Meta-Psychology, 2022, vol 6, MP.2021.2720
https://doi.org/10.15626/MP.2021.2720
Article type: Original Article
Published under the CC-BY4.0 license

Open data: Not Applicable
Open materials: Yes
Open and reproducible analysis: Yes
Open reviews and editorial process: Yes
Preregistration: No

Edited by: Daniël Lakens
Reviewed by: Joseph Hilgard, Erik van Zwet
Analysis reproduced by: André Kalmendal
All supplementary files can be accessed at OSF:
https://doi.org/10.17605/OSF.IO/V2KW5

Z-curve 2.0:
Estimating Replication and Discovery Rates

František Bartoš
University of Amsterdam; Faculty of Arts, Charles University

Ulrich Schimmack
University of Toronto, Mississauga

Abstract

Selection for statistical significance is a well-known factor that distorts the published literature and challenges the cumulative progress in science. Recent replication failures have fueled concerns that many published results are false-positives. Brunner and Schimmack (2020) developed z-curve, a method for estimating the expected replication rate (ERR) – the predicted success rate of exact replication studies based on the mean power after selection for significance. This article introduces an extension of this method, z-curve 2.0. The main extension is an estimate of the expected discovery rate (EDR) – the estimate of a proportion that the reported statistically significant results constitute from all conducted statistical tests. This information can be used to detect and quantify the amount of selection bias by comparing the EDR to the observed discovery rate (ODR; observed proportion of statistically significant results). In addition, we examined the performance of bootstrapped confidence intervals in simulation studies. Based on these results, we created robust confidence intervals with good coverage across a wide range of scenarios to provide information about the uncertainty in EDR and ERR estimates. We implemented the method in the `zcurve` R package (Bartoš & Schimmack, 2020).

Keywords: Publication Bias, Selection Bias, Expected Replication Rate, Expected Discovery Rate, File-Drawer, Power, Mixture Models

It has been known for decades that the published record in scientific journals is not representative of all studies that are conducted. For a number of reasons, most published studies are selected because they reported a theoretically interesting result that is statistically significant; $p < .05$ (Rosenthal & Gaito, 1964; Scheel et al., 2021; Sterling, 1959; Sterling et al., 1995). This selective publishing of statistically significant results introduces a bias in the published literature. At the very least, published effect sizes are inflated. In the most extreme cases, a false-positive result is supported by a large number of statistically significant results (Rosenthal, 1979).

Some sciences (e.g., experimental psychology) tried to reduce the risk of false-positive results by demanding replication studies in multiple-study articles (c.f., Wegner, 1992). However, internal replication studies provided a false sense of replicability because researchers used questionable research practices to produce successful internal replications (Francis, 2014; John et al., 2012; Schimmack, 2012). The pervasive presence of publication bias at least partially explains replication failures in social psychology (Open Science Collaboration, 2015; Pashler & Wagenmakers, 2012; Schimmack, 2020; Schimmack, 2012) medicine (Begley & Ellis, 2012; Prinz et al., 2011), and economics (Camerer et al., 2016; Chang & Li, 2015).

In meta-analyses, the problem of publication bias is usually addressed by one of the different methods for its detection and a subsequent adjustment of effect size estimates. However, many of them (Egger et al., 1997; Ioannidis & Trikalinos, 2007; Schimmack, 2012) perform poorly under conditions of heterogeneity (Renkewitz & Keiner, 2019), whereas others employ a meta-analytic model assuming that the studies are conducted on a single phenomenon (e.g., Hedges, 1992; Maier et al., 2022; Vevea & Hedges, 1995). Moreover, while the aforementioned methods test for publication bias (return a p-value or a Bayes factor), they usually do not provide a quantitative estimate of selection bias. An exception would be the publication probabilities/ratios estimates from selection models (e.g., Hedges, 1992). Maximum likelihood selection models work well when the distribution of effect sizes is consistent with model assumptions, but can be biased when the distribution when the actual distribution does not match the expected distribution (e.g., Brunner & Schimmack, 2020; Hedges, 1992; Vevea & Hedges, 1995). Brunner and Schimmack (2020) introduced a new method that does not require a priori assumption about the distribution of effect sizes. The z-curve method uses a finite mixture model to correct for selection bias. We extended z-curve to also provide information about the amount of selection bias. To distinguish between the new and old z-curve methods, we refer to the old z-curve as z-curve 1.0 and the new z-curve as z-curve 2.0. Z-curve 2.0 has been implemented in the open statistic program R as the `zcurve` package that can be downloaded from CRAN (Bartoš & Schimmack, 2020).

Before we introduce z-curve 2.0, we would like to introduce some key statistical terms. We assume that readers are familiar with the basic concepts of statistical significance testing; normal distribution, null-hypothesis, alpha, type-I error, and false-positive result (see Bartoš & Maier, In press, for discussion of some of those concepts and their relation).

Glossary

Power is defined as the long-run relative frequency of statistically significant results in a series of exact replication studies with the same sample size when the null-hypothesis is false. For example, in a study with two groups (n = 50), a population effect size of Cohen's d = 0.4 has 50.8% power to produce a statistically significant result. Thus, 100 replications of this study are expected to produce approximately 50 statistically significant results. The actual frequency will approach 50.8% as the study is repeated infinitely.

Unconditional power extends the concept of power to studies where the null-hypothesis is true. Typically, power is a conditional probability assuming a non-zero effect size (i.e., the null-hypothesis is false). However, the long-run relative frequency of statistically significant results is also known when the null-hypothesis is true. In this case, the long-run relative frequency is determined by the significance criterion, alpha. With alpha = 5%, we expect that 5 out of 100 studies will produce a statistically significant result. We use the term unconditional power to refer to the long-run frequency of statistically significant results without conditioning on a true effect. When the effect size is zero and alpha is 5%, unconditional power is 5%. As we only consider unconditional power in this article, we will use the term power to refer to unconditional power, just like Canadians use the term hockey to refer to ice hockey.

Mean (unconditional) power is a summary statistic of studies that vary in power. Mean power is simply the arithmetic mean of the power of individual studies. For example, two studies with power = 0.4 and power = 0.6, have a mean power of 0.5.

Discovery rate is a relative frequency of statistically significant results. Following Sorić (1989), we call statistically significant results discoveries. For example, if 100 studies produce 36 statistically significant results, the discovery rate is 36%. Importantly, the discovery rate does not distinguish between true or false discoveries. If only false-positive results were reported, the discovery rate would be 100%, but none of the discoveries would reflect a true effect (Rosenthal, 1979).

Selection bias is a process that favors the publication of statistically significant results. Consequently, the published literature has a higher percentage of statistically significant results than was among the actually conducted studies. It results from significance testing that creates two classes of studies separated by the significance criterion alpha. Those with a statistically significant result, $p < .05$, where the null-hypothesis is rejected, and those with a statistically non-significant result, where the null-hypothesis is not rejected, $p > .05$. Selection for statistical significance limits the population of all studies that were conducted to the population of studies with statistically significant results. For example, if two studies produce p-values of .20 and .01, only the study with the p-value .01 is retained. Selection bias is often called publication bias. Studies show that authors are more likely to submit findings for publication when the results are statistically significant (Franco et al., 2014)

Observed discovery rate (ODR) is the percentage of statistically significant results in an observed set of studies. For example, if 100 published studies have 80 statistically significant results, the observed discovery rate is 80%. The observed discovery rate is higher than the true discovery rate when selection bias is present.

Expected discovery rate (EDR) is the mean power before selection for significance; in other words, the mean power of all conducted studies with statistically significant and non-significant results. As power is the long-run relative frequency of statistically significant results, the mean power before selection for significance is the expected relative frequency of statistically significant results. As we call statistically significant results discoveries, we refer to the expected percentage of statistically significant results as the expected discovery rate. For example, if we have two studies with power of 0.05 and 0.95, we are expecting 1 statistically significant result and an EDR of 50%, (0.95 + 0.05)/2 = 0.50.

Expected replication rate (ERR) is the mean power after selection for significance, in other words, the mean power of only the statistically significant studies. Furthermore, since most people would declare a replication successful only if it produces a result in the same direction, we base ERR on the power to obtain a statistically significant result in the same direction. Using the prior example, we assume that the study with 5% power produced a statistically non-significant result and the study with 95% power produced a statistically significant result. In this case, we end up with only one statistically significant result with 95% power. Subsequently, the mean power after selection for significance is 95% (there is almost zero chance that a study with 95% power would produce replication with an outcome in the opposite direction). Based on this estimate, we would predict that 95% of exact replications of this study with the same sample size, and therefore with 95% power, will be statistically significant in the same direction.

As mean power after selection for significance predicts the relative frequency of statistically significant results in replication studies, we call it the expected replication rate. The ERR also corresponds to the “aggregate replication probability” discussed by Miller (2009).

Numerical Example

Before introducing the formal model, we illustrate the concepts with a fictional example. In the example, researchers test 100 true hypotheses with 100% power (i.e., every test of a true hypothesis produces $p < .05$) and 100 false hypotheses ($\mathcal{H}_0$ is true) with 5% power which is determined by alpha = .05. Consequently, the researchers obtain 100 true positive results and 5 false-positive results, for a total of 105 statistically significant results. The expected discovery rate is $(1 \times 100 + 0.05 \times 100)/(100 + 100) = 105/200 = 52.5\%$ which corresponds to the observed discovery rate when all conducted studies are reported.

So far, we have assumed that there is no selection bias. However, let us now assume that 50 of the 95 statistically non-significant results are not reported. In this case, the observed discovery rate increased from 105/200 to 105/150 = 70%. The discrepancy between the EDR, 52.5%, and the ODR, 70%, provides quantitative information about the amount of selection bias.

As shown, the EDR provides valuable information about the typical power of studies and about the presence of selection bias. However, it does not provide information about the replicability of the statistically significant results. The reason is that studies with higher power are more likely to produce a statistically significant result in replications (Brunner & Schimmack, 2020; Miller, 2009). The main purpose of z-curve 1.0 was to estimate the mean power after selection for significance to predict the outcome of exact replication studies. In the example, only 5 of the 100 false hypotheses were statistically significant. In contrast, all 100 tests of the true hypothesis were statistically significant. This means that the mean power after selection for significance is $(5 \times 0.025 + 100 \times 1)/(5 + 100) = 100.125/105 \approx 95.4\%$, which is the expected replication rate.

Formal Introduction

Unfortunately, there is no standard symbol for power, which is usually denoted as $1 - \beta$, with β being the probability of a type-II error. We propose to use epsilon, ϵ, as a Greek symbol for power because one Greek word for power starts with this letter ($\epsilon\xi o\nu\sigma\iota\alpha$). We further add subscript 1 or 2, depending on whether the direction of the outcome is relevant or not. Therefore, ϵ_2 denotes power of a study regardless of the direction of the outcome and ϵ_1 denotes power of a study in a specified direction.

The EDR,

$$\text{EDR} = \frac{\sum_{k=1}^{K} \epsilon_{2,k}}{K}, \tag{1}$$

is defined as the mean power (ϵ_2) of a set of K studies, independent on the outcome direction. Following Brunner and Schimmack (2020), the expected replication rate (ERR) is defined as the ratio of mean squared power and mean power of all studies, statistically significant and non-significant ones. We modify the definition here by taking the direction of the replication study into account. The mean square power in the nominator is used because we are computing the expected relative frequency of statistically significant studies produced by a set of already statistically significant studies – if a study produces a statistically significant result with probability equal to its power, the chance that the same study will again be significant is power squared. The mean power in the denominator is used because we are restricting our selection to only already statistically significant studies which are produced at the rate corresponding to their power (see also Miller, 2009). The ratio simplifies by omitting division by K in both the nominator and denominator to:

$$\text{ERR} = \frac{\sum_{k=1}^{K} \epsilon_{2,k} \times \epsilon_{1,k}}{\sum_{k=1}^{K} \epsilon_{2,k}}, \tag{2}$$

which can also be read as a weighted mean power, where each power is weighted by itself. The weights originate from the fact that studies with higher power are more likely to produce statistically significant results. The weighted mean power of all studies is therefore equal to the unweighted mean power of the studies selected for significance (ksig; cf. Brunner & Schimmack, 2020).

If we have a set of studies with the same power (e.g., set of exact replications with the same sample size) that test for an effect with a z-test, the p-values converted to z-statistics follow a normal distribution with mean μ_z and a standard deviation equal to 1. Using an alpha level α, the power is the tail area of a standard normal distribution (Φ) centered over a mean, (μ_z) on the left and right side of the z-statistics corresponding to alpha, -1.96 and 1.96 (with the usual alpha = .05),

$$\epsilon_{2,z} = 1 - \Phi(1.96 - \mu_z) + \Phi(-1.96 - \mu_z), \tag{3}$$

or the tail area on the right side of the z-statistics corresponding to alpha, when we are also considering whether the directionality of the effect,

$$\epsilon_{1,z} = 1 - \Phi(1.96 - \mu_z). \tag{4}$$

Two-sided p-values do not preserve the direction of the deviation from null and we cannot know whether a z-statistic comes from the lower or upper tail of the distribution. Therefore, we work with absolute values of z-statistics, changing their distribution from normal to folded normal distribution (Elandt, 1961; Leone et al., 1961).

Figure 1 illustrates the key concepts of z-curve with various examples. The first three density plots in the first row show the sampling distributions for studies with low ($\epsilon = 0.3$), medium ($\epsilon = 0.5$), and high ($\epsilon = 0.8$) power, respectively. The last density plots illustrate the distribution that is obtained for a mixture of studies with low, medium, and high power with equal frequency (33.3% each). It is noteworthy that all four density distributions have different shapes. While Figure 1 illustrates how differences in power produce differences in the shape of the distributions, z-curve works backward and uses the shape of the distribution to estimate power.

Although z-curve can be used to fit the distributions in the first row, we assume that the observed distribution of all z-statistics is distorted by the selection bias. Even if some statistically non-significant p-values are reported, their distribution is subject to unknown selection effects. Therefore, by default z-curve assumes that selection bias is present and uses only the distribution of statistically significant results. This changes the distributions of z-statistics to folded normal distributions that are truncated at the z-statistic corresponding to the significance criterion, which is typically $z = 1.96$ for $p = .05$ (two-tailed). The second row in Figure 1 shows these truncated folded normal distributions. Importantly, studies with different levels of power produce different distributions despite the truncation. The different shapes of truncated distributions make it possible to estimate power by fitting a model to the truncated distribution.

The third row of Figure 1 illustrates the EDR as a proportion of statistically significant studies from all conducted studies. We use Equation 3 to re-express EDR (Equation 1), which equals the mean unconditional power, of a set of K heterogeneous studies using the means of sampling distributions of their z-statistics, $\mu_{z,k}$,

$$\text{EDR} = \frac{\sum_{k=1}^{K} \epsilon_{2,z,k}}{K}. \tag{5}$$

Z-curve makes it possible to estimate the shape of the distribution in the region of statistically non-significant results on the basis of the observed distribution of statistically significant results. That is, after fitting a model to the grey area of the curve, it extrapolates the full distribution.

The fourth row of Figure 1 visualizes a distribution of expected z-statistics if the statistically significant studies were to be exactly replicated (not depicting the small proportion of results in the opposite direction than the original, significant, result). The full line highlights the portion of studies that would produce a statistically significant result, with the distribution of statistically non-significant studies drawn using the dashed line. An exact replication with the same sample size of the studies in the grey area in the second row is not expected to reproduce the truncated distribution again because truncation is a selection process. The replication distribution is not truncated and produces statistically signif-

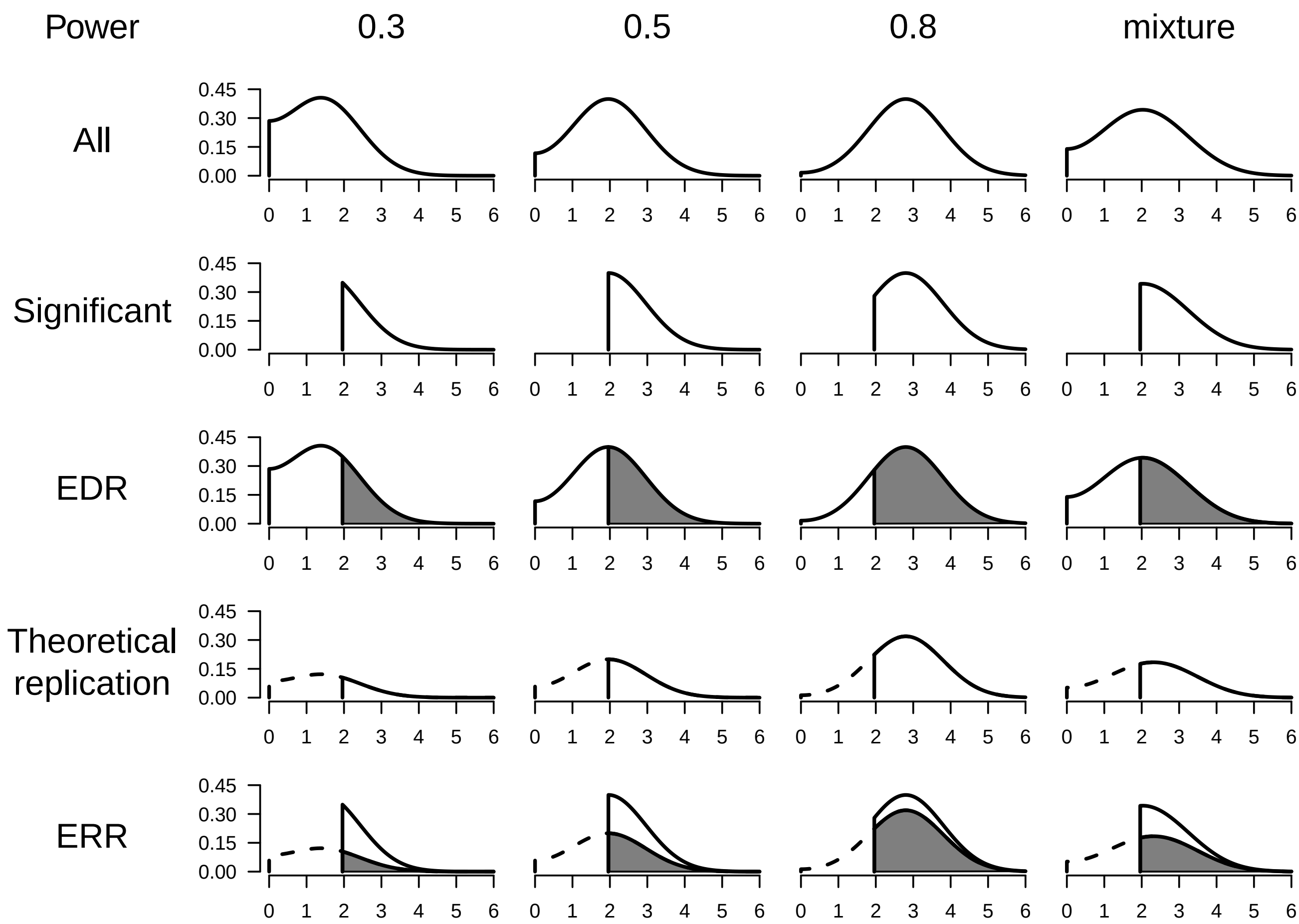

Figure 1. Density (y-axis) of z-statistics (x-axis) generated by studies with different powers (columns) across different stages of the publication process (rows). The first row shows a distribution of z-statistics from z-tests homogeneous in power (the first three columns) or by their mixture (the fourth column). The second row shows only statistically significant z-statistics. The third row visualizes EDR as a proportion of statistically significant z-statistics out of all z-statistics. The fourth row shows a distribution of z-statistics from exact replications of only the statistically significant studies (dashed line for non-significant replication studies). The fifth row visualizes ERR as a proportion of statistically significant exact replications out of statistically significant studies.

icant and non-significant results. By modeling the selection process, z-curve predicts the non-truncated distributions in the fourth row from the truncated distributions in the second row.

The fifth row of Figure 1 visualizes ERR as a proportion of statistically significant exact replications in the expected direction from a set of the previously statistically significant studies. The ERR (Equation 2) of a set of heterogeneous studies can be again re-expressed using Equations 3 and 4 with the means of sampling distributions of their z-statistics,

$$\mathrm{ERR} = \frac{\sum_{k=1}^{K} \epsilon_{2,z,k} \times \epsilon_{1,z,k}}{\sum_{k=1}^{K} \epsilon_{2,z,k}}. \tag{6}$$

Z-curve 2.0

Z-curve is a finite mixture model (Brunner & Schimmack, 2020). Finite mixture models leverage the fact that an observed distribution of statistically significant z-statistics is a mixture of K truncated folded normal distribution with means $\mu_{z,k}$ and standard deviations 1. Instead of trying to estimate $\mu_{z,k}$ of every single observed z-statistic, a finite mixture model approximates the observed distribution based on K studies with a smaller set of J truncated folded normal distributions, $f(z;\theta)$, with $J < K$ components,

$$f(z;\theta) = \sum_{j=1}^{J} \pi_j f_{j[a,b]}(z;\theta_j). \tag{7}$$

Each mixture component j approximates a proportion of π_j observed z-statistics with a probability density function, $f_{j[a,b]}$, of truncated folded normal distribution with parameters θ_j – a mean $\mu_{z,j}$ and standard deviation equal to 1. For example, while actual studies may vary in power from 40% to 60%, a mixture model may represent all of these studies with a single component with 50% power.

Z-curve 1.0 used three components with varying means. Extensive testing showed that varying means produced poor estimates of the EDR. Therefore, we switched to models with fixed means and increased the number of components to seven. The seven components are equally spaced by one standard deviation from $z = 0$ (power = alpha) to 6 (power ≈ 1). As power for z-statistics greater than 6 is essentially 1, it is not necessary to model the distribution of z-statistics greater than 6, and all z-statistics greater than 6 are assigned a power value of 1 (Brunner & Schimmack, 2020). The power values implied by the 7 components are 0.05, 0.17, 0.50, 0.85, 0.98, 0.999, 0.99997. We also tried a model with equal spacing of power, and we tried models with fewer or more components, but neither did improve performance in simulation studies. We use the model parameter estimates to compute the estimated EDR and ERR as the weighted average of seven truncated folded normal distributions centered over $z = 0$ to 6,

$$\widehat{\text{EDR}} \approx \sum_{j=1}^{J} \hat{\pi}_j \times \epsilon_{2,z,j}, \tag{8}$$

$$\widehat{\text{ERR}} \approx \frac{\sum_{j=1}^{J} \hat{\pi}_j \times \epsilon_{2,z,j} \times \epsilon_{1,z,j}}{\sum_{j=1}^{J} \hat{\pi}_j \times \epsilon_{2,z,j}}.$$

Curve Fitting

Z-curve 1.0 used an unorthodox approach to find the best fitting model that required fitting a truncated kernel-density distribution to the statistically significant z-statistics (Brunner & Schimmack, 2020). This is a non-trivial step that may produce some systematic bias in estimates. Z-curve 2.0 makes it possible to fit the model directly to the observed z-statistics using the well-established expectation maximization (EM) algorithm that is commonly used to fit mixture models (Dempster et al., 1977; Lee & Scott, 2012). Using the EM algorithm has the advantage that it is a well-validated method to fit mixture models. It is beyond the scope of this article to explain the mechanics of the EM algorithm (Bishop, 2006, e.g.,), but it is important to point out some of its potential limitations. The main limitation is that it may terminate the search for the best fit before the best fitting model has been found. In order to prevent this, we run 20 searches with randomly selected starting values and terminate the algorithm in the first 100 iterations, or if the criterion falls below 10¯3. We then select the outcome with the highest likelihood value and continue until 1000 iterations or a criterion value of 10¯5 is reached. To speed up the fitting process, we optimized the procedure using Rcpp (Eddelbuettel & François, 2011).

Information about point estimates should be accompanied by information about uncertainty whenever possible. The most common way to do so is by providing confidence intervals. We followed the common practice of using bootstrapping to obtain confidence intervals for mixture models (Ujeh et al., 2016). As bootstrapping is a resource-intensive process, we used 500 samples for the simulation studies. Users of the z-curve package can use more iterations to analyze actual data.

Simulations

Brunner and Schimmack (2020) compared several methods for estimating mean power and found that z-curve performed better than three competing methods. However, these simulations were limited to the estimation of the ERR. Here we present new simulation studies to examine the performance of z-curve as a method to estimate the EDR as well. One simulation directly simulated power distributions, the other one simulated t-tests. We report the detailed results of both simulation studies in a Supplement. For the sake of brevity, we focus on the simulation of t-tests because readers can more easily evaluate the realism of these simulations. Moreover, most tests in psychology are t-tests or F-tests and Brunner and Schimmack (2020) already showed that the numerator degrees of freedom of F-tests do not influence results. Thus, the results for t-tests can be generalized to F-tests and z-tests.

The simulation was a complex $4 \times 4 \times 4 \times 3 \times 3$ design with 576 cells. The first factor of the design that was manipulated was the mean effect size with Cohen's ds ranging from 0 to 0.6 (0, 0.2, 0.4, 0.6). The second factor in the design was heterogeneity in effect sizes was simulated with a normal distribution around the mean effect size with SDs ranging from 0 to 0.6 (0, 0.2, 0.4, 0.6). Preliminary analysis with skewed distributions showed no influence of skew. The third factor of the design was sample size for between-subject design with $N = 50$, 100, and 200. The fourth factor of the design was the percentage of true null-hypotheses that ranged from 0 to 60% (0%, 20%, 40%, 60%). The last factor of the design was the number of studies with sets of $k = 100$, 300, and 1,000 statistically significant studies.

Each cell of the design was run 100 times for a total of 57,600 simulations. For the main effects of this design there were $57{,}600/4 = 14{,}400$ or $57{,}600/3 = 19{,}200$

simulations. Even for two-way interaction effects, the number of simulations is sufficient, $57,600/16 = 3,600$. For higher interactions the design may be underpowered to detect smaller effects. Thus, our simulation study meets recommendations for sample sizes in simulation studies for main effects and two-way interactions, but not for more complex interaction effects (Morris et al., 2019). The code for the simulations is accessible at https://osf.io/r6ewt/.

Evaluation

For a comprehensive evaluation of z-curve 2.0 estimates, we report bias (i.e., mean distance between estimated and true values), root mean square error (RMSE; quantifying the error variance of the estimator), and confidence interval coverage (Morris et al., 2019). To check the performance of the z-curve across different simulation settings, we analyzed the results of the factorial design using analyses of variance (ANOVAs) for continuous measures and logistic regression for the evaluation of confidence intervals (0 = true value not in the interval, 1 = true value in the interval). The analysis scripts and results are accessible at https://osf.io/r6ewt/.

Results

We start with the ERR because it is essentially a conceptual replication study of Brunner and Schimmack (2020) simulation studies with z-curve 1.0.

ERR. Visual inspection of the z-curves ERR estimates plotted against the true ERR values did not show any pathological behavior due to the approximation by a finite mixture model (Figure 2).

Figure 2 shows that even with $k = 100$ studies, z-curve estimates are clustered close enough to the true values to provide useful predictions about the replicability of sets of studies. Overall bias was less than one percentage point, -0.88 ($SE_{MCMC} = 0.04$). This confirms that z-curve has high large-sample accuracy (Brunner & Schimmack, 2020). RMSE decreased from 5.14 ($SE_{MCMC} = 0.03$) percentage points with $k = 100$ to 2.21 ($SE_{MCMC} = 0.01$) percentage points with $k = 1,000$. Thus, even with relatively small sample sizes of 100 studies, z-curve can provide useful information about the ERR.

The Analysis of Variance (ANOVA) showed no statistically significant 5-way interaction or 4-way interactions. A strong three-way interaction was found for effect size, heterogeneity of effect sizes, and sample size, $z = 9.42$. Despite the high statistical significance, effect sizes were small. Out of the 36 cells of the 4 × 3 × 3 design, 32 cells showed less than one percentage point bias. Larger biases were found when effect sizes were large, heterogeneity was low, and sample sizes were small. The largest bias was found for Cohen's $d = 0.6$, SD = 0, and $N = 50$. In this condition, ERR was 4.41 ($SE_{MCMC} = 0.11$) percentage points lower than the true replication rate. The finding that z-curve performs worse with low heterogeneity replicates findings by Brunner and Schimmack (2020). One reason could be that a model with seven components can easily be biased when most parameters are zero. The fixed components may also create a problem when true power is between two fixed levels. Although a bias of 4 percentage points is not ideal, it also does not undermine the value of a model that has very little bias across a wide range of scenarios.

The number of studies had a two-way interaction with effect size, $z = 3.8$, but bias in the 12 cells of the 4 × 3 design was always less than 2 percentage points. Overall, these results confirm the fairly good large sample accuracy of the ERR estimates.

We used logistic regression to examine patterns in the coverage of the 95% confidence intervals. This time a statistically significant four-way interaction emerged for effect size, heterogeneity of effect sizes, sample size, and the percentage of true null-hypotheses, $z = 10.94$. Problems mirrored the results for bias. Coverage was low when there were no true null-hypotheses, no heterogeneity in effect sizes, large effects, and small sample sizes. Coverage was only 31.3% ($SE_{MCMC} = 2.68$) when the percentage of true $\mathcal{H}_0$ was 0, heterogeneity of effect sizes was 0, the effect size was Cohen's $d = 0.6$, and the sample size was $N = 50$.

In statistics, it is common to replace confidence intervals that fail to show adequate coverage with confidence intervals that provide good coverage with real data; these confidence intervals are often called robust confidence intervals (Royall, 1996). We suspected that low coverage was related to systematic bias. When confidence intervals are drawn around systematically biased estimates, they are likely to miss the true effect size by the amount of systematic bias, when sampling error pushes estimates in the same direction as the systematic bias. To increase coverage, it is therefore necessary to take systematic bias into account. We created robust confidence intervals by adding three percentage points on each side. This is very conservative because the bias analysis would suggest that only adjustment in one direction is needed.

The logistic regression analysis still showed some statistically significant variation in coverage. The most notable finding was a 2-way interaction for effect size and sample size, $z = 4.68$. However, coverage was at 95% or higher for all 12 cells of the design. Further inspection showed that the main problem remained scenarios

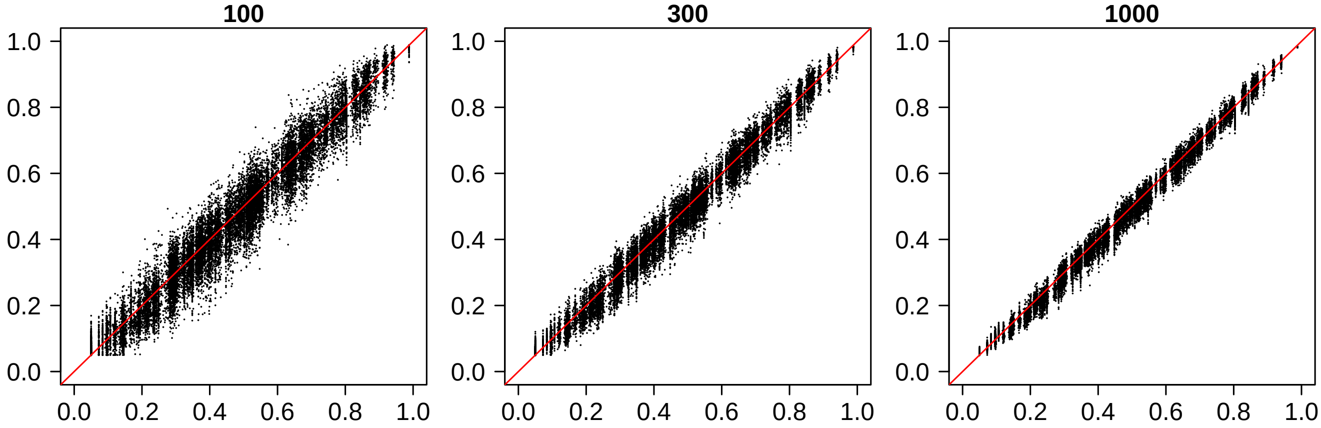

Figure 2. Estimated (y-axis) vs. true (x-axis) ERR in simulation U across a different number of studies.

with high effect sizes (d = 0.6) and no heterogeneity (SD = 0), but even with small heterogeneity, SD = 0.2, this problem disappeared. We therefore recommend extending confidence intervals by three percentage points. This is the default setting in the z-curve package, but the package allows researchers to change these settings. Moreover, in meta-analyses of studies with low heterogeneity, alternative methods that are more appropriate for homogeneous methods (e.g., selection models; Hedges, 1992) may be used or the number of components could be reduced.

EDR. Visual inspection of EDRs plotted against the true discovery rates (Figure 3) showed a noticeable increase in uncertainty. This is to be expected as EDR estimates require estimation of the distribution for statistically non-significant z-statistics solely on the basis of the distribution of statistically significant results.

Despite the high variability in estimates, they can be useful. With the observed discovery rate in psychology being often over 90% (Sterling, 1959), many of these estimates would alert readers that selection bias is present. A bigger problem is that the highly variable EDR estimates might lack the power to detect selection bias in small sets of studies.

Across all studies, systematic bias was small, 1.42 (SE_{MCMC} = 0.08) for 100 studies, 0.57 (SE_{MCMC} = 0.06) for 300 studies, 0.16 (SE_{MCMC} = 0.05) percentage points for 1000 studies. This shows that the shape of the distribution of statistically significant results does provide valid information about the shape of the full distribution. Consistent with Figure 3, RMSE values were large and remained fairly large even with larger number of studies, 11.70 (SE_{MCMC} = 0.11) for 100 studies, 8.88 (SE_{MCMC} = 0.08) for 300 studies, 6.49 (SE_{MCMC} = 0.07) percentage points for 1000 studies. These results show how costly selection bias is because more precise estimates of the discovery rate would be available without selection bias.

The main consequence of high RMSE is that confidence intervals are expected to be wide. The next analysis examined whether confidence intervals have adequate coverage. This was not the case; coverage = 87.3% (SE_{MCMC} = 0.14). We next used logistic regression to examine patterns in coverage in our simulation design. A notable 3-way interaction between effect size, sample size, and percentage of true $\mathcal{H}_0$ was present, z = 3.83. While the pattern was complex, not a single cell of the design showed coverage over 95%.

As before, we created robust confidence intervals by extending the interval. We settled for an extension by five percentage points. The 3-way interaction remained statistically significant, z = 3.36. Now 43 of the 48 cells showed coverage over 95%. For reasons that are not clear to us, the main problem occurred for an effect size of Cohen's d = 0.4 and no true $\mathcal{H}_0$, independent of sample size. While improving the performance of z-curve remains an important goal and future research might find better approaches to address this problem, for now, we recommend using z-curve 2.0 with these robust confidence intervals, but users can specify more conservative adjustments.

Application to Real Data

It is not easy to evaluate the performance of z-curve 2.0 estimates with actual data because selection bias is ubiquitous and direct replication studies are fairly rare (Zwaan et al., 2018). A notable exception is the Open Science Collaboration project that replicated 100 studies from three psychology journals (Open Science Collaboration, 2015). This unprecedented effort has attracted attention within and outside of psychological science and the article has already been cited over 1,000 times. The key finding was that out of 97 statistically signifi-

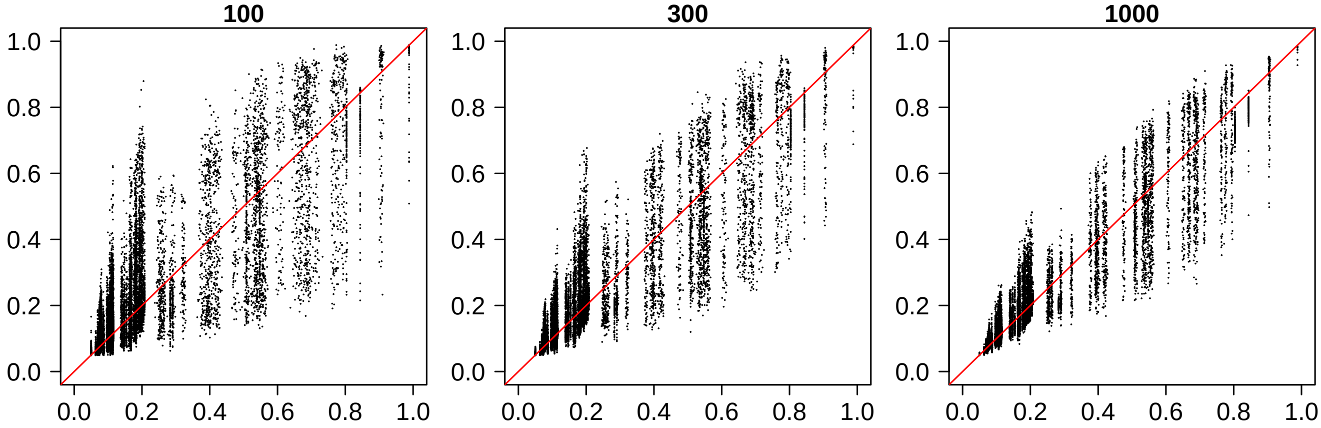

Figure 3. Estimated (y-axis) vs. true (x-axis) EDR across a different number of studies.

cant results, including marginally significant ones, only 36 replication studies (37%) reproduced a statistically significant result in the replication attempts.

This finding has produced a wide range of reactions. Often the results are cited as evidence for a replication crisis in psychological science, especially social psychology (Schimmack, 2020). Others argue that the replication studies were poorly carried out and that many of the original results are robust findings (Bressan, 2019). This debate mirrors other disputes about failures to replicate original results. The interpretation of replication studies is often strongly influenced by researchers' a priori beliefs. Thus, they rarely settle academic disputes. Z-curve analysis can provide valuable information to determine whether an original or a replication study is more trustworthy. If a z-curve analysis shows no evidence for selection bias and a high ERR, it is likely that the original result is credible and the replication failure is a false negative result or the replication study failed to reproduce the original experiment. On the other hand, if there is evidence for selection bias and the ERR is low, replication failures are expected because the original results were obtained with questionable research practices.

Another advantage of z-curve analyses of published results is that it is easier to obtain large representative samples of studies than to conduct actual replication studies. To illustrate the usefulness of z-curve analyses, we focus on social psychology because this field has received the most attention from meta-psychologists (Schimmack, 2020). We fitted z-curve 2.0 to two studies of published test statistics from social psychology and compared these results to the actual success rate in the Open Science Collaboration project (k = 55).

One sample is based on Motyl et al. (2017) assessment of the replicability of social psychology (k = 678). The other sample is based on the coding of the most highly cited articles by social psychologists with a high H-Index ($k = 2,208$; Schimmack, 2021). The ERR estimates were 44%, 95% CI $[35, 52]$%, and 51%, 95% CI $[45, 56]$%. The two estimates do not differ significantly from each other, but both estimates are considerably higher than the actual discovery rate in the OSC replication project, 25%, 95% CI $[13, 37]$%. We postpone the discussion of this discrepancy to the discussion section.

The EDRs estimates were 16%, 95% CI $[5, 32]$%, and 14%, 95% CI $[7, 23]$%. Again, both of the estimates overlap and do not significantly differ. At the same time, the EDR estimates are much lower than the ODRs in these two data sets (90%, 89%). The z-curve analysis of published results in social psychology shows a strong selection bias that explains replication failures in actual replication attempts. Thus, the z-curve analysis reveals that replication failures cannot be attributed to problems of the replication attempts. Instead, the low EDR estimates show that many non-significant original results are missing from the published record.

Discussion

A previous article introduced z-curve as a viable method to estimate mean power after selection for significance (Brunner & Schimmack, 2020). This is a useful statistic because it predicts the success rate of exact replication studies. We therefore call this statistic the expected replication rate. Studies with a high replication rate provide credible evidence for a phenomenon. In contrast, studies with a low replication rate are untrustworthy and require additional evidence.

We extended z-curve 1.0 in two ways. First, we implemented the expectation maximization algorithm to fit the mixture model to the observed distribution of z-statistics. This is a more conventional method to fit mixture models. We found that this method produces good estimates, but it did not eliminate some of the systematic biases that were observed with z-curve 1.0. More important, we extended z-curve to estimate the mean power before selection for significance. We call this statistic the expected discovery rate because mean power predicts the percentage of statistically significant results for a set of studies. We found that EDR estimates have satisfactory large sample accuracy, but vary widely in smaller sets of studies. This limits the usefulness for meta-analysis of small sets of studies, but as we demonstrated with actual data, the results are useful when a large set of studies is available. The comparison of the EDR and ODR can also be used to assess the amount of selection bias. A low EDR can also help researchers to realize that they test too many false hypotheses or test true hypotheses with insufficient power.

In contrast to Miller (2009), who stipulates that estimating the ERR ("aggregated replication probability") is unattainable due to selection processes, Brunner and Schimmack (2020) z-curve 1.0 addresses the issue by modeling the selection for significance.

Finally, we examined the performance of bootstrapped confidence intervals in simulation studies. We found that coverage for 95% confidence intervals was sometimes below 95%. To improve the coverage of confidence intervals, we created robust confidence intervals that added three percentage points to the confidence interval of the ERR and five percentage points to the confidence interval of the EDR.

We demonstrate the usefulness of the EDR and confidence intervals with an example from social psychology. We find that ERR overestimates the actual replicability in social psychology. We also find clear evidence that power in social psychology is low and that high success rates are mostly due to selection for significance. It is noteworthy that while the Motyl et al. (2017) dataset is representative for social psychology, Schimmack (2021) dataset sampled highly influential articles. The fact that both sampling procedures produced similar results suggests that studies by eminent researchers or studies with high citation rates are no more replicable than other studies published in social psychology.

Z-curve 2.0 does provide additional valuable information that was not provided by z-curve 1.0. Moreover, z-curve 2.0 is available as an R-package, making it easier for researchers to conduct z-curve analyses (Bartoš & Schimmack, 2020). This article provides the theoretical background for the use of the z-curve package. Subsequently, we discuss some potential limitations of z-curve 2.0 analysis and compare z-curve 2.0 to other methods that aim to estimate selection bias or power of studies.

Bias Detection Methods

In theory, bias detection is as old as meta-analysis. The first bias test showed that Mendel's genetic experiments with peas had less sampling error than a statistical model would predict (Pires & Branco, 2010). However, when meta-analysis emerged as a widely used tool to integrate research findings, selection bias was often ignored. Psychologists focused on fail-safe N (Rosenthal, 1979), which did not test for the presence of bias and often led to false conclusions about the credibility of a result (Ferguson & Heene, 2012). The most common tools to detect bias rely on correlations between effect sizes and sample size. A key problem with this approach is that it often has low power and that results are not trustworthy under conditions of heterogeneity (Inzlicht et al., 2015; Renkewitz & Keiner, 2019). The tests are also not useful for meta-analysis of heterogeneous sets of studies where researchers use larger samples to study smaller effects, which also introduces a correlation between effect sizes and sample sizes. Due to these limitations, evidence of bias has been dismissed as inconclusive (Cunningham & Baumeister, 2016; Inzlicht & Friese, 2019).

It is harder to dismiss evidence of bias when a set of published studies has more statistically significant results than the power of the studies warrants; that is, the ODR exceeds the EDR (Sterling et al., 1995). Aside from z-curve 2.0, there are two other bias tests that rely on a comparison of the ODR and EDR to evaluate the presence of selection bias, namely the Test of Excessive Significance (TES; Ioannidis & Trikalinos, 2007) and the Incredibility Test (IT; Schimmack, 2012).

Z-curve 2.0 has several advantages over the existing methods. First, TES was explicitly designed for meta-analysis with little heterogeneity and may produce biased results when heterogeneity is present (Renkewitz & Keiner, 2019). Second, both the TES and the IT take observed power at face value. As observed power is inflated by selection for significance, the tests have low power to detect selection for significance, unless the selection bias is large. Finally, TES and IT rely on *p*-values to provide information about bias. As a result, they do not provide information about the amount of selection bias.

Z-curve 2.0 overcomes these problems by correcting the power estimate for selection bias, providing quantitative evidence about the amount of bias by comparing the ODR and EDR, and by providing evidence about statistical significance by means of a confidence interval around the EDR estimate. Thus, z-curve 2.0 is a valuable tool for meta-analysts, especially when analyzing a large sample of heterogenous studies that vary widely in designs and effect sizes. As we demonstrated with our example, the EDR of social psychology studies is very low. This information is useful because it alerts readers to the fact that not all p-values below .05 reveal a true and replicable finding.

Nevertheless, z-curve has some limitations. One limitation is that it does not distinguish between significant results with opposite signs. In the presence of multiple tests of the same hypothesis with opposite signs, researchers can exclude inconsistent significant results and estimate z-curve on the basis of significant results with the correct sign. However, the selection of tests by the meta-analyst introduces additional selection bias, which has to be taken into account in the comparison of the EDR and ODR. Another limitation is the assumption that all studies used the same alpha criterion (.05) to select for significance. This possibility can be explored by conducting multiple z-curve analyses with different selection criteria (e.g., .05, .01). The use of lower selection criteria is also useful because some questionable research practices produce a cluster of just significant results. However, all statistical methods can only produce estimates that come with some uncertainty. When severe selection bias is present, new studies are needed to provide credible evidence for a phenomenon.

Predicting Replication Outcomes

Since 2011, many psychologists have learned that published significant results can have a low replication probability (Open Science Collaboration, 2015). This makes it difficult to trust the published literature, especially older articles that report results from studies with small samples that were not pre-registered. Should these results be disregarded because they might have been obtained with questionable research practices? Should results only be trusted if they have been replicated in a new, ideally pre-registered, replication study? Or should we simply assume that most published results are probably true and continue to treat every *p*-value below .05 as a true discovery?

The appeal of z-curve is that we can use the published evidence to distinguish between credible and "incredible" (biased) statistically significant results. If a meta-analysis shows low selection bias and a high replication rate, the results are credible. If a meta-analysis shows high selection bias and a low replication rate, the results are incredible and require independent verification.

As appealing as this sounds, every method needs to be validated before it can be applied to answer substantive questions. This is also true for z-curve 2.0. We used the results from the OSC replicability project for this purpose. The results suggest that z-curve predictions of replication rates may be overly optimistic. While the expected replication rate was between 44% and 51% (35% - 56% CI range), the actual success rate was only 25%, 95% CI $[13, 37]$%. Thus, it is important to examine why z-curve estimates are higher than the actual replication rate in the OSC project.

One possible explanation is that there is a problem with the replication studies. Social psychologists quickly criticized the quality of the replication studies (Gilbert et al., 2016). In response, the replication team conducted the new replications of contested replication studies. Based on the effect sizes in these much larger replication studies, not a single original study would have produced statistically significant results (Ebersole et al., 2020). It is therefore unlikely that the quality of replication studies explains the low success rate of replication studies in social psychology.

A more interesting explanation is that social psychological phenomena are not as stable as boiling distilled water under tightly controlled laboratory conditions. Rather, effect sizes vary across populations, experimenters, times of day, and a myriad of other factors that are difficult to control (Stroebe & Strack, 2014). In this case, selection for significance produces additional regression to the mean because statistically significant results were obtained with the help of favorable hidden moderators that produced larger effect sizes that are unlikely to be present again in a direct replication study.

The worst-case scenario is that studies that were selected for significance are no more powerful than studies that produced statistically non-significant results. In this case, the EDR predicts the outcome of actual replication studies. Consistent with this explanation, the actual replication rate of 25%, 95% CI $[13, 37]$%, was highly consistent with the EDR estimates of 16%, 95% CI $[5, 32]$%, and 14%, 95% CI $[7, 23]$%. More research is needed once more replication studies become available to see how closely actual replication rates are to the EDR and the ERR. For now, they should be considered the worst and the best possible scenarios and actual replication rates are expected to fall somewhere between these two estimates.

A third possibility for the discrepancy is that questionable research practices change the shape of the z-curve in ways that are different from a simple selection model. For example, if researchers have several statistically significant results and pick the highest one, the selection model underestimates the amount of selection that occurred. This can bias z-curve estimates and inflate the ERR and EDR estimates. Unfortunately, it is also possible that questionable research practices have the opposite effect and that ERR and EDR estimates underestimate the true values. This uncertainty does not undermine the usefulness of z-curve analyses. Rather it shows how questionable research practices undermine the credibility of published results. Z-curve 2.0 does not alleviate the need to reform research practices and to ensure that all researchers report their results honestly.

Conclusion

Z-curve 1.0 made it possible to estimate the replication rate of a set of studies on the basis of published test results. Z-curve 2.0 makes it possible to also estimate the expected discovery rate; that is, how many tests were conducted to produce the statistically significant results. The EDR can be used to evaluate the presence and amount of selection bias. Although there are many methods that have the same purpose, z-curve 2.0 has several advantages over these methods. Most importantly, it quantifies the amount of selection bias. This information is particularly useful when meta-analyses report effect sizes based on methods that do not consider the presence of selection bias.

Author Contact

František Bartoš; f.bartos96@gmail.com; Department of Psychological Methods, University of Amsterdam; Department of Arts, Faculty of Arts, Charles University; ORCID: 0000-0002-0018-5573

Ulrich Schimmack; ulrich.schimmack@utoronto.ca; Department of Psychology; University of Toronto, Mississauga; ORCID: 0000-0001-9456-5536

Acknowledgments

Access to computing and storage facilities owned by parties and projects contributing to the National Grid Infrastructure MetaCentrum provided under the program “Projects of Large Research, Development, and Innovations Infrastructures” (CESNET LM2015042), is greatly appreciated. We would like to thank Maximilian Maier, Erik W. van Zwet, and Leonardo Tozzi for valuable comments on a draft of this manuscript.

Data Availability Statement

Supplementary materials are accessible at https://osf.io/r6ewt/ and the R-package is accessible at https://cran.r-project.org/web/packages/zcurve/.

Conflict of Interest and Funding

No conflict of interest to report. This work was not supported by a specific grant.

Author Contributions

Most of the ideas in the manuscript were developed jointly. The main idea behind the z-curve method and its density version was developed by Dr. Schimmack. Mr. Bartoš implemented the EM version of the method and conducted the extensive simulation studies.

Open Science Practices

This article earned the Open Materials badge for making the materials openly available. It has been verified that the analysis reproduced the results presented in the article. The entire editorial process, including the open reviews, is published in the online supplement.

References

Bartoš, F., & Maier, M. (In press). Power or alpha? The better way of decreasing the false discovery rate. https://doi.org/10.31234/osf.io/ev29a

Bartoš, F., & Schimmack, U. (2020). zcurve: An R package for fitting z-curves [R package version 1.0.0]. https://CRAN.R-project.org/package=zcurve

Begley, C. G., & Ellis, L. M. (2012). Raise standards for preclinical cancer research. *Nature*, *483*(7391), 531–533. https://doi.org/10.1038/483531a

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.

Bressan, P. (2019). Confounds in “failed” replications. *Frontiers in Psychology*, *10*. https://doi.org/10.3389/fpsyg.2019.01884

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

Camerer, C. F., Dreber, A., Forsell, E., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Almenberg, J., Altmejd, A., Chan, T., et al. (2016). Evaluating replicability of laboratory experiments in economics. *Science*, *351*(6280), 1433–1436. https://doi.org/10.1126/science.aaf0918

Chang, A. C., & Li, P. (2015). Is economics research replicable? Sixty published papers from thirteen journals say “usually not”. *Finance and Economics Discussion Series 2015-083*. http://dx.doi.org/10.17016/FEDS.2015.083

Cunningham, M. R., & Baumeister, R. F. (2016). How to make nothing out of something: Analyses of the impact of study sampling and statistical interpretation in misleading meta-analytic conclusions. *Frontiers in Psychology*, *7*, 1639. https://doi.org/10.3389/fpsyg.2016.01639

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, *39*(1), 1–22. https://doi.org/10.1111/j.2517-6161.1977.tb01600.x

Ebersole, C. R., Mathur, M. B., Baranski, E., Bart-Plange, D.-J., Buttrick, N. R., Chartier, C. R., Corker, K. S., Corley, M., Hartshorne, J. K., IJzerman, H., et al. (2020). Many Labs 5: Testing pre-data-collection peer review as an intervention to increase replicability. *Advances in Methods and Practices in Psychological Science*, *3*(3), 309–331. https://doi.org/10.1177/2515245920958687

Eddelbuettel, D., & François, R. (2011). Rcpp: Seamless R and C++ integration. *Journal of statistical software*, *40*, 1–18. https://doi.org/10.18637/jss.v040.i08

Egger, M., Smith, G. D., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *BMJ*, *315*(7109), 629–634. https://doi.org/10.1136/bmj.315.7109.629

Elandt, R. C. (1961). The folded normal distribution: Two methods of estimating parameters from moments. *Technometrics*, *3*(4), 551–562. https://doi.org/10.2307/1266561

Ferguson, C. J., & Heene, M. (2012). A vast graveyard of undead theories: Publication bias and psychological science's aversion to the null. *Perspectives on Psychological Science*, *7*(6), 555–561. https://doi.org/10.1177/1745691612459059

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

Franco, A., Malhotra, N., & Simonovits, G. (2014). Publication bias in the social sciences: Unlocking the file drawer. *Science*, *345*(6203), 1502–1505. https://doi.org/10.1126/science.1255484

Gilbert, D., King, G., Pettigrew, S., & Wilson, T. (2016). Comment on "Estimating the reproducibility of psychological science". *Science*, *351*, 1037. http://dx.doi.org/10.1126/science.aad7243

Hedges, L. V. (1992). Modeling publication selection effects in meta-analysis. *Statistical Science*, 246–255. https://doi.org/10.1214/SS%5C%2F1177011364

Inzlicht, M., & Friese, M. (2019). The past, present, and future of ego depletion. *Social Psychology*. https://doi.org/10.1027/1864-9335/a000398

Inzlicht, M., Gervais, W., & Berkman, E. (2015). Bias-correction techniques alone cannot determine whether ego depletion is different from zero: Commentary on Carter, Kofler, Forster, & McCullough, 2015. *Kofler, Forster, & McCullough*. http://dx.doi.org/10.2139/ssrn.2659409

Ioannidis, J. P., & Trikalinos, T. A. (2007). An exploratory test for an excess of significant findings. *Clinical Trials*, *4*(3), 245–253. https://doi.org/10.1177/1740774507079441

John, L. K., 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

Lee, G., & Scott, C. (2012). Em algorithms for multivariate Gaussian mixture models with truncated and censored data. *Computational Statistics & Data Analysis*, *56*(9), 2816–2829. https://doi.org/10.1016/j.csda.2012.03.003

Leone, F. C., Nelson, L. S., & Nottingham, R. (1961). The folded normal distribution. *Technometrics*, *3*(4), 543–550. https://doi.org/10.1080/00401706.1961.10489974

Maier, M., Bartoš, F., & Wagenmakers, E.-J. (2022). Robust Bayesian meta-analysis: Addressing publication bias with model-averaging. *Psychological Methods*. 10.1037/met0000405

Miller, J. (2009). What is the probability of replicating a statistically significant effect? *Psychonomic Bulletin & Review*, *16*(4), 617–640. https://doi.org/10.3758/PBR.16.4.617

Morris, T. P., White, I. R., & Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, *38*(11), 2074–2102. https://doi.org/10.1002/sim.8086

Motyl, M., Demos, A. P., Carsel, T. S., Hanson, B. E., Melton, Z. J., Mueller, A. B., Prims, J., Sun, J., Washburn, A. N., Wong, K. M., et al. (2017). The state of social and personality science: Rotten to the core, not so bad, getting better, or getting worse? *Journal of Personality and Social Psychology*, *113*(1), 34–58. https://doi.org/10.1037/pspa0000084

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251).

Pashler, H., & Wagenmakers, E.-.-J. (2012). Editors' introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science*, *7*(6), 528–530. https://doi.org/10.1177/1745691612465253

Pires, A. M., & Branco, J. A. (2010). A statistical model to explain the Mendel—Fisher controversy. *Statistical Science*, 545–565. https://doi.org/10.1214/10-STS342

Prinz, F., Schlange, T., & Asadullah, K. (2011). Believe it or not: How much can we rely on published data on potential drug targets? *Nature Reviews Drug Discovery*, *10*(9), 712–712. https://doi.org/10.1038/nrd3439-c1

Renkewitz, F., & Keiner, M. (2019). How to detect publication bias in psychological research. *Zeitschrift für Psychologie*, *227*(4), 261–279. https://doi.org/10.1027/2151-2604/a000386

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, *86*(3), 638–641. https://doi.org/10.1037/0033-2909.86.3.638

Rosenthal, R., & Gaito, J. (1964). Further evidence for the cliff effect in interpretation of levels of significance. *Psychological Reports*, *15*(2), 570. https://doi.org/10.2466/pr0.1964.15.2.570

Scheel, A. M., Schijen, M. R., & Lakens, D. (2021). An excess of positive results: Comparing the standard psychology literature with registered reports. *Advances in Methods and Practices in Psychological Science*, *4*(2), 1–12. https://doi.org/10.1177/25152459211007467

Schimmack, U. (2020). The Replicability Index is the most powerful tool to detect publication bias in meta-analyses. https://replicationindex.com/2020/01/01/the-replicability-index-is-the-most-powerful-tool-to-detect-publication-bias-in-meta-analyses/

Schimmack, U. (2021). Rotten to the core II: A replication and extension of Motyl et al. https://replicationindex.com/2021/03/10/rotten-to-the-core-ii-a-replication-and-extension-of-motyl-et-al/

Schimmack, U. (2012). The ironic effect of significant results on the credibility of multiple-study articles. *Psychological Methods*, *17*(4), 551. https://doi.org/10.1037/a0029487

Sorić, B. (1989). Statistical “discoveries” and effect-size estimation. *Journal of the American Statistical Association*, *84*(406), 608–610. https://doi.org/10.2307/2289950

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/10.1080/00031305.1995.10476125

Stroebe, W., & Strack, F. (2014). The alleged crisis and the illusion of exact replication. *Perspectives on Psychological Science*, *9*(1), 59–71. http://dx.doi.org/10.1177/1745691613514450

Ujeh, C., Satitvipawee, P., Sillabutra, J., Soontornpipit, P., Kitidamrongsuk, P., & Viwatwongkasem, C. (2016). Bootstrapping with R to determine variances of mixture model estimates in predicting confidence intervals for population sizes. *Procedia Computer Science*, *86*, 216–219. https://doi.org/10.1016/j.procs.2016.05.103

Vevea, J. L., & Hedges, L. V. (1995). A general linear model for estimating effect size in the presence of publication bias. *Psychometrika*, *60*(3), 419–435. https://doi.org/10.1007/BF02294384

Wegner, D. M. (1992). The premature demise of the solo experiment. *Personality and Social Psychology Bulletin*, *18*(4), 504–508. https://doi.org/10.1017/S0140525X17001972

Zwaan, R. A., Etz, A., Lucas, R. E., & Donnellan, M. B. (2018). Making replication mainstream. *Behavioral and Brain Sciences*, *41*, 120. https://doi.org/10.1017/S0140525X17001972