Action**
1. Primary hypothesis quote	Copy pasted quote of the hypothesis statement
2. Does the primary hypothesis predict the presence of effect or the absence of an effect?	Select from a list (*effect*/*no effect).*
3. Identify study design	Select from a list (*paired t-test, unpaired t-test, one-way within-subject ANOVA, one-way between-subject ANOVA*, *two-way within-subject ANOVA*, *two-way between-subject ANOVA*, *two-way mixed ANOVA, another)*
4. Based on the primary hypothesis, what is the key statistical result?	Select from a list (*differences of means, simple effect, interaction effect, unclear*)
5. Statistical result quote	Copy pasted quote of the statistical result of interest based on the effect of interest
6. Recompute the exact *p*-value based on the reported test statistic	Report recomputed *p*-value

4.4. Recomputing *p*-values

The z-curve method requires exact *p*-values (e.g., $p = 0.002$) as input parameter. If the corresponding *p*-value is reported relatively (e.g., $p < 0.05$), the *p*-value will be recomputed where sufficient information is available (i.e., degrees of freedom and *F*-ratio or *t*-statistic). *P*-values will be 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 statistic test and degrees of freedom. In case where a *t*-statistic or *F*-ratio from a one-way ANOVA with two levels is reported but not the degrees of freedom, degrees of freedom will be estimated using the sample size per group and study design reported in the original study. For instance, in case the key statistical test is a within-subject *t*-test or a one-way within-subject ANOVA with two levels, we would calculate the degrees of freedom as N – 1. In case the statistical test is a between-subject (unpaired) *t*-test or a one-way between-subject ANOVA with two levels, we will calculate the degrees of freedom as N – 2. Once the degrees of freedom are calculated, the *p*-value will then be recomputed.

In case that the exact *p*-value and the corresponding *t*-statistic is not reported, but an effect size is available, authors will attempt to convert effect sizes into *p*-values for study designs involving a *t*-test. Converting the effect sizes into *p*-values is different depending on whether the authors present their effect sizes as Cohen's *d*, Cohen's d_z, or Hedges' *g*. To compute *p*-values for between-subject designs, it is necessary first to convert Hedges' *g* back into Cohen's *d*. The basic formula for the conversion given by Hedges (1984) will be used:

$$d = \frac{g}{c(m)}$$

where c(m) = 1 – (3/((4 * m), where *m* is the degrees of freedom for the study. *m* will be calculated as *N* – 2 for studies that used between-subjects designs and *N* – 1 for studies that used within-subjects designs. If Cohen's *d* is what the authors report in their study, then no conversion is necessary and the Cohen's *d* can be converted directly into a *t*-statistic.

If the *t*-statistic is from a between-subjects design and sample sizes for the treatment and control conditions are provided, it is converted to a *t*-score using the following formula:

$$t = \frac{d}{\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$$

If the sample sizes for the treatment and control groups are *not* provided, the *t*-statistic will be calculated as follows:

$$t \approx \frac{d}{2} * \sqrt{n}$$

In the case of within-subjects designs, the Cohen's d_z will be converted to a *t*-statistic using the following formula:

$$t = d_z * \sqrt{n}.$$

If the statistical test is a one-way ANOVA with two levels, and *F*-ratios and degrees of freedom are reported, we will first compute η_p^2 as follows:

$$\eta_p^2 = \frac{F\ x\ DF\ effect}{F\ x\ DF\ effect + DF\ error}$$

And then η_p^2 will be converted to a Cohen's *d* using the following formula:

$$d = \frac{2 * \sqrt{n^2}}{\sqrt{1 - n_2}}$$

Note that for a one-way ANOVA with two levels, η_p^2 equals η^2. Thus, if the F-ratio is not reported, we could also estimate Cohen's d from either η_p^2 or η^2. Finally, we won't attempt to compute other ANOVA effect sizes (i.e., ω^2, ω_p^2) because they require information that is seldom reported in published studies such as mean-square (MS) and sum-of-squares (SS) errors. If we are unable to recompute a *p*-value reported relatively (e.g., $p < 0.05$) by any of above procedures, the *p*-value will not be included in the z-curve analysis.

4.5. Data analysis

All statistical tests will be conducted in R (R Core Team, 2019). *P*-values extracted from studies that tested for the hypothesis of no effect (these studies were labeled in preregistration 2) will not be used in the z-curve analysis because they can distort the results of the analysis. This is because just like researchers can be tempted to find evidence in favor of the hypothesis of no effect by exploiting researchers' degrees of freedom, researchers testing the hypothesis of no effect can also introduce bias to find evidence of in favor of the null hypothesis. However, if a sufficient number of studies (n > 10) tested the hypothesis of no effect, the corresponding *p*-values will be used for an alternative z-curve analysis. Research aims 1-2 (**RA1** and **RA2**) will be addressed by conducting a z-curve analysis using the *z-curve 2.0* R package (Bartoš & Schimmack, 2022). The z-curve method allows to estimate the following parameters (Bartoš & Schimmack, 2022):

- **Observed Discovery Rate (ODR):** the relative frequency of significant results. For instance, if a sample of 10 studies produce 4 significant results, the ODR is 40%.
- **Expected Discovery Rate (EDR):** the average observed power of all studies with both significant and non-significant results.
- **Expected Replicability Rate (ERR)**: the average power of only significant results. It represents the estimated proportion of significant studies that would yield another significant effect if subjected to a direct replication.
- **Sorić false discovery rate (SFDR)**: the maximum percentage of studies that could be false positives.

RA1: will be answered by considering the following outcomes: a) whether the point estimate of ODR lies within the 95% CI of the EDR or not. When the point estimate of ODR lies within the 95% CI of the EDR it can be concluded that there is statistical evidence in favor of the presence of publication bias.

RA2: will be addressed by measuring the EDR which is an estimate of the observed average power. Besides the parameters required to answer RQ1 and RQ2, we will also report the estimate for **SFDR** given that the presence of studies with underpowered designs can increase the type I error rate.

5. References

Aert, R. C. M. van, Wicherts, J. M., & Assen, M. A. L. M. van. (2019). Publication bias examined in meta-analyses from psychology and medicine: A meta-meta-analysis. *PLOS ONE*, *14*(4), e0215052. https://doi.org/10.1371/journal.pone.0215052

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

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

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. https://doi.org/10.1136/bjsports-2019-101863

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

Carter, E. C., Schönbrodt, F. D., Gervais, W. M., & Hilgard, J. (2019). Correcting for bias in psychology: A comparison of meta-analytic methods. *Advances in Methods and Practices in Psychological Science*, *2*(2), 115–144. https://doi.org/10.1177/2515245919847196

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

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

Cristea, I. A., Georgescu, R., & Ioannidis, J. P. A. (2022). Effect Sizes Reported in Highly Cited Emotion Research Compared With Larger Studies and Meta-Analyses Addressing the Same Questions. *Clinical Psychological Science*, *10*(4), 786–800. https://doi.org/10.1177/21677026211049366

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

Fritz, C., Morris, P., & Richler, J. (2011). Effect Size Estimates: Current Use, Calculations, and Interpretation. *Journal of Experimental Psychology. General*, *141*, 2–18. https://doi.org/10.1037/a0024338

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

Hedges, L. V. (1984). Estimation of Effect Size under Nonrandom Sampling: The Effects of Censoring Studies Yielding Statistically Insignificant Mean Differences. *Journal of Educational Statistics*, *9*(1), 61–85. https://doi.org/10.2307/1164832

Kühberger, A., Fritz, A., & Scherndl, T. (2014). Publication Bias in Psychology: A Diagnosis Based on the Correlation between Effect Size and Sample Size. *PLOS ONE*, *9*(9), e105825. https://doi.org/10.1371/journal.pone.0105825

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis* (pp. ix, 247). Sage Publications, Inc.

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

Nuijten, M. B., van Assen, M. A. L. M., Augusteijn, H. E. M., Crompvoets, E. A. V., & Wicherts, J. M. (2020). Effect Sizes, Power, and Biases in Intelligence Research: A Meta-Meta-Analysis. *Journal of Intelligence*, *8*(4), E36. https://doi.org/10.3390/jintelligence8040036

Rosenthal, R., & DiMatteo, M. R. (2001). Meta-analysis: Recent developments in quantitative methods for literature reviews. *Annual Review of Psychology*, *52*, 59–82. https://doi.org/10.1146/annurev.psych.52.1.59

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

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

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

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

Stanley, T. D. (2017). Limitations of PET-PEESE and Other Meta-Analysis Methods. *Social Psychological and Personality Science*, *8*, 194855061769306. https://doi.org/10.1177/1948550617693062

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

Stanley, T., D., & Doucouliagos, H. (2014). Meta-regression approximations to reduce publication selection bias. *Research Synthesis Methods*, *5*(1), 60–78. https://doi.org/10.1002/jrsm.1095

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

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