An Average Power Primer:

Clarifying Misconceptions about Average Power and Replicability

Maria D. Soto[1] and Ulrich Schimmack[1]

[1] Department of Psychology, University of Toronto

Submitted to *Meta-Psychology*. Participate in open peer review by sending an email to open.peer.reviewer@gmail.com. The full editorial process of all articles under review at Meta-Psychology can be found following this link:

https://tinyurl.com/mp-submissions

You will find this preprint by searching for the first author's name.

Author Note

Ulrich Schimmack https://orcid.org/0000-0001-9456-5536. Maria D. Soto https://orcid.org/0000-0001-6825-3985. We are grateful for the financial support of this work by the Canadian Social Sciences and Humanities Research Council (SSHRC). This research project was supported by a standard research grant awarded to Ulrich Schimmack by the Canadian Social Sciences and Humanities Research Council. The Open Science Framework repository link is available at https://osf.io/7ck4v/. We have no known conflict of interest to disclose. Correspondence regarding this article should be addressed to ulrich.schimmack@utoronto.ca, 3359 Mississauga Road, Mississauga, Ontario, Canada L5L 1C6.

Abstract

The replication crisis heightened interest in methods for assessing the credibility of published research. One approach to evaluate published results is to estimate the average power of original studies based on observed data. Recent criticisms, however, have challenged the validity and usefulness of this approach, arguing that it involves a fundamental "ontological error," fails to predict replication outcomes accurately, and yields imprecise estimates. This article aims to address these critics. We clarify that using observed data to estimate true power is a standard inferential practice and does not constitute an ontological error. We argue that the primary purpose of average power estimation is not to predict the outcome of future replication studies, but the hypothetical outcome if original researchers had to replicate their studies with new samples. Lastly, we demonstrate that even when uncertainty is substantial, average power estimates provide valuable diagnostic information about the credibility of literatures, especially when selection for significance is present. An applied example using a Z-curve analysis of terror management research shows that seemingly strong evidence of over 800 significant results does not rule out the possibility that all results are false positives. We conclude that, despite limitations, average power estimation remains a valid and useful tool for evaluating the evidential value of published research.

Keywords: statistical power, replicability, meta-analysis, credibility, uncertainty

An Average Power Primer:

Clarifying Misconceptions about Average Power and Replicability

During the 2010s, the field of psychology became increasingly aware of the difficulty in replicating many published findings. The most notable demonstration of replication problems is the Reproducibility Project (Open Science Collaboration, 2015). In this project, 100 studies published in three journals were replicated as closely as possible and only 36% of the replication studies obtained a statistically significant result. Success rates were slightly higher in cognitive psychology (50%) than in social psychology (25%) (Open Science Collaboration, 2015). From then on, efforts have been directed to assess the replicability of published results in psychology. In some cases, even large-scale replication studies with thousands of participants failed to replicate results despite hundreds of significant results in peer-reviewed journals (Vohs et al., 2021). The underwhelming results from seemingly robust findings backed by dozens of significant results have led to some researchers to declare a replication crisis or a crisis of confidence in psychological science (Schimmack, 2021; Vazire, 2018).

In response to the replication crisis, psychologists started to examine scientific practices that may inflate the probability of publishing a false positive result (John et al., 2012). Statistical tools such as the Test of Excess Significance (Francis, 2014; Ioannidis & Trikalinos, 2007; Schimmack, 2012), p-curve (Simonsohn et al., 2014), and z-curve (Bartoš & Schimmack, 2022; Brunner & Schimmack, 2020) were developed to diagnose publication bias and questionable research practices, and to evaluate the credibility of published research after correcting for publication bias. These methods are based on solid mathematical foundations (see Brunner & Schimmack, 2020, for proofs), and have been validated with extensive simulation studies (Bartoš

& Schimmack, 2022). They also have shown predictive validity with actual replication studies (Sotola, 2023).

However, some critical articles have raised concerns about the validity and usefulness of retrospective estimation of average power (e.g., McShane et al., 2020; Pek et al., 2024), raising three major concerns: an alleged ontological error when applying statistical power concepts post-study for evaluation, limited predictive validity of average power for future replication outcomes, and high uncertainty in average power estimates.

In this article, we carefully address these criticisms against power analyses of published studies to evaluate their credibility or to predict the outcome of replication studies. First, we show that the ontological error argument is based on a misunderstanding on the use of statistical power to examine the evaluate completed studies. We clarify that the primary purpose of estimating average power is not to predict outcomes of future replication studies, but to evaluate the credibility of the published studies by estimating their true average power. That is, if researchers repeated their original studies exactly, using the same methods and sample sizes, how likely would they be to obtain significant results again in a new sample with new sampling error? Second, we argue that average power estimates and their corresponding confidence intervals can be used to assess the credibility of original results despite uncertainty. Lastly, we showcase through a case study how average power can serve to assess the evidential value of a literature with a high discovery rate. The observed estimates and uncertainty can serve to inform future replication efforts. These estimates of average power provide powerful (pun intended) information about significant results in psychology journals. Misleading claims about methods that estimate average power are therefore problematic because they may be used by researchers

who produced significant results with questionable research practices to dismiss evidence that their evidence has low credibility; that is, low power and a high false positive risk.

Ontological Confusion: Pre-Outcome and Post-Outcome Probabilities

Traditionally, statistical power is defined as the long-run probability that a study produces a statistically significant result, conditional on a hypothetical population effect size that is different from zero (Cohen, 1988). The conditioning on a non-zero effect size makes sense for a priori power analysis, but it cannot be applied to estimates of true power because some studies may have population effect sizes of zero. We therefore define power (ϵ) as the unconditional probability of producing a significant result (Bartoš & Schimmack, 2022). This definition includes studies where the null hypothesis is true. The probability of obtaining a significant result when the null hypothesis is true is equals alpha. Henceforth, we refer to unconditional power as power. We distinguish among three types of power, hypothetical power, $\epsilon_{\mathrm{H}} = f(\alpha, \mathrm{N}, \theta_{\mathrm{H}})$, observed power, $\hat{\epsilon} = f(\alpha, \widehat{\mathrm{N}}, \hat{\theta})$, and true power, $\epsilon_{\mathrm{T}} = f(\alpha, \mathrm{N}, \theta_{\mathrm{T}})$. We use the subscript " H " to denote a hypothetical population parameter that is neither true nor observed, and " T " to denote a true population parameter, (H = Hypothetical, T = True).

Hypothetical power (ϵ_{H}) describes a power estimate where θ_{H} is a speculation of the population effect size. Hypothetical power is often used to plan sample sizes during the pre-study phase using an assumed true population effect size. Observed power ($\hat{\epsilon}$) is estimated using the observed effect size and sampling error. The true power (ϵ_{T}) represents the unconditional probability of obtaining a significant result, based on the unknown true population effect size (θ_{T}).

We refrain from the use of terms such as a prior and post-hoc power because power calculations can be conducted before and after a study, and power calculations after a study can

use hypothetical values or observed data. We also do not discuss the problematic use of observed power to evaluate the results of a single study (Hoenig & Heisey, 2001). Our focus is on the use of observed average power from a set of studies as estimates of the true average power of these studies.

Pek et al. (2024) argue that it is a fatal ontological error to estimate the power of a study after the study has been completed, "Applying a probability over random data to fixed (observed) data is a fatal ontological error" (p. 5). We agree that an observed outcome (e.g., a significant p-value) is a fixed fact. However, inferential statistics routinely use observed data to estimate unknown population parameters. Average power calculations do not make claims about the probability of the conducted studies. For example, the original studies in the Reproducibility Project had 97 significant results, which is typical for psychology journals (Sterling, 1959; Sterling et al., 1995). Estimating the true power of these studies is not a prediction of the probability of the completed studies to produce significant results because the percentage of significant results is known to be 97%. The goal is not to assign a probability to the set of realized studies with a fixed outcome, but to estimate the expected success rate if the same study were repeated under identical conditions with a new random sample. This is conceptually analogous to estimating the true proportion of any binary outcome (heads vs. tails of a coin toss) on the basis of a sample of events (Brunner & Schimmack, 2020)

A fair coin has a 50% probability to obtain heads or tails. After tossing the coin, the outcome is either heads or tails. The true power parameter of the coin determines the long-run probability of that coin to produce head again in the next toss or the long-run frequency of heads and tails of the coin. We cannot infer this coin's long-run probability to draw heads (ϵ) from a single coin toss ($\hat{\epsilon}$), in the same manner that we cannot estimate the true population effect size

(θ) present in a study from the single observed effect size ($\hat{\theta}$). If we continue randomly selecting and tossing coins from different populations with different probabilities, we will have a series observed outcomes. Then, the observed rate of heads $\hat{P}(Heads)$ can be used to estimate the mean probability to draw heads if we were to toss all the coins again, the population mean true power ($\overline{\epsilon_T}$). Therefore, it is also possible to estimate the probability of drawing heads if we were to toss another coin from the exact same populations where we drew for the original coin set. The same logic can be applied to published research, the observed success rate of a set of studies $\hat{P}(Significant)$ can be used to estimate the population mean true power ($\overline{\epsilon_T}$), that is, the long-run probability of the studies to produce the observed outcome.

In summary, it is not an ontological error to infer power from the results of observed studies and to make predictions about the outcome of hypothetical or future studies drawn from this set of completed studies.

Average Power and Replicability

Another concern has been "average power cannot quantify the replicability of an effect" (Pek et al, 2024, p. 11). It is often assumed that the purpose of power estimation is to predict outcomes of real-world replication studies, which are often non-exact due to changes in context, samples, or procedures (McShane et al., 2020). However, this argument does not hold for hypothetical prospective replication studies that by definition are identical to the completed studies. We agree that it is not very valuable to predict the outcome of actual replication studies, especially when average power of original studies is low. In this case, it would be futile to replicate the original studies exactly with the same sample sizes. Rather, actual replication studies would need to increase sample sizes and might also improve methodological weaknesses of original studies (e.g., Doyen et al., 2012).

However, predicting outcomes of new replication studies is not the primary goal of estimating average power. The key question is what results one would expect in a hypothetical replication project where the original authors redo their studies exactly as they were done, but with a new sample. This information is particularly interesting when results are selected for significance. When selection for significance is present, statistical significance provides no information about the false positive risk (Sterling, 1959). Estimates of average power, however, make it possible to distinguish credible results that were driven by true effect sizes from sets of studies with low average power that may contain a large percentage of false positive results (Sorić, 1989). In short, average power estimation provides a diagnostic tool for evaluating the evidential value of sets of studies.

Actual replication studies can be used, however, to evaluate predictions based on average power. The original studies used for the Reproducibility Project had a 97% success rate (Open Science Collaboration, 2015). Taken at face value, this finding implies that the studies had an average power of 97%. However, the replication studies had only a success rate of 36%. This suggests that the replication studies had an average power of 36%.

The issue with the original estimate is that it does not take in consideration that just like average observed effect size can be biased by selective publication so can the observed success rate. It is imminent that selection bias is accounted for when estimating average power, as selection bias systematically inflates the observed success rate. Thus, we cannot use the observed success rate of a set of studies at face value. Actual replication study outcomes can be used, however, to evaluate predictions based on average power. Either the replication studies had similar sample sizes (Open Science Collaboration, 2015) or the observed effect sizes of a

replication studies can be used to compute the hypothetical power with the sample sizes of the original studies.

Average Power Estimates are Too Variable to be Useful

Pek et al. (2024) draw on McShane et al.'s (2020) simulation studies to argue that estimates of average power are too variable and imprecise to be useful. They cite McShane et al.'s (2020) article to support this claim, but even McShane et al. (2020) point out that there is no fixed amount of uncertainty that invalidates empirical results, "Although what constitutes a tolerable degree of variability varies by context, we view a 95% sampling distribution width of .2 as the worst tolerable for estimating average power" (p. 191).

We agree that uncertainty is a concern, especially in small samples or highly heterogeneous literatures. However, uncertainty does not negate the informational value of power estimates. For example, if an average power estimate ranges from 10% to 40%, we can safely conclude that the original studies are underpowered and that future studies need to increase sample sizes to avoid false negative results, even if there is uncertainty about the true power of the studies. From this finding, we can also infer that effect size estimates are likely inflated because the true population effect size would not produce a significant result. This conclusion is implied by the statistical fact that a p-value equal to alpha (.05) corresponds to 50% power. Thus, a study with 40% power requires an inflated effect size estimate to achieve significance.

The main problem with estimates of observed power is the fallacy to confuse observed and true parameters; and the term observed does not help. Point estimates of average power are not observations of the true average power. To avoid this confusion, all estimates of true average power should be accompanied by uncertainty indicators such as 95% confidence intervals. The

width of these intervals depends on the set of studies and some other factors. The estimation error shrinks as the number of studies increases, just like sampling error shrinks with the number of participants in a study. Uncertainty can sometimes be less than McShane et al.'s (2020) arbitrary criterion of 20 percentage points and sometimes more. Whether the results are useful depends on the research question and the actual width of a confidence interval. It is impossible to state categorically that they are always too wide to be useful (Pek et al., 2024).

In conclusion, uncertainty in estimates of average power can be estimated with confidence intervals and average power estimates can be useful even if there is uncertainty in the estimates. The following section will use terror management research to showcase how average power can be useful evaluate the credibility of completed studies.

Terror Management Meta-Analysis

Chen et al. (2025) published a meta-analysis of over 800 terror management studies, in which the authors cite a chapter by Pek et al. (2022) warning readers that "power estimates are subject to the usual theoretical objections to estimating power from a fixed sample of data" (Chen et al., 2025, p. 16). Terror management studies use primes of death to examine the effect of mortality salience on people's beliefs, attitudes, and behaviors. We are concerned that misleading claims like these can be used to dismiss the results of a z-curve analysis. As pointed out above, it is not an ontological error to estimate true power based on observed data and to wonder how many significant results terror management researchers would find if they redid their studies the same way with new samples.

We reproduce the Z-curve analysis reported in their article here and explain for readers unfamiliar with Z-curve how results of a Z-curve analysis can be used to evaluate the credibility of published results (see also Soto & Schimmack, 2024). The data used for this case study was

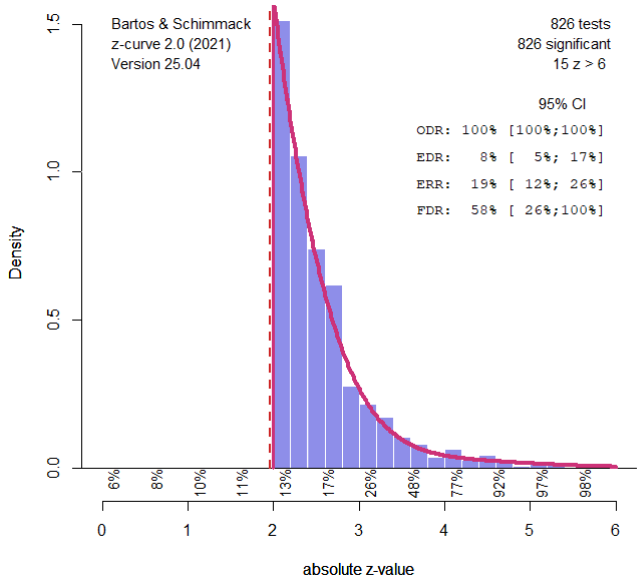
made publicly available by Chen et al. (2025) and the code to replicate the Z-curve results has been made available in an associated OSF repository, https://osf.io/7ck4v/. The re-analysis of the data was not preregistered.

Z-curve is a selection model that uses the distribution of z-scores to estimate average power. Unlike the meta-analytical approaches criticized by Pek et al. (2024) that estimate average power from a meta-analytic effect size, Z-curve estimates the population mean true power of a random sample of tests that have been selected for significance. By modelling the distribution of underlying the true power values across studies, it accommodated for heterogeneous true effect sizes without assuming the true distribution and selection bias (see Bartoš & Schimmack, 2022b and Brunner & Schimmack, 2020, for further details).

A Z-curve plot is helpful to evaluate the presence of selection bias. In this case, Chen et al. (2025) coded only significant results, which explains why there are no observations on the left side of the significance criterion z = 1.96 that corresponds to alpha = .05, two-tailed. However, it is well known that psychology journals mostly report significant results with success rates over 90% (Sterling, 1959). Thus, selection bias is a problem and observed power overestimates true power (Schimmack, 2012).

Figure 1

A Z-curve plot for Terror Management Studies



During visual inspection, a peak at z = 1.96, a steep decrease, and few studies with high z-scores (z > 4) imply that average power is low. Studies with 80% power would have a peak at 2.8. Z-curve provides an estimate of average power by fitting a model to this distribution and assuming a simple selection process, namely non-significant results are not reported, whereas significant results are reported. This model implies that there is no use of p-hacking to lower p-values to the level of significance. With this assumption it is possible to estimate the average power of all studies that were conducted, including the non-significant results that are missing in Figure 1. This estimate is sometimes called unconditional power because it does not condition on significance. Bartoš and Schimmack (2022) call this estimate of average power the Expected Discovery Rate (EDR), because it is an estimate of the actual percentage of significant results that is expected without selection for significance. The point-estimate is 8%, as reported in Chen et al. (2025).

The first observation is that the confidence interval is not symmetrical. The difference from 8% to 5% is only 3 percentage points, whereas the difference between 8% and 17% is 9

percentage points. The asymmetry is caused by a floor effect. Power is limited by alpha and cannot be less than 5%. This makes it problematic to express uncertainty by the width of the interval (McShane et al., 2020). Uncertainty also has different substantial implications depending on the direction of sampling error. In this example, the lower limit is theoretically more important than the upper limit. The lower limit includes alpha = 5%. This suggests, it is not possible to rule out the possibility that all these significant results were produced without a real effect; That is, the null-hypothesis is true in all 825 studies**.**

The EDR can also be used with the formula developed by Sorić (1989) to estimate the maximum false discovery rate that is compatible with the estimated EDR. Schimmack and Bartoš, (2023) refer to this estimate as the False Discovery Risk (FDR). The relationship between EDR and FDR is not linear. Even a low EDR of 17% implies that no more than 26% of the significant results can be false positive results. The range of possible FDR values is very wide and ranges from 26% to 100%, with a rather meaningless point estimate of 58%. Given the width of the confidence interval, it would be false to conclude that 58% of the results are false positives, but 42% are not false positives. The important empirical conclusion is that the data do not rule out the possibility that the entire literature rests on false positive results. The fact that the confidence interval is very wide does not undermine this conclusion because the burden of proof is on researchers who want to provide evidence for their theory. We might say that an FDR of 100% creates reasonable doubt about the credibility of the evidence. To convince skeptics, we might demand a much lower FDR.

The EDR estimate is based on assumptions about the selection process that may be false. If researchers used questionable practices to get significance in their studies, they need to test fewer studies. Moreover, questionable research practices can inflate the false positive risk

considerably (Simmons et al., 2011). Thus, this scenario also raises concerns about the credibility of the published results.

Alternatively, it is possible to focus on the average power of the published studies with significant results. This is also the aim of p-curve (PCURVE). This estimate is sometimes called conditional power, average power after selection for significance or the Expected Replicability Rate (ERR) as described in Z-curve (Brunner & Schimmack, 2020). It is an estimate of the population's long-run probability of producing a significant result again if another hypothetical study were conducted under the exact same conditions. Best understood as the expected success rate if we had a time machine and asked the original researchers to redo their studies with new samples and without the use of questionable research practices.

The ERR is 19% with a confidence interval ranging from 12% to 26%. Thus, the confidence interval even meets McShane et al.'s (2020) definition of precision that deviations from the point estimate should be less than 10 percentage points. This further affirms that Pek et al.'s (2024) claims that these estimates are always imprecise is inaccurate.

The confidence interval of the ERR does not include a value of 5%. Thus, the ERR rejects the pessimistic conclusion based on the EDR that all studies could be false positives. However, the upper limit of the ERR also shows that average power is low and that replication studies with the same sample sizes are more likely to produce non-significant results than significant ones. This is valuable information because it means that future studies need to use much larger samples to provide meaningful tests of the theory, even if average power does not predict the outcome of these new studies. If it could do so, we would not need to conduct costly actual replication studies.

A higher ERR than the EDR implies that studies vary in power. That is, if all studies had an average power of 10%, selection for significance could not select for more powerful studies. Heterogeneity in power also implies that some studies have a higher chance of producing a significant result in an actual replication study than others. To capitalize on this heterogeneity, it is possible to lower the criterion for significance to reduce the false positive risk (Soto & Schimmack, 2024).

We used *Zing* for the following section rather than the *zcurve* package (Bartoš & Schimmack, 2020) because *Zing* has some new developmental features that are not yet implemented in the current version of the *zcurve* package. *Zing* provides some information about heterogeneity in power because it computes the local power for subsets of studies with different z-scores. These estimates are shown below the x-axis. Studies with non-significant results are expected to have very low power (< 11%). However, even studies with z-scores between 2 and 2.5 have only 13% power. Average power estimates reach acceptable levels of power with z-scores greater than 4. There are also 15 studies with z-scores greater than 6 that practically have 100% power. Terror-management researchers might want to focus on these studies to see whether there are moderating factors that separate these studies from other studies.

According to Pek et al. (2024), Chen et al.'s (2025) z-curve analysis of terror management studies should be ignored because it makes an ontological error, does not predict outcomes of actual replication studies, and produces imprecise estimates. We provided strong counterarguments to these three claims. First, we realize that the published studies have a 100% success rate and that it is nonsensical to make probability statements about the outcome of these studies. However, we can make probability statements about researchers' ability to hypothetically replicate their findings under the same conditions as the original studies with the

same sample sizes, but with a new sample and following exactly the same statistical procedure that they used the first time. Our results show that researchers will not be able to produce significant results again in most studies. In fact, it would be surprising if they can produce significant results again in more than 50% of these studies. These results are helpful to evaluate the existing literature and raise concerns about the terror management literature.

While we do not agree with Chen et al. (2025) that "there must be some nonzero underlying effects in the studies we examined," we agree with their conclusion that "overall, the heterogeneous set of significant findings in TMT likely includes either inflated estimates or false positives, due to a combination of insufficient statistical power and publication bias" (p. 18). The real fatal error would be to ignore these findings and to assume that a standard effect-size meta-analysis that does not take publication bias into account provides credible evidence about this literature.

Conclusion

In this commentary, we addressed three criticism of statistical methods that estimate average power to assess the credibility of published results. First, we showed that the ontological error argument is invalid because it confuses observed power and true power. The observed power of a set of completed studies is not a probability. Maybe the term "observed" is misleading. We do not observe the true average power. What we observe are biased estimates of population effect sizes and statistical power. When these observed data are used to estimate true power, true power remains a probability because we did not observe the true power. We merely have an estimate of true power based on observed data. This distinction between estimates and population parameters is often blurred, but crucial. Once we make a clear distinction, the ontological error argument vanishes.

Next, we pointed out that true power is related to the outcome of exact replication studies with the same sample sizes as original studies, but not inexact replication studies with larger sample sizes. The main focus of estimating average power is not to predict the outcome of future studies, but to evaluate the credibility of the studies that were conducted. Average power is an estimate of the success rate of hypothetical exact replication studies that only differ in sampling error. Their ability to do so relies entirely on their correspondence with true power. Thus, arguments that replication studies are never exact are irrelevant for the use of average power to evaluate published studies. This conclusion is opposite to Pek et al.'s (2024) claim that average power may have some use as a tool to plan future studies but should never be used to evaluate completed studies. "Power should not be used to evaluate the results of completed studies from an N–P perspective, because to do so is based on flawed logic" (p. 4). We argue that this is the most important use of average power, especially when publication bias renders statistical significance meaningless and inflates effect size estimates.

Lastly, we addressed the argument against average power estimates based on an arbitrary criterion that estimates of average true power have to be within 10 percentage points of the true average power to be useful. Even McShane et al. (2020) pointed out that in some research context larger confidence intervals can still provide useful information. We, further, demonstrated this with an estimate of the average power of terror management studies. The width of the confidence intervals is a function of the number of studies and the usefulness of estimates must be evaluated on a case-by-case basis. An issue with uncertainty of point estimates, arises when the point estimates are falsely interpreted as the true population parameters. A point estimate of a 58% false positive risk should not be interpreted as evidence that 42% of all studies are not false positives. For the FDR, the upper limit of the 95%CI is

important even if the interval is wide. A wide confidence interval that includes a 100% false positive risk is informative because it undermines claims that a large number of significant results can only be obtained with a real effect. A credible literature would have a low FDR indicating that most published significant results are correct rejections of the null hypothesis. For example, in medical research and emotion research the upper FDR limit was 21% and 30%, respectively, suggesting credible evidence that most published significant results are correct rejections of the null hypothesis (Schimmack & Bartoš, 2023; Soto & Schimmack, 2024). Furthermore, in both instances it was possible to achieve an upper bound for the false positive risk below 5%, once one lowered alpha to .01.

Z-curve is not perfect, and it is still developing. Other researchers are developing similar methods (van Zwet et al., 2023). More research on the method and validation with real data is welcome. All this work will advance our ability to evaluate published results. Although Z-curve has only recently been published in a peer-reviewed journal with open peer-review (Brunner & Schimmack, 2020), it is already being used to probe the credibility of published results in several fields with mixed results. Some authors have documented statistical evidence of publication bias in dishonest research (Bartoš, 2024), gambling intervention tools (McAuliffe et al., 2021), system justification theory (SJT) (Sotola & Credé, 2022) and Construal Level Theory (CLT) (Maier et al., 2022). While others have document little to no evidence of publication bias in technology education (Buckley et al., 2023), social media use and self-esteem (van Anen, 2022), and, mental fatigue and exercise (Holgado et al., 2023). Z-curve analyses have been used to assess the credibility of results published in journals such as *Cognition & Emotion* and *Emotion* (Soto & Schimmack, 2024), *Journal of Sports Science* (Mesquida et al., 2023), organizational

journals (Crede & Sotola, 2024), personality journals (Sotola & Credé, 2023) and, there is an incoming review of the 10 most impactful journals in occupational health (Veen et al., 2024).

Furthermore, the method has been previously validated through simulation studies (Bartoš & Schimmack, 2022; Brunner & Schimmack, 2020). Comparisons between Z-curve estimates and large-scale replication efforts have shown that average power can be a valid predictor of actual replication outcomes (Sotola, 2023).

In this commentary, we addressed fundamental objections to the use of these methods and showed that they are based on misunderstandings of average power and replicability. This article should help researchers who are using Z-curve, reviewers, and readers of Z-curve studies to understand the method and its limitations, and to respond to false criticisms of the method.

References

Bartoš, F. (2024). The Untrustworthy Evidence in Dishonesty Research. *Meta-Psychology*, *8*. https://doi.org/10.15626/MP.2023.3987

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

Bartoš F, Schimmack U (2020). "zcurve: An R Package for Fitting Z-curves." R package version 2.4.2, https://CRAN.R-project.org/package=zcurve

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

Buckley, J., Hyland, T., & Seery, N. (2023). Estimating the replicability of technology education research. *International Journal of Technology and Design Education*, *33*(4), 1243–1264. https://doi.org/10.1007/s10798-022-09787-6

Chen, L., Benjamin, R., Guo, Y., Lai, A., & Heine, S. J. (2025). Managing the terror of publication bias: A systematic review of the mortality salience hypothesis. *Journal of Personality and Social Psychology*. https://doi.org/10.1037/pspa0000438

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). L. Erlbaum Associates.

Crede, M., & Sotola, L. K. (2024). All is well that replicates well: The replicability of reported moderation and interaction effects in leading organizational sciences journals. *The Journal of Applied Psychology*, *109*(10), 1659–1667. https://doi.org/10.1037/apl0001197

Doyen, S., Klein, O., Pichon, C.-L., & Cleeremans, A. (2012). Behavioral Priming: It's All in the Mind, but Whose Mind? *PLOS ONE*, *7*(1), e29081. https://doi.org/10.1371/journal.pone.0029081

Francis, G. (2014). The frequency of excess success for articles in Psychological Science. *Psychonomic Bulletin & Review*, *21*(5), 1180–1187. https://doi.org/10.3758/s13423-014-0601-x

Hoenig, J. M., & Heisey, D. M. (2001). The Abuse of Power: The Pervasive Fallacy of Power Calculations for Data Analysis. *The American Statistician*, *55*(1), 19–24. https://doi.org/10.1198/000313001300339897

Holgado, D., Mesquida, C., & Román-Caballero, R. (2023). Assessing the Evidential Value of Mental Fatigue and Exercise Research. *Sports Medicine*, *53*(12), 2293–2307. https://doi.org/10.1007/s40279-023-01926-w

Ioannidis, J. P. A., & Trikalinos, T. A. (2007). The appropriateness of asymmetry tests for publication bias in meta-analyses: A large survey. *CMAJ: Canadian Medical Association Journal = Journal de l'Association Medicale Canadienne*, *176*(8), 1091–1096. https://doi.org/10.1503/cmaj.060410

John, L., Loewenstein, G., & Prelec, D. (2012). Measuring the Prevalence of Questionable Research Practices With Incentives for Truth Telling. *Psychological Science*, *23*(5), 524–532. https://doi.org/10.1177/0956797611430953

Maier, M., Bartoš, F., Oh, M., Wagenmakers, E.-J., Shanks, D., & Harris, A. (2022). *Adjusting for Publication Bias Reveals That Evidence for and Size of Construal Level Theory Effects is Substantially Overestimated*. OSF. https://doi.org/10.31234/osf.io/r8nyu

McAuliffe, W. H. B., Edson, T. C., Louderback, E. R., LaRaja, A., & LaPlante, D. A. (2021). Responsible product design to mitigate excessive gambling: A scoping review and z-curve analysis of replicability. *PLOS ONE*, *16*(4), e0249926. https://doi.org/10.1371/journal.pone.0249926

McShane, B. B., Böckenholt, U., & Hansen, K. T. (2020). Average Power: A Cautionary Note. *Advances in Methods and Practices in Psychological Science*, *3*(2), 185–199. https://doi.org/10.1177/2515245920902370

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

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

Pek, J., Hoisington-Shaw, K. J., & Wegener, D. T. (2022). Avoiding Questionable Research Practices Surrounding Statistical Power Analysis. In W. O'Donohue, A. Masuda, & S. Lilienfeld (Eds.), *Avoiding Questionable Research Practices in Applied Psychology* (pp. 243–267). Springer International Publishing. https://doi.org/10.1007/978-3-031-04968-2_11

Pek, J., Hoisington-Shaw, K. J., & Wegener, D. T. (2024). Uses of Uncertain Statistical Power: Designing Future Studies, Not Evaluating Completed Studies. *Psychological Methods*. https://doi.org/10.1037/met0000577

Schimmack, U. (2012). The Ironic Effect of Significant Results on the Credibility of Multiple-Study Articles. *Psychological Methods*, *17*(4), 551–566. https://doi.org/10.1037/a0029487

Schimmack, U. (2021). The Validation Crisis in Psychology. *Meta-Psychology*, *5*. https://doi.org/10.15626/MP.2019.1645

Schimmack, U., & Bartoš, F. (2023). Estimating the false discovery risk of (randomized) clinical trials in medical journals based on published p-values. *PLOS ONE*, *18*(8), e0290084. https://doi.org/10.1371/journal.pone.0290084

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. (2014). P-curve: A key to the file-drawer. *Journal of Experimental Psychology. General*, *143*(2), 534–547. https://doi.org/10.1037/a0033242

Soto, M. D., & Schimmack, U. (2024). Credibility of results in emotion science: A Z-curve analysis of results in the journals Cognition & Emotion and Emotion. *Cognition and Emotion*, *0*, 1–17. https://doi.org/10.1080/02699931.2024.2443016

Sotola, L. (2023). How Can I Study from Below, that which Is Above? : Comparing Replicability Estimated by Z-Curve to Real Large-Scale Replication Attempts. *Meta-Psychology*, *7*. https://doi.org/10.15626/MP.2022.3299

Sotola, L. K., & Credé, M. (2022). On the predicted replicability of two decades of experimental research on system justification: A Z-curve analysis. *European Journal of Social Psychology*, *52*(5–6), 895–909. https://doi.org/10.1002/ejsp.2858

Sotola, L. K., & Credé, M. (2023). Estimating the replicability of statistically significant moderation effects in personality research using z-curve analysis. *Journal of Research in Personality*, *107*, 104435. https://doi.org/10.1016/j.jrp.2023.104435

Sterling, T. D. (1959). Publication Decisions and their Possible Effects on Inferences Drawn from Tests of Significance—Or Vice Versa. *Journal of the American Statistical Association*, *54*(285), 30–34. https://doi.org/10.1080/01621459.1959.10501497

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

van Anen, A. (2022). *How strong is our evidence? Evidential value and publication bias in research on social media use and self-esteem* [Master's thesis]. Tilburg University. http://arno.uvt.nl/show.cgi?fid=158963

van Zwet, E., Gelman, A., Greenland, S., Imbens, G., Schwab, S., & Goodman, S. N. (2023). A New Look at P Values for Randomized Clinical Trials. *NEJM Evidence*, *3*(1), EVIDoa2300003. https://doi.org/10.1056/EVIDoa2300003

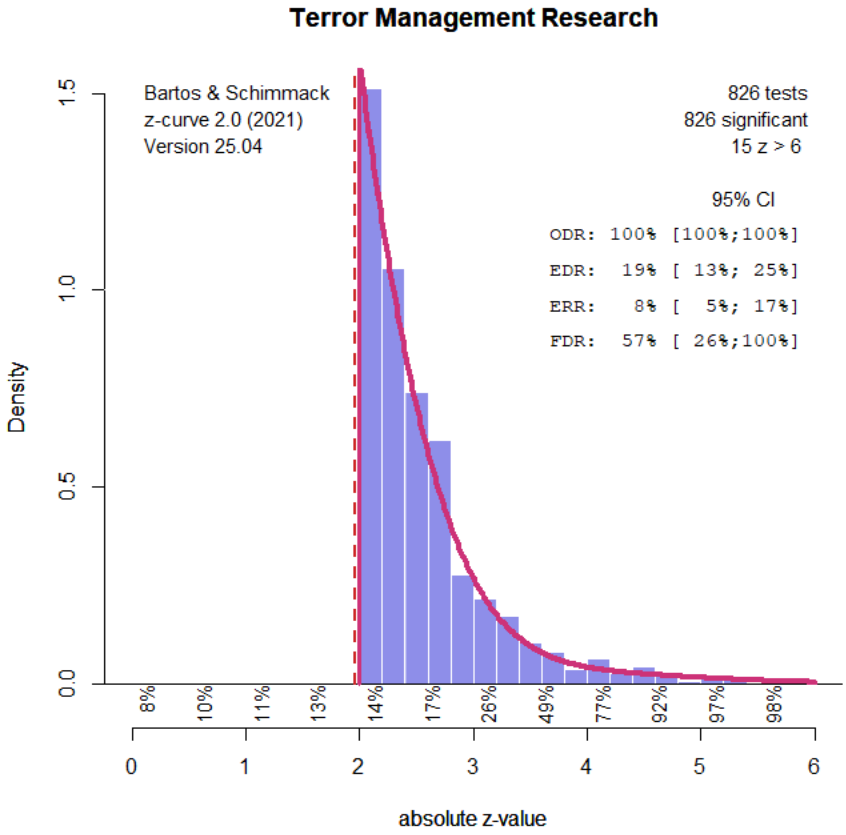
Vazire, S. (2018). Implications of the Credibility Revolution for Productivity, Creativity, and Progress. *Perspectives on Psychological Science*, *13*(4), 411–417. https://doi.org/10.1177/1745691617751884

Veen, M. van, Bartoš, F., Sarafoglou, A., Schelvis, R., Bouter, L., & Coenen, P. (2024). *Are there indications of publication bias in occupational health research? An examination of the literature and suggestions for future improvements*. https://doi.org/10.17605/OSF.IO/WFG7C

Vohs, K. D., Schmeichel, B. J., Lohmann, S., Gronau, Q. F., Finley, A. J., Ainsworth, S. E., Alquist, J. L., Baker, M. D., Brizi, A., Bunyi, A., Butschek, G. J., Campbell, C., Capaldi, J., Cau, C., Chambers, H., Chatzisarantis, N. L. D., Christensen, W. J., Clay, S. L., Curtis, J., … Albarracín, D. (2021). A Multisite Preregistered Paradigmatic Test of the Ego-Depletion Effect. *Psychological Science*, *32*(10), 1566–1581. https://doi.org/10.1177/0956797621989733

Figure 1

A Z-curve plot for Terror Management Studies



Note. ODR = Observed Discovery Rate, EDR = Expected Discovery Rate, ERR = Expected Replication Rate, FDR = False Discovery Risk.