We thank both reviewers for taking the time to review this manuscript and for giving us the opportunity to respond to their comments.

Responses to Reviewer #1

Comment 1

However, I have concerns about their conceptualization of replicability, contribution, and methodological approach. The authors primarily conceptualize replicability in terms of retrospective (or post-hoc) average power (also known as the "expected discovery rate" / EDR).

Response

The most direct way to estimate replication rates is through large-scale direct replication projects. However, such projects require substantial resources that are often unavailable to many sports scientists. An alternative approach is to examine the average statistical power of a body of studies and assess the potential presence of bias.

The probability that a previous finding will replicate depends on both statistical power and the probability that the tested hypothesis is true. True statistical power is unknown because the true effect size is unknown. Nevertheless, the observed effect size from a prior study—although potentially distorted by sampling error—provides a more informative estimate of the true effect than arbitrary benchmarks such as Cohen's *d* thresholds.

When the estimated average power of a set of studies is very low while the proportion of statistically significant results is high, this discrepancy is consistent with selective reporting, or other forms of selection bias. This discrepancy can't emerge under other conditions than selection bias. The central finding of the present study is that average statistical power in the sampled literature is very low. This estimate does not predict the replicability of future individual studies in the field, nor does it provide an estimate of the replication rate in future replication projects. The probability that a study replicates can't be predicted (see Miller, 2011), as we do not know the proportion of true effects and null effects in the literature. Nevertheless, even though average power cannot be used to make claims about the replicability of any individual study, it does indicate that, within the sampled literature, studies are severely underpowered. If the 269 studies were replicated using the same methods and sample sizes, the expected average replication rate would therefore be low assuming all investigated hypotheses are true, regardless of the exact proportion of true and null effects. Low statistical power is also problematic because non-significant results become difficult to interpret, and, when combined with selection bias, it leads to inflated estimates of true effect sizes—one of the key findings of the Sports Science Replication Project. Thus, estimating the average power of a published set of studies provides valuable information about the robustness and credibility of findings in the literature.

Miller, J., & Schwarz, W. (2011). Aggregate and individual replication probability within an explicit model of the research process. Psychological Methods, 16(3), 337–360. https://doi.org/10.1037/a0023347

Comment 2

Average power is a meta-analytic analogue of single study post hoc power. Single study post hoc power has been greatly lampooned for many decades now (Hoenig & Heisey, 2001; Yuan & Maxwell, 2005). For example, Greenland (2012) writes that post hoc power computed from completed studies is:

"Irrelevan[t]: Power refers only to future studies done on populations that look exactly like our sample with respect to the estimates from the sample used in the power calculation; for a study as completed (observed), it is analogous to giving odds on a horse race after seeing the outcome."

In addition, average power is not relevant to the replicability of actual prospective replication studies. As McShane, Bockenholt, and Hansen (2020) write: "Average power is relevant to replicability if and only if replication is defined in terms of statistical significance within the classical frequentist repeated sampling framework. As this framework is both purely hypothetical and ontologically impossible, average power is not relevant to the replicability of actual prospective replication studies."

Response

The reviewer cites Hoenig and Heisey (2001) and Yuan and Maxwell (2005) to criticize the use of observed power for single studies. We ourselves have explained the problems of post-hoc power (Lakens, 2022), and there are few researchers in the world that have developed a better understanding of statistical power than we have. We agree that calculating observed power for individual studies is not very informative, as the estimated observed power does not provide more information than the reported *p*-value. Moreover, when statistical power is low, the confidence interval around an observed power estimate from a single study will be very wide, further limiting its interpretability (https://replicationindex.com/2015/03/24/an-introduction-to-observed-power-based-on-yuan-and-maxwell-2005/).

We cannot see how this argument is relevant to the present study, as our analysis does not rely on observed power estimates from individual studies. Instead, analogous to how meta-analysis combines effect size estimates across studies to increase precision and reduce uncertainty about the true effect size, *z*-curve aggregates information from the distribution of *p*-values (or *z*-scores) across studies. This approach enables estimation of average observed power and corresponding confidence intervals at the meta-analytic level, where uncertainty is substantially reduced relative to single-study estimates.

We both strongly agree and strongly disagree with McShane and colleagues' claim that the frequentist framework is purely hypothetical, and ontologically impossible, depending on how the quote is meant. First, of course all statistical models are false. So the statement that they are ontologically impossible is empty and trivial. All knowledge we have is wrong. In that sense, we agree. But we disagree in the sense that many models are useful. We also disagree that power is not relevant to the replicability of prospective replication studies. Again, in theory, we could defend their argument. If we would perform infinitely large studies (say we collect a billion observations for every study we will do from now on in sport science) then observed power is irrelevant, and the only thing that matters in a replication study is if there is an effect large enough to matter, or not. But the field of sport and exercise science is not going to collect a billion observations in the future. The sample sizes have been stable across decades, and this seems unlikely to change. In a research area where scientists collect relatively similar sample sizes in their studies, the average statistical power is an important predictor of replicability.

McShane et al. (2020) argue that average power is not relevant to replicability and is ontologically problematic because exact replications are rarely possible in psychological research. Their argument is based on two main points: first, replication studies typically employ larger sample sizes than original studies; second, effect size heterogeneity can influence replication outcomes. Although we strongly support the idea that sports and exercise scientists should collect larger studies when they build on (or directly replicate) earlier work, this is currently not the case, because our own work has shown that studies with a-priori power analyses do not lead to noticeably larger sample sizes (Mesquida et al. 2025). The expected discovery rate (average power) and the expected replication rate (average power among studies reporting statistically significant results) are calculated assuming the same experiment using the same methods, procedures, and sample size. Under these definitions, heterogeneity should be small, and the assumption that sample sizes remain stable is in line with what happens in the literature. We agree that if the field would collectively increase sample sizes, we should expect higher replicability rates. This is also why in the discussion we wrote: "Researchers should also design their studies with high power by conducting rigorous a priori power analyses. Unfortunately, only 41% of studies in our sample performed an a priori power analysis to justify the sample size, and of those, many were poorly conducted [40].".

As noted by McShane et al. (2020), "average power is relevant to replicability if and only if replication is defined in terms of statistical significance within the classical frequentist repeated sampling framework." This is precisely the framework adopted in our definitions of both expected discovery rate and replication rate. Accordingly, estimating average observed power remains informative, as it provides guidance for future research planning—for example, by indicating the need for larger sample sizes or for study designs capable of detecting smaller effect sizes.

McShane et al. (2020) raise additional criticisms in their article, such as that the average power on analyses of relatively small sets of studies (approximately 30), for which estimates are necessarily imprecise. Precision increases with the number of studies analyzed. In the present work, our analysis includes 269 studies, so this

criticism does not apply, and the precision of the estimates is explicitly quantified using 95% confidence intervals. Accordingly, the uncertainty of our results should be evaluated on the basis of the reported estimates and their confidence intervals, rather than by reference to findings from studies that did not examine *z*-curve–based estimates.

Mesquida, C., Murphy, J., Warne, J. & Lakens, D. (2025). Prevalence, reporting practices, and methodological quality of a priori power analyses in sports and exercise science research. SportRXiv.

Comment 3

Pek et al (2022) also note ontological concerns with average power. Pek et al (2024) further note that (as per the present authors' approach) "using power for evaluating completed studies can be counterproductive."

Response

From the reviewer's comment, it is not entirely clear what the precise object of critique is. Pek et al. (2023) cite McShane et al. (2020) to raise ontological concerns about the use of average power, and we have already expressed our disagreement with those arguments above.

The reviewer further quotes the statement from the abstract of Pek et al. (2023) that "using power for evaluating completed studies can be counterproductive." In their discussion of the power of published research (see also the last column in Table 1), Pek et al. (2024) describe several common approaches to estimating observed power: (a) using fixed effect size benchmarks such as Cohen's *d* thresholds, (b) using empirically derived benchmark effect sizes, and (c) using effect sizes obtained from meta-analyses. The benchmark approach based on Cohen's *d* thresholds is problematic because these thresholds were developed in the context of psychology and social sciences and are not necessarily applicable to sports and exercise science. The use of empirically derived effect sizes also has important limitations, as such estimates may be affected by selection bias and may not generalize across research contexts (e.g., using effect sizes from strength and conditioning studies to design studies on physical activity and cognitive performance).

These critiques, however, do not apply to *z*-curve estimates. Unlike the approaches discussed by Pek et al., *z*-curve estimates expected values based on the distribution of statistically significant results and the implied population effect sizes, rather than on arbitrary benchmarks or empirically derived effect size distributions from potentially biased literatures.

Finally, Pek et al. (2023) also criticize the use of observed power on the grounds that such estimates are highly imprecise due to uncertainty in observed effect sizes. We agree that this concern is particularly relevant for single

studies when statistical power is very low, leading to wide confidence intervals for individual power estimates. However, as also noted previously, we calculate average power based on 269 studies, so this criticism does not apply.

If the reviewer can explain in more detail why the criticisms by Pek et al would be relevant for our research, we could provide a more targeted rebuttal, but as it stands, the points raised in Pek et al are not relevant for our paper.

Comment 4

While I have thus far focused on the primary manner in which the authors conceptualize replicability (i.e., average power / EDR), exactly the same concerns apply to the secondary manner (i.e., the "expected replication rate" / ERR).

Response

Both the expected discovery rate and the expected replication rate refer to the average statistical power of a set of studies, with the latter being conditional on studies that support the tested hypothesis. Accordingly, the same considerations and responses outlined above apply equally to this comment. Observing an expected replication rate of 50% among studies that supported the hypothesis is therefore informative, as it indicates that, on average, only about half of these studies would be expected to yield statistically significant results if repeated under the same conditions and with the same sample sizes. This information is valuable for planning future studies, for example, by highlighting the need for larger sample sizes, which is one of our main recommendations in the discussion.

Comment 5

Rosenthal was a pioneer studying replication in psychology. Drawing on his work dating from the 1960s, Rosenthal (1990) dismissed evaluations of replicability that are dichotomous and based on significance testing as "the traditional, not very useful view of replication" and advocated evaluations of replicability that are continuous and based on effect sizes as "the newer, more useful view of replication".
The authors' approach in this paper is dichotomous and based on significance testing and thus falls squarely in what Rosenthal thirty-five years ago today already termed "the traditional, not very useful view of replication."

Response

We disagree with Rosenthal's criticism of NHST. Using NHST to make dichotomous claims is an important tool within methodological falsificationism (see Tunç, Tunç and Lakens, 2023), and it is the approach employed in all studies included in this project as well as in most research in sports and exercise science. Dichotomous claims allow for the corroboration or falsification of theoretical predictions. In addition, they serve a pragmatic function by facilitating scrutiny and critique among peers through contestable and testable claims. For example, *z*-curve provides a formal test for selection bias: if the upper bound of the confidence interval for the expected discovery

rate does not include the observed discovery rate, this indicates evidence consistent with selective reporting in the literature. Although we respect Rosenthal's work on publication bias, we are not impressed by his criticism on significance testing. His criticism is underdeveloped, and might have been educational for psychologists 35 years ago, but we do not consider his views the state of the art in 2026. We have pushed back on the rather simplistic criticism of NHST by Rosenthal and others in Lakens (2021) and Tunc, Tunc, and Lakens (2023). Although we greatly enjoy explaining why the criticism on NHST is largely unfounded, we do not believe this article is the place for this discussion.

Lakens, D. (2021). The practical alternative to the *p* value is the correctly used *p* value. *Perspectives on Psychological Science, 16*(3), 639–648. https://doi.org/10.1177/1745691620958012

Tunç, D. U., Tunç, M. N., & Lakens, D. (2023). The epistemic and pragmatic function of dichotomous claims based on statistical hypothesis tests. *Theory & Psychology, 33*(3), 403–423. https://doi.org/10.1177/09593543231160112

Comment 6

"It is therefore not surprising that a common finding among replication projects is that unbiased replication studies with larger sample sizes produce much smaller effect sizes [9,25,26]. For instance, the ### 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% [9]."

As can be seen, the ### replication project takes a continuous quantitative view based on effect sizes, reporting that the median decrease in the effect size estimates was 75% and going on to characterize the full distribution of effect size differentials in Figures 1 and 2 of that paper.

I do not find the present authors' retrospective and dichotomous approach based on significance testing to be an advance over the ### replication project's prospective and continuous approach based on effect sizes. Indeed, I view it as retrograde.

Response

The goal of this project was not to estimate the overestimation of effect sizes, as this would have required conducting direct replications of the studies, and comparing the new effect sizes against the original effect sizes. Our goal was to estimate average power and to assess selection bias. We both test for the presence of bias, provide a descriptive overview of this bias (in Figure 4), and estimate the average power. We agree that the results of the Sports Science Replication Centre are extremely interesting, and provide important information, such as a median decrease in effect sizes of 75%. However, performing replications is a lot of work (the Sports and Exercise Replication project took more than 5 years), and only 25 replication studies were performed. We believe our current analysis, based on 269 studies, provides important complementary knowledge about the state of our field. While the results of the replication

project in sports and exercise science was informative, it raises the question: why did effect sizes decrease so dramatically? Our study addresses this question by showing that a combination of selection bias and underpowered study designs leads to inflated effect size estimates in a different, and much larger, set of studies than was the focus of the replication project. Specifically, our findings provide evidence that the sampled studies had very low statistical power and clear indications of selection bias. This offers a potential explanation for the observed decrease in effect sizes in the Sports Science Replication Center. The argument presented here seems to suggest that investigating the same scientific problem through multiple approaches is a waste of resources; however, we do not agree. Both converging and diverging findings across approaches foster critical evaluation of evidence and methods, providing valuable additional insights.

Comment 7

Even for those who prefer a dichotomous approach based on significance testing, when such is applied to the sports science replication project, we get a result similar to the present authors' result (see middle of page 12 of their manuscript). Therefore, in a very important sense, the present authors' result is already known (or at least cannot be said to be novel).

Response

This article used *z*-curve to examine the credibility of a literature based on a different and much larger set of articles than those included in the actual replication project. The reviewer seems to treat both studies as similar, in the sense that they both point to a problem with the replicability of findings. But the two projects actually provide different novel, converging, results. Our project shows widespread low power, and selection bias. The replication project was not able to provide these specific insights. The replication project showed replication rates are low, and effect sizes are inflated. When two different methods using different datasets yield consistent results, it provides evidence that the findings are not driven solely by sampling error (e.g., replication studies may have selected studies with easy or inexpensive designs) or methodological biases (e.g., weaker effects in replication studies due to researchers' expertise).

Comment 8

The authors' use the forensic Z-curve meta-analytic procedure of Brunner & Schimmack (2020) and Bartos & Schimmack (2022).

On page 3 of their manuscript, they note that they could use the forensic P-curve meta-analytic procedure of Simonsohn, Nelson, and Simmons instead.

In a forthcoming Journal of the American Statistical Association paper, Morey and Davis-Stober provide a formal

analysis that proves that the P-curve has poor statistical properties. For example, they prove that the P-curve produces inconsistent estimates of average power / EDR.

One might question the relevance of this to the Z-curve and thus the present manuscript. I quote the final paragraph of Morey and Davis-Stober:

"As a final point, we suggest that meta-scientists be more skeptical of procedures like the P-curve in the meta-scientific literature. Papers introducing them are often light on statistical exposition, using metaphors [and] a few simulations to make sweeping arguments. Simulation is a powerful tool and can help build intuition, but it is not a substitute for formal analysis. Simulation may provide hints of problems with a procedure, but only if the simulator's formal knowledge helps guide the choice of simulations. A simulator might quit after running a few simulations that tell them what they think is true while problems remain uncovered. Given the implications of poor forensic procedures for science, all such procedures demand deeper formal scrutiny."

This forthcoming paper is extremely relevant to the present manuscript because the very paragraph above could be written about the Z-curve.

Response

Brunner and Schimmack (2020) directly compared *p*-curve and *z*-curve and demonstrated that *p*-curve fails when data are heterogeneous, as they typically are and as they are in the present article (heterogeneity: ERR > EDR; homogeneity: ERR = EDR). Schimmack and Brunner have also published several subsequent criticisms of *p*-curve. Morey and Davis-Stober's article further reinforces these criticisms, and the authors of *p*-curve have not defended their method against them. Thus, while *p*-curve was an early attempt to estimate the true power of a set of studies, it has been shown to be unreliable.

It is therefore not valid to assume that any criticism of *p*-curve automatically applies to a fundamentally different method such as *z*-curve. The z-curve method has been extensively validated in simulation studies, and performs well with typical datasets, including of the type analyzed in this article. The convergence between *z*-curve predictions and the results of the actual replication project further supports the validity of *z*-curve. If the method were flawed, it would not produce estimates that align with outcomes observed in large-scale replication studies.

We believe the criticism on p-curve was useful in educating researchers about how all publication bias methods have strong assumptions, and these are never perfectly aligned with reality. The criticism of Morey and Davis-Stober did remind us that people need to be reminded of the limitations of all bias detection methods. Therefore, we added this as a fourth limitation:

"First, all bias-detection methods rely on simplified models of selection bias and therefore rest on assumptions that will inevitably be violated to some extent. Although z-curve provides one of the most informative tools currently available for drawing inferences about selection bias in sports and exercise science, the true state of the field cannot be known in the absence of complete transparency. Consequently, z-curve results should be interpreted as our best available indication of selection bias, while recognizing that no statistical method can perfectly quantify the degree of bias in a scientific literature."

Comment 9

Turning back to this manuscript and its use of the Z-curve, in short, we at present know next to nothing about the statistical properties of the z-curve (just as we knew next to nothing about the statistical properties of the P-curve until Morey and Davis-Stober came along). The statistical properties of the z-curve may be as poor or worse than those of the p-curve. Or they may be solid. We simply cannot say. Morey and Davis-Stober write:

"Given the stated purpose of the P-curve—evaluating the trustworthiness of scientific literature—the stakes are too high to use tests with such poor, or poorly-understood, properties."
The same applies to the Z-curve which has the same stated purpose. As a consequence, I remain very skeptical of any use of the Z-curve until its properties have been investigated formally and shown not to be wanting—especially given the very high stakes involved.

Response

The aphorism by statistician George Box, "All models are wrong, but some are useful," reminds us that all statistical tools have inherent limitations. This applies to the z-curve as well, which, like other methods, relies on specific assumptions and conditions (Brunner & Schimmack, 2020; Bartoš & Schimmack, 2022). However, these limitations do not diminish its usefulness. In fact, the convergence between z-curve predictions and the results of actual replication projects supports its validity: if the method were fundamentally flawed, it would not produce estimates that align with outcomes observed in large-scale replication studies. Additionally, the z-curve provides confidence intervals, allowing researchers to assess the uncertainty surrounding its estimates. We fully support scrutiny of scientific methods, but there is no need to throw out the baby with the bathwater. If the reviewer wishes that we only use statistical methods that yield results that perfectly align with reality, we can't use any statistical tests available to scientists.

The properties and performance of z-curve have been carefully evaluated by its original authors (Brunner & Schimmack, 2020; Bartoš & Schimmack, 2022), and it is therefore neither accurate nor nuanced to claim that we know "next to nothing" about the properties of the test. When a method is first published, its popularity and potential for critique may not be immediately evident. For example, p-curve analysis was first published in 2014 (Simonsohn et al., 2014) and updated in 2015 (Simonsohn et al., 2015). A strong critique highlighting major

limitations of p-curve appeared as a preprint in 2017 and in print in 2019 (Carter et al., 2019), demonstrating that p-curve performs poorly under heterogeneity—a common feature of empirical data. Methods such as z-curve were subsequently developed and shown to perform better under heterogeneity (Brunner & Schimmack, 2020), provided that failure conditions are realistic.

Given this evidence, it seems a stretch to claim that "the statistical properties of the z-curve may be as poor or worse than those of the p-curve," particularly since z-curve has demonstrated superior performance (Brunner & Schimmack, 2020; Bartoš & Schimmack, 2022). Nonetheless, we acknowledge all bias detection methods have limitations, and included the fourth limitation noted above to clarify this more strongly for readers.

Comment 10

1. Page 2, Table 1, etc.: You refer to these four quantities as "parameters" but they are not parameters. The word parameter has a formal definition within the context of a statistical model and these do not qualify. These are outputs or estimands but not parameters.

Response

Thanks for pointing out this inaccuracy. ODR, EDR ERR and FDR are estimates of population parameters rather than parameters themselves. Accordingly, we have replaced the term "parameter" for "quantity".

Comment 11

Page 3: You assert (arguably rather blithely) that the z-curve's independence assumption is met in your analysis because only one p-value per study is included in the analysis. This is of course not necessarily true. If, for example, the 269 studies share authors or sets of authors, that could induce dependence. There are of course many additional sources of possible dependence. One simply cannot say.

Response

This statement is incorrect, and statistical independence is not affected by whether studies share authors or have other non-statistical interrelationships. The independence assumption is about the sampling error of studies and each new sample has a new sampling error. If all studies used z-tests and had the same effect size and sample size, we expect an average sampling error of 1. When studies are heterogeneous, there is additional variation due to real differences in the non-centrality parameters (the location of the normal distribution on the x-axis of z-values) that describes the sampling distribution, but that is irrelevant for z-curve because it makes no assumptions about that distribution. Some studies may be close to z = 0 and others may be close to 3. That is heterogeneity, not dependence in sampling errors. Dependence of sampling errors would only be a problem if we included multiple tests based on the same data (e.g., correlated dependent variables), but we can exclude this possibility with certainty.

Comment 12

The authors discuss many subjective choices or value judgments as if they were objective. An example that recurs throughout the manuscript is the discussion and use of alpha = 0.05 and power = 0.80. As is well known, any choice of alpha and power reflects a particular tradeoff between the relative costs of Type I errors versus Type II errors. Except in very narrow circumstances where these relative costs can be objectively quantified (e.g. industrial quality control), these relative costs reflect a particular subjective utility (or loss) function. This subjective function will in turn vary by context or even by different people working within the same context (Neyman, 1977). This is why some have made calls for researchers to "justify their alpha" and power in light of their subjective preferences and idiosyncratic research contexts (see, for example, Lakens et al, 2018). It would be helpful if the authors discussed a range of possible (alpha, power) pairs. Alternatively, if they believe (alpha = 0.05, power = 0.80) are objectively justified in their setting, please state that and argue in favor of it. This comment applies more broadly to other quantities that the authors tend to suggest are objective (e.g., the percentage of studies with "statistically significant" results, the replication rate, etc.): either recognize the subjectivity involved or justify the values of these quantities that you believe are objectively optimal.

Response

We agree that researchers should justify their chosen Type I and Type II error rates. However, this requirement does not apply to the present study. We use a 5% significance criterion in the z-curve because all sampled studies applied this threshold. We therefore can't freely decide which alpha level to use - we have to take over the alpha levels used unanimously by researchers in this literature, and all used 5%. Similarly, we do not use 80% power to design any novel study, but use it merely as a conventional example. Although we could use any random power level, our extensive experience in educating scientists about these topics has made us realize that using non-conventional numbers (e.g., 93%) confuses readers, as they will think there is a special reason we are using a non-conventional number instead of what is typical (i.e., 80%).

Responses to Reviewer #3

Comment 1

Interpretation of the estimated discovery rate and observed discovery rate. The authors assert that a big difference between the expected discovery rate and observed discovery rate indicates the presence of publication bias. This is not accurate. The difference between the EDR and ODR indicates the presence of p-hacking. If the degree to which the original studies were powered to detect true effects (i.e., the EDR) is significantly lower than the percent of studies reported that showed a significant effect (i.e., the ODR), then this suggests that the only way such a high ODR could have been achieved in light of the average power of the original studies was through some kind of p-hacking. Schimmack (2012) is a good reference for understanding this. This also means that the authors assertion on page 9 that, "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," is inaccurate.

Response

The reviewer is correct that we should not have claimed that the difference between the EDR and ODR reflects solely publication bias, but the cause is also not solely due to p-hacking. Factually, *Z*-curve cannot distinguish between publication bias and *p*-hacking. As Bartos and Schimmack (2025) note: “Publication bias may co-occur with questionable research practices (QRPs) (Johnetal.,2012; Stefan & Schönbrodt, 2023; Nagy et al., 2025; Mathur, 2024). Some QRPs are effectively equivalent to publication bias (e.g., selective reporting); however, others, such as aggressive optional stopping, might create more severe deviations of the distribution of test statistics than publication bias. While this pattern will be clearly detectable on the z-curve plot, it is probably not distinguishable from the similar patterns introduced by publication bias.” Accordingly, we have replaced the term “publication bias” with “selection bias” throughout the manuscript.

This point has been made repeatedly in the literature and in other writing, such as the ROBMA package: “Publication bias and QRP might produce similar patterns of results. The z-curve diagnostics cannot distinguish between them” (https://cran.r-project.org/web/packages/RoBMA/vignettes/ZCurveDiagnostics.html)

Both phenomena contribute to an inflated discovery rate. While *z*-curve provides a test for selection bias, it cannot determine whether the bias arises from publication practices or *p*-hacking, which Schimmack often refers to as selection bias.

Bartos, F. & Schimmack, U. (2025). Z-Curve Plot: A Visual Diagnostic for Publication Bias in Meta-Analysis. https://doi.org/10.48550/arXiv.2509.07171

Comment 2

Omission of the file drawer ratio. Given the authors' interest in publication bias in this research—an interest which is, of course, more than warranted—I am puzzled by the fact that they at no point mention, nor do they report, the file drawer ratio. This is z-curve analysis's most straightforward measure of publication bias. The file drawer ratio is a ratio indicating how many unpublished, nonsignificant studies are predicted to exist for every one published, significant finding. A high FDR suggests high publication bias. Given the other z-curve results the authors report, I would predict the FDR would be quite high, too, but it should be reported, as it typically is in reporting the results of z-curve analyses (e.g., Fremling et al., 2025).

Response

We have now included the estimate in the Results section and addressed its implications in the Discussion.

In the Results section, we added:

"The point estimate of the File-Drawer Ratio was 8 (95% CI [2, 19]), suggesting that for every published significant result, z-curve predicts eight unpublished studies with non-significant results."

In the Discussion section, we added:

"The concept of the file drawer was introduced by Rosenthal (1979) and refers to unpublished studies that produced non-significant results. If studies had 80% power, there would be only one non-significant study in the file drawer for every four published significant studies (File-Drawer Ratio = 1:4 or 0.25:1). If studies had only 20% power, there would be four non-significant studies for every published significant study (File-Drawer Ratio = 4:1). Thus, the impact of file-drawer bias increases as study power decreases.

The file-drawer problem has important implications for accessing an unbiased literature. Underpowered studies can be combined in a meta-analysis to estimate the true effect size accurately, but this requires that all effect sizes are published. If some effect sizes remain unreported, meta-analytic estimates of the true effect size may be biased."

Comment 3

Coding procedure. I have only a couple quibbles with the coding procedure the authors used, but I think they are important to address. The authors write that they decided that for studies where no exact p-value could be calculated, if the p-value for a particular statistical test was only reported as "p < .05" or "p > .05," then they did

not include them in the analysis at all. However, they then go on to write that p-values reported as "p < .001" were imputed as .0001, and "p < .005" were imputed as .0005 and included in the analysis. They do not justify the fact that they decided not to impute any value for "p < .05" but did so for p-values indicated as below .001 or .005. I believe it would be totally acceptable to impute values of .025 for cases where the p-value for a test was only reported as "p < .05." It would be the same method as for "p < .001" and "p < .005" where a p-value that was halfway between the threshold indicated and 0 was imputed. In addition, if "p > .05" was reported for a nonsignificant test, this could be imputed as a p-value of exactly .05. Methods like those I describe have been used for past z-curve analyses (see pp. 66-67 of the supplemental file for Salfate et al., 2025).

Response

The decision to impute only imprecisely reported small *p*-values (e.g., $p < .001$, $p < .005$) was based on the fact that these values are likely to reflect studies with very high power. Consequently, imputing them as $p = .0001$ or $p = .0005$ has a negligible impact on the *z*-curve results, as demonstrated in Sensitivity Analysis 1. We did not initially impute *p*-values reported as $p < .05$ or $p > .05$ because these categories encompass a much wider range of possible values, making any imputation more arbitrary and more likely to influence the *z*-curve estimates. Researchers can use p < .05 for a p-value of 0.0001, or for a p-value of 0.049, but the average expected power for both tests differs substantially. Imputing p = 0.025 for all these tests adds a very strong assumption into the results, which we do not feel comfortable with. At the same time, only 13 out of 269 tests reported imprecise p < .05 results, so the impact of any decision does not meaningfully change our results.

Following the reviewer's recommendation, we conducted three additional sensitivity analyses. Sensitivity Analysis 2 included the *p*-values from the primary *z*-curve analysis (269 *p*-values) plus *p*-values reported as $p < .05$, which were imputed as .025 (13 *p*-values). Sensitivity Analysis 3 included the *p*-values from the primary *z*-curve analysis (269 *p*-values) plus *p*-values reported as $p > .05$, which were imputed as .05 (13 *p*-values), where p = 0.05 is chosen as the most conservative value (i.e., it leads to the highest average power). Sensitivity Analysis 4 included the *p*-values from the primary *z*-curve analysis (269 *p*-values) plus *p*-values reported as $p < .05$ and $p > .05$ (26 *p*-values), which were imputed as .025 and .05, respectively. None of these additional sensitivity analyses yielded materially different estimates.

The results of these sensitivity analyses are provided as a Rmarkdown file (*zcurve_sensitivity_analysis.Rmd*) and a rendered document (*zcurve_sensitivity_analysis.docx*) at https://osf.io/d7wyc/.

We have also reported these sensitivity analyses in the Results section:

"In addition to the sensitivity analysis excluding these p-values, further sensitivity analyses were conducted at the reviewer's request. These included: (2) a sensitivity analysis in which 13 p-values reported as p < 0.05 that could not

be recomputed and were not included in the primary z-curve were imputed as 0.25; (3) a sensitivity analysis in which 13 p-values reported as $p > 0.05$ that could not be recomputed and were not included in the primary z-curve were imputed as 0.5; and (4) a sensitivity analysis combining both sets of imputed p-values. All four sensitivity analyses returned results consistent with those of the primary z-curve. The results of the sensitivity analyses can be found at https://osf.io/d7wyc/."

Comment 4

Search for studies for inclusion. The authors write on page 10, "We followed the protocol described in Murphy et al. [18] to select studies, but we deviated on two points from the original protocol." This is extremely unclear and seems to be the only information included on how the authors found the 350 studies considered for inclusion. Much more detail about how they searched the 10 journals for studies to consider for inclusion in this paper should be reported.

Response

Earlier in the paragraph, we specified that studies had to be applied within the sub-disciplines of physiology, sports performance, physical activity, injury prevention, and psychology, be confirmatory in the sense that they tested a hypothesis, and use an experimental or quasi-experimental design. In response to your suggestion, we have now added additional details regarding the journals considered, the number of studies sampled from each journal, and the requirement that selected studies employed an F-test or t-test.

The revised paragraph now reads as follows:

"The 350 studies were sampled from 10 journals ranked in quartile 1 according to www.scimagojr.com (as of 13th September 2022). The list of journals, along with the number of studies sampled from each, is depicted in Figure 3. All studies were published between 2024 and 2018. We started at a given issue and worked backwards. The study selection protocol was based on the Proposal of a Selection Protocol for Replication of Studies in Sports and Exercise Science (Murphy et al. 2022). First, only applied sport and exercise science studies (studying changes in human performance in response to physical activity, exercise, and sport) in the sub-disciplines of physiology, sports performance, physical activity, injury prevention, and psychology were considered. Second, only confirmatory studies that tested a hypothesis with an experimental (randomized controlled trials) or quasi-experimental design (non-randomized controlled trials) were included. Third, studies had to use an F-test (i.e., ANOVA) or t-test as an inferential test to evaluate the hypothesis; studies that employed correlations, mixed models or Bayesian statistics were excluded. We followed the protocol described in Murphy et al. (2022) to select studies, with two deviations from the original protocol. First, whereas the original protocol only selected studies that reported a statistically significant main effect, we considered both statistically significant and non-significant effects. This is because the z-

curve analysis uses both significant and non-significant p-values. Second, we also considered interaction effects, which were not considered in the original selection protocol."

Comment 5

Suggestions for improving sports psychology research. The recommendations for improving sports psychology research in the Discussion are good, as far as they go. However, in addition to a priori power analyses and registered reports, there are other remedies that bear discussing that the authors do not mention. These include: pre-registering methods, hypotheses, and analyses; making data openly available; journals being more open to accepting exploratory research; and a wider acceptance of replication studies. Given that the authors pre-registered their own work and made their own data openly available, I trust I do not have to elaborate much on any of these remedies. But I think they warrant inclusion in the discussion of remedies for the problems with research on sports psychology.

Response

We now mention these recommendations in the Discussion, which reads as follows: "Beyond these practices, the field should increasingly adopt higher standard for open data and code (Borg et al. 2020), transparently report exploratory research (Ditroilo et al. 2025), and collaborate with statisticians (Sainini et al. 2021). Together, these measures can substantially increase the reproducibility, replicability and therefore informational value of research in sports and exercise science."

Writing quality. I think the paper at times would be unclear and a bit hard to follow for readers unfamiliar with z-curve analysis and the logic behind it. Even I at times got slightly confused, and I have read and written up many z-curves myself. I think the paper could do with some revisions that elaborate on some major concepts important to this work.

Response

Thank you for pointing this out. We have reviewed the manuscript and made minor changes throughout to improve readability and clarity. We did not receive any specific suggestions on sections or concepts, and so hope that these changes reflect increased clarity in the document.

Comment 6:

There is something of an inconsistency in how the authors articulate the goal of their work. At first, they make it sound like publication bias is their main focus, but then do discuss p-hacking later on, even though they (erroneously, as I indicated above) assert that z-curve analysis cannot offer insights into the presence of p-hacking. I think the authors should make it clearer sooner and more consistently that their interest is both.

Response

Thank you for pointing this out. We have clarified in the manuscript that the primary goal of our work is to assess selection bias using the z-curve method. We explicitly discuss that selection bias encompasses two major forms: publication bias and p-hacking. While the z-curve method cannot distinguish between these two forms, both contribute to an inflated discovery rate, and our discussion now consistently reflects that our interest lies in understanding the broader phenomenon of selection bias.

Comment 7

This sentence from page 10 is hard to follow: "Specifically, we assess the presence of publication bias and average power in a sample of 269 studies published across 10 quartile 1 applied sports and exercise science journals using a z-curve analysis of primary statistical results."

Response

Thanks for pointing out this issue. We have now revised the sentence, which now reads as follows:
"Specifically, we assess the presence of publication bias and average power in a sample of 269 studies published across ten applied sports and exercise science journals using a z-curve analysis of primary statistical results."

Comment 8 (Methods)

I don't really understand what the authors are saying they did on page 10 when they write, "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://eur02.safelinks.protection.outlook.com/?url=https%3A%2F%2Fosf.io%2Fmqbr2%2F&data=05%7C02%7Cc.mesquida.caldentey%40tue.nl%7C0e9dce0bce1046dc35a408de1043aacb%7Ccc7df24760ce4a0f9d75704cf60efc64%7C0%7C0%7C638966079997265917%7CUnknown%7CTWFpbGZsb3d8eyJFbXB0eU1hcGkiOnRydWUsIlYiOiIwLjAuMDAwMCIsIlAiOiJXaW4zMiIsIkFOIjoiTWFpbCIsIldUIjoyfQ%3D%3D%7C0%7C%7C%7C&sdata=htcDryZwnr0tnao3eQDPjsq6MHbrbjhmtYNoPytP2Kc%3D&reserved=0). 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." I do not understand this precision analysis, and the link to the OSF provided does not give me permission to look in detail at what they wrote there. I do not think the authors should assume the reader will be familiar with such material, and so should elaborate on what this is more clearly.

Response

Thanks for pointing out the lack of clarity regarding the precision analysis. A precision analysis can estimate the number of observations required to determine an expected proportion within a specified margin of error. In a previous study assessing the prevalence and reproducibility of a priori power analyses, this approach was used to determine the number of studies needed to estimate the proportion of studies reporting an a priori power analysis within a margin of error. The same sample size and set of studies were used in the present study. We have now included a definition of precision analysis and make explicit that the same set of studies used in the previous study (assessing the prevalence and reproducibility of a priori power analyses) was used in the present study.

The revised paragraph now reads as follows:

“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 (Mesquida et al. 2025), which aimed to estimate the prevalence, reporting practices and reproducibility of a priori power analyses in sports and exercise science journals. Specifically, the precision analysis was conducted to estimate the number of studies required to detect an expected proportion of studies reporting an a priori power analysis (https://osf.io/mqbr2/). Assuming an expected proportion of 30% with a margin of error of 5%, the analysis indicated a required sample of 323 studies, which was subsequently rounded up to 350 studies. For convenience, the same set of 350 studies was used in both the previous study (Mesquida et al. 2025) and the present study (https://osf.io/d7wyc/).”

Mesquida, C., Murphy, J., Warne, J. & Lakens, D. (2025). Prevalence, reporting practices, and methodological quality of a priori power analyses in sports and exercise science research. SportRXiv.

Comment 9 (Methods)

This quote from page 12 is also worded in a way that is somewhat confusing: "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." This makes it sound like 57% of the 350 p-values were excluded, which obviously was not the case. So I think this should be reworded to be less confusing.

Response

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

Because I have cited some past work relevant to this topic, I just want to assert that none of the research I have cited in this review is my own. I am not a primary or co-author on any of the papers cited below.

References

Fremling, L., Strauel, C., & Bognar, E. (2025). Z-curve analysis of studies involving moderation published in leading health psychology journals. Health Psychology. https://eur02.safelinks.protection.outlook.com/?url=https%3A%2F%2Fdoi.org%2F10.1037%2Fhea0001534&data=05%7C02%7Cc.mesquida.caldentey%40tue.nl%7C0e9dce0bce1046dc35a408de1043aacb%7Ccc7df24760ce4a0f9d75704cf60efc64%7C0%7C0%7C638966079997280163%7CUnknown%7CTWFpbGZsb3d8eyJFbXB0eU1hcGkiOnRydWUsIlYiOiIwLjAuMDAwMCIsIlAiOiJXaW4zMiIsIkFOIjoiTWFpbCIsIldUIjoyfQ%3D%3D%7C0%7C%7C%7C&sdata=a6s8rmxv%2F%2BwIjD608o6gO3GNDvp7rS9nxyMfUubksA4%3D&reserved=0

Salfate, V. S., Spielmann, J., & Briley, D.A. (2024). Supporting the status quo is weakly associated with subjective well-being: A comparison of the palliative function of ideology across social status groups using a meta-analytic approach. [Supplemental material]. Psychological Bulletin, 150(11), 1318-1346. https://eur02.safelinks.protection.outlook.com/?url=https%3A%2F%2Fdoi.org%2F10.1037%2Fbul0000446&data=05%7C02%7Cc.mesquida.caldentey%40tue.nl%7C0e9dce0bce1046dc35a408de1043aacb%7Ccc7df24760ce4a0f9d75704cf60efc64%7C0%7C0%7C638966079997293941%7CUnknown%7CTWFpbGZsb3d8eyJFbXB0eU1hcGkiO

nRydWUsIlYiOiIwLjAuMDAwMCIsIlAiOiJXaW4zMiIsIkFOIjoiTWFpbCIsIldUIjoyfQ%3D%3D%7C0%7C%7C%7C&sdata=z8XO10cHUQZ5hsWJybHfhZKxPDedtgvn2LCjqlP2huQ%3D&reserved=0

Schimmack, U. (2012). The ironic effect of significant results on the credibility of multiple-study articles. Psychological Methods, 17(4), 551-566. https://eur02.safelinks.protection.outlook.com/?url=https%3A%2F%2Fdoi.org%2F10.1037%2Fa0029487&data=05%7C02%7Cc.mesquida.caldentey%40tue.nl%7C0e9dce0bce1046dc35a408de1043aacb%7Ccc7df24760ce4a0f9d75704cf60efc64%7C0%7C0%7C638966079997307206%7CUnknown%7CTWFpbGZsb3d8eyJFbXB0eU1hcGkiOnRydWUsIlYiOiIwLjAuMDAwMCIsIlAiOiJXaW4zMiIsIkFOIjoiTWFpbCIsIldUIjoyfQ%3D%3D%7C0%7C%7C%7C&sdata=oTL%2FSA2jkS5V4KANyAWk75PAoMsXkdkkwrmIaLkL790%3D&reserved=0

Editorial Comments

This manuscript has been well written, from an editorial perspective, and our submission guidelines have been well adhered to. I have no further requests for changes. Please ensure that Eline Ensinck has consented to being named in the Acknowledgements statement. Thank you.

Response

Eline has consented to being named in the Acknowledgements statement.