Z-curve 2.0: Estimating Replication Rates and Discovery Rates

František Bartoš[1,2,*], Ulrich Schimmack[3]

1 University of Amsterdam

2 Faculty of Arts, Charles University

3 University of Toronto, Mississauga

Correspondence concerning this article should be addressed to: František Bartoš, University of Amsterdam, Department of Psychological Methods, Nieuwe Achtergracht 129-B, 1018 VZ Amsterdam, The Netherlands, fbartos96@gmail.com

Author Note

Access to computing and storage facilities owned by parties and projects contributing to the National Grid Infrastructure MetaCentrum provided under the programme "Projects of Large Research, Development, and Innovations Infrastructures" (CESNET LM2015042), is greatly appreciated. We would like to thank Maximilian Maier for helpful comments and suggestions on previous versions of this manuscript.

Abstract

Publication bias, the fact that published studies are not necessarily representative of all conducted studies, poses a significant threat to the credibility of scientific literature. To mitigate the problem, we introduce z-curve 2.0 as a method that estimates two interpretable measures for the credibility of scientific literature based on test-statistics of published studies - the expected replication rate (ERR) and the expected discovery rate (EDR). Z-curve 2.0 extends the work by Brunner and Schimmack (2020) in several ways. First, we extended z-curve to estimate the number of all studies that were conducted, including studies with statistically non-significant results that may not have been reported, solely on the basis of statistically significant results. This allows us to estimate the EDR; that is, the percentage of statistically significant results that were obtained in all studies. EDR can be used to assess the size of the file-drawer, estimate the maximum number of false positive results, and may provide a better estimate of the success rate in actual replication studies than the ERR because exact replications are impossible. Second, add bootstrapped confidence intervals to provide information about the uncertainty in the estimates. We show in two simulation studies that new estimation methods outperform the original version of z-curve 1.0 and *p*-curve, and illustrate the usage on the example of the Reproducibility Project: Psychology.

Keywords: Publication Bias, Replicability, Expected Replication Rate, Expected Discovery Rate, File-Drawer

Z-curve 2.0: Estimating Replication Rates and Discovery Rates

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 was statistically significant (Rosenthal & Gaito, 1964; Scheel, Schijen, & Lakens, 2020; Sterling, 1959; Sterling et al., 1995). This selective publishing of statistically significant results introduces a bias in the published literature. At 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 (cf. Wegner, 1992). However, internal replication studies provided a false sense of replicability because researchers reported only successful replication attempts (Francis, 2014; John, Lowenstein, & Prelec, 2012; Schimmack, 2012). The pervasive presence of publication bias is a crucial reason of a replication crisis in many sciences (psychology: Open Science Collaboration, 2015; Pashler & Wagenmakers, 2012, medicine: Begley & Ellis, 2012; Prinz, Schlange, & Asadullah 2011, and economics: Camerer et. al., 2016; Chang & Li, 2015).

To address the problem of publication bias, different methods for the detection of and adjustment for publication bias were developed. However, many of them (Egger, Smith, Schneider, & Minder, 1997; Ioannidis and 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; Hedges & Vevea, 1996; Maier, Bartoš & Wagenmakers, 2020). Moreover, while the aforementioned methods test for publication bias (return a *p*-value or a Bayes factor), they do not

provide an interpretable measure of credibility (i.e., an effect size). Z-curve is aspiring to lift those limitations by modeling the observed test statistics with a finite mixture model. Therefore, no homogeneity or meta-analytical model assumptions are needed, and it may be a superior method for the detection of publication bias when heterogeneity is present. Furthermore, z-curve provides two interpretable measures for the credibility of a set of published studies - Expected Replicability Rate (ERR) and Expected Discovery Rate (EDR).

In the following sections, we first explain ERR and EDR - the two measures for assessing the credibility of literature. Second, we introduce the new z-curve estimation methods. Third, we evaluate the performance of the original z-curve (z-curve 1.0, Brunner & Schimmack, 2020), *p*-curve (Simonsohn, Nelson, & Simmons, 2014), and the new z-curve 2.0 using two simulation studies. Finally, we present an applied example where we fit the z-curve 2.0 to the test statistics from original studies whose replication was attempted by the Reproducibility Project: Psychology (Open Science Collaboration, 2015). We provide implementation of the presented methods in zcurve R package (Bartoš & Schimmack, 2020).

Expected Replication Rate

Our work builds on the mathematical foundations and simulation studies conducted by Brunner and Schimmack (2020). Brunner and Schimmack showed that the success rate for a set of exact replication studies is equivalent to the mean power of the original studies, which can be estimated from test statistics (e.g., *t*-test, *F*-test, etc.) of published studies. They explained that the term power can be a bit confusing because it is usually used for the conditional probability of obtaining a statistically significant result when the null-hypothesis is false. Z-curve 1.0 cannot estimate this conditional probability because it does not distinguish between true and false hypothesis. Thus, mean power referred to the unconditional probability of obtaining statistically

significant results, which includes true null-hypothesis with a probability of alpha. That is, mean power is 5% when alpha is .05 and all studies are false positives. To avoid confusion, we distinguish conditional and unconditional power. It is also important to distinguish between two populations of studies (Brunner & Schimmack, 2020). One population of studies are all studies that were conducted, including studies with statistically non-significant results. The other population is the population of studies that produced a statistically significant result. The unconditional mean power of the studies with a significant outcome determines the success rate if all of these studies were replicated exactly. Z-curve 1.0 estimates the unconditional mean power of a set of studies that were selected to be statistically significant to predict the percentage of significant results if the studies were exactly replicated. We call this probability the expected replication rate (ERR).

Expected Discovery Rate

A new feature of z-curve 2.0 is that it estimates the expected discovery rate (EDR). As noted above, it is important to distinguish between the population of all studies that were conducted, and the population of studies selected for statistical significance. These two populations have different levels of mean power unless all studies have the same power. The reason is that selection for statistical significance favors studies with high power because studies with high power are more likely to produce statistically significant results. Consequently, the unconditional mean power of the full population is lower than the unconditional mean power after selection for statistical significance (Brunner & Schimmack, 2020). The discovery rate is simply the unconditional mean power of all studies that were conducted, including those with statistically non-significant results. If all studies were published, the discovery rate is simply the percentage of studies with statistically significant results. However, when selection bias is

present, the observed discovery rate overestimates the true discovery rate. Z-curve 2.0 aims to estimate the discovery rate based on a finite mixture model of the studies with statistically significant results.

For example, if researchers tested 40 false hypotheses (H0 is true) and 60 true hypotheses with 80% power, we would observe 40*.05 + 60*.80 = 2 + 48 = 50 statistically significant results. Thus, the true discovery rate is 50/100 = 50%. If only 10 out of the 40 statistically non-significant results are reported, the observed discovery rate is 50/(50+10) = 83%. This observed discovery rate provides a false impression of the robustness of a literature. We examine whether z-curve 2.0 can recover the true discovery rate of 50% by fitting a finite mixture model to the test statistics of the 50 studies with statistically significant results. We refer to this estimate as the expected discovery rate (EDR).

Imagine a set of studies with the same unconditional power *p* (probability of obtaining a statistically significant result irrespective of the null hypothesis being false or true). If all of them test the same effect with a two-sided z-test, their *p*-values converted to z-scores follow a normal distribution with mean μ_z and standard deviation equal to 1. Using an alpha level α, the relationship between *p* and μ_z can be depicted in Equation 1. It describes that the power *p* is equal to the sum of probability of a z-statistic higher than the cutoff z-score corresponding to α on the right and left side of the distribution with Φ standing for cumulative density function of standard normal distribution,

$$p = 1 - \Phi(\Phi^{-1}\left(1 - \frac{a}{2}\right) - \mu_z) + \Phi(\Phi^{-1}\left(\frac{a}{2}\right) - \mu_z). \quad (1)$$

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, Nelson, & Nottingham, 1961). Switching from z-statistics to absolute z-statistics does not impact Equation 1.

If we have *K* studies with heterogenous power due to variation in effect sizes or sample sizes, the distribution of expected test statistics can be approximated with a mixture of *K* normal distributions. These normal distributions are centered at studies individual means μ_{zk} (k = 1:K) corresponding to their powers p_k with standard deviation 1, respectively to the corresponding folded normal distributions.

However, statistically non-significant p-values are often not published. Publication bias works as a censoring mechanism on p-values that leaves only p-values less than the statistical significance criterion, alpha. Even if some non-significant p-values are reported, their distribution is subject to unknown selection effects. Therefore, only observed p-values lower than alpha are used to estimate ERR and EDR, making the folded normal distribution truncated from left at z-score corresponding to alpha.

Figure 1 illustrates key concepts. The first row of Figure 1 shows folded standard normal distributions for studies with 0.3, 0.5, and 0.8 power and a mixture of these studies with equal weights assigned to the three power values. It can be seen that the mode of the distributions moves to the right with increasing levels of power. The second row illustrates the effect of selection for statistical significance which is z = 1.96 with p = .05, two-sided.

The third row illustrates the discovery rate, which is the proportion of the area under the curve on the right side of the statistical significance criterion. Z-curve aims to estimate the full area under the curve, including the area for statistically non-significant results on the basis of the shape of the truncated distribution on the right side of the statistical significance criterion.

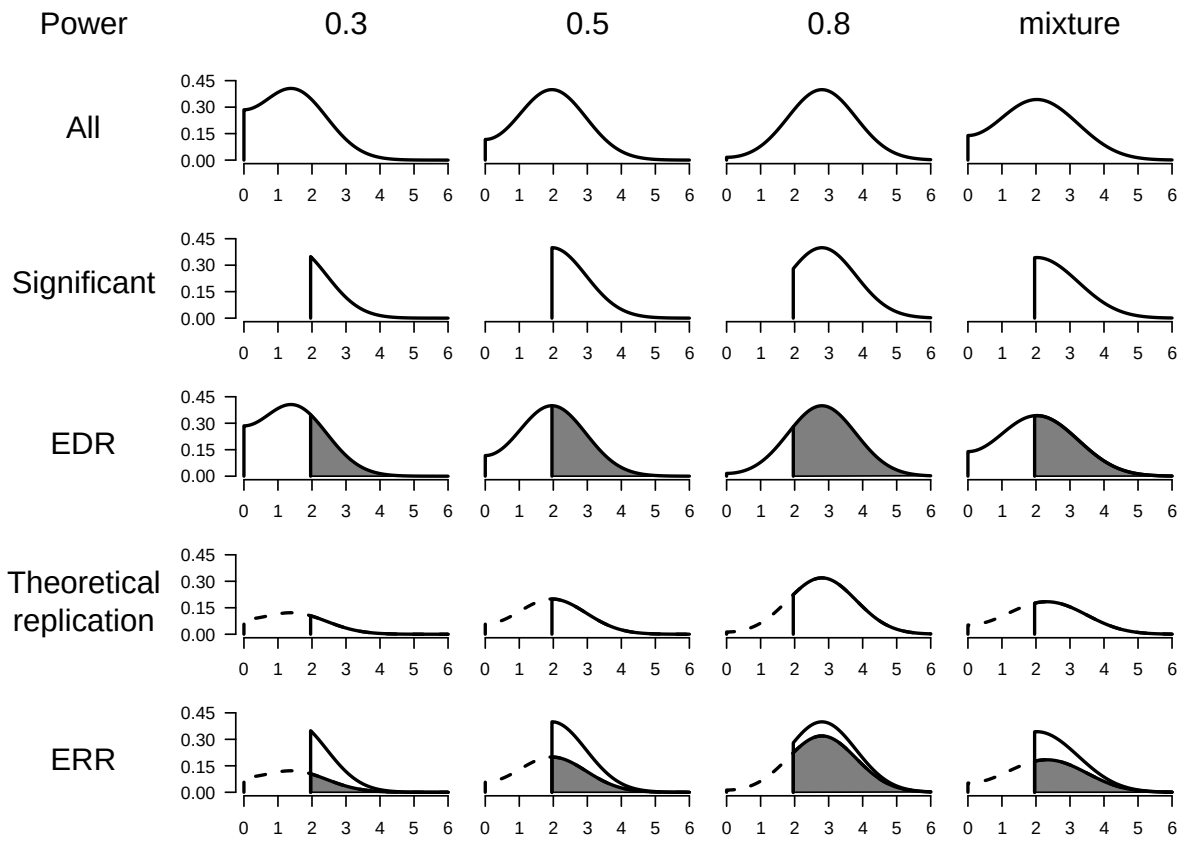

Figure 1. Explanation of EDR and ERR using distributions of z-scores for studies with different power (columns).

The fourth row shows the distribution of z-statistics that is expected if only the statistically significant studies are replicated exactly. As shown, some of these studies will produce statistically non-significant results even though they are exact replications of studies that were selected for statistical significance. The reason is that selection for statistical significance produces inflated estimates of effect sizes that are bound to regress to the mean in the replication studies. The fifth row illustrates the replicability rate the proportion of studies that did reproduce a statistically significant result in the exact replication attempts. As shown, the ERR increases with increasing power of studies.

We can also define the EDR and ERR using the power of original studies. In case of EDR, the proportion of statistically significant studies from *K* conducted studies is simply the mean of the individual studies' power p_k,

$$EDR = \frac{\sum_{k=1}^{K} p_k}{K}. \quad (2)$$

And the ERR is a proportion of successfully replicated statistically significant studies from all statistically significant studies. It can be written as a probability obtaining a statistically significant result twice in a row ($p_k * p_k$) dived by the number of all statistically significant studies,

$$ERR = \frac{\sum_{k=1}^{K} p_k * p_k}{\sum_{k=1}^{K} p_k}. \quad (3)$$

The EDR is useful for three reasons. First, it can be used to compare the EDR with the observed discovery rate (i.e., the percentage of statistically significant results that are published or were retrieved for a meta-analysis). Discrepancies between EDR and ODR suggest that publication bias is present. EDR can also be used to examine the false positive rate; that is, the percentage of statistically significant results that are false positives. Although it is theoretically impossible to determine the false positive rate, the maximum false positive rate is a function of the discovery rate (Sorić, 1989). When publication bias is present, the observed discovery rate would underestimate the risk of false positive results. However, z-curve estimates of the EDR correct for publication bias and provide some information about the maximum number of false positive results. Finally, it is possible that the EDR is a better predictor of success rates in actual replication studies than the expected replication rate (ERR). ERR estimates predict the outcome of replication studies under ideal conditions where replication studies are exact copies of the original studies. However, if replication studies are merely sampled from a population of similar studies with varying effect sizes, selection bias will favor studies with larger effect sizes, and

regression to the mean will reduce the success rate of replication studies. As EDR is not based on a single study, but rather a population of studies that were attempted, it is more likely to reflect the success rate when regression to the mean is taken into account.

Z-curve 2.0

We developed two versions of z-curve that are based on z-curve 1.0 (see Brunner & Schimmack, 2020 for technical details of z-curve 1.0). Rather than trying to estimate the powers of individual studies directly, z-curve uses the observed statistically significant results (the second row in Figure 1) to estimate the whole distribution of all conducted studies (the first row in Figure 1). Z-curve 2.0 does that by using a finite mixture model of $J = 7$ truncated folded normal distributions, with probability density function of an observed test statistic z,

$$f(z, \Theta) = \sum_{j=1}^{J} \pi_j f_{j[a,b]}(z; \theta_j). \quad (4)$$

Each mixture component j has its own weight π_j and probability density function $f_{j[a,b]}$ with parameters θ_j. We set the standard deviations of all components to 1 and space their means μ_j equally across the z-scores at values 0, 1, 2, 3, 4, 5, and 6[1]. We truncate the distribution function from left at a corresponding to the statistical significance level α and from left at $b = 6$. All z-scores higher than 6 are very likely to come from distribution with power essentially equal to 1 and their removal mitigates estimation issue. After the model is estimated, we add the held-out z-statistics back with p_k equal to 1 and rescale the estimated weights $\hat{\pi}_j$ to add to 1.

We use the model parameters and Equation (1) and (2) to estimate the EDR,

1 Comparable results (for the EM algorithm) could be obtained if the number and/or location of the components were estimated and selected using AIC/BIC. The described model setting is presented for simplicity and computational efficiency.

$$\widehat{EDR} = \sum_{j=1}^{J} \hat{\pi}_j * (1 - \Phi(\Phi^{-1}\left(1 - \frac{a}{2}\right) - \mu_j) + \Phi(\Phi^{-1}\left(\frac{a}{2}\right) - \mu_j)), \quad (5)$$

and Equation (1) and (3) to estimate the ERR,

$$\widehat{ERR} = \frac{\sum_{j=1}^{J} (\hat{\pi}_j * (1 - \Phi(\Phi^{-1}\left(1 - \frac{a}{2}\right) - \mu_j) + \Phi(\Phi^{-1}\left(\frac{a}{2}\right) - \mu_j)))^2}{\sum_{j=1}^{J} \hat{\pi}_j * (1 - \Phi(\Phi^{-1}\left(1 - \frac{a}{2}\right) - \mu_j) + \Phi(\Phi^{-1}\left(\frac{a}{2}\right) - \mu_j))}. \quad (6)$$

Density based z-curve

The first version, KD2, is a direct extension of z-curve 1.0. That is, the distribution of the observed z-statistics is first approximated with a truncated kernel-density distribution for z-statistics ranging from the statistical significance criterion (z = 1.96) to the value for extreme z-statistics (z > 6). Then, we estimate weights π_j of the mixture model outlined in Equation (4), by minimizing root mean square error (RMSE) of the estimated z-statistics' density and the mixture model's density using the nlminb package in R (Wuertz, 2014).

Expectancy-maximization (EM) z-curve

The second method does not require fitting a kernel-density distribution to the observed z-statistics. Instead, we fit the model (Equation 3) directly to the observed z-scores using the EM algorithm (Dempster, Laird, & Rubin, 1977, Lee & Scott, 2012). The EM algorithm maximizes the logarithmic likelihood of data given the model parameters. After initiation with starting values, it proceeds in two steps. First, the "E" step, computes the posterior probability of the individual data points belonging to a given component using the components parameters. Second, it maximizes the components parameter values given the posterior probabilities of individual components. The algorithm oscillates between the two steps until it reaches a convergence criterion or a prespecified number of iteration (Bishop, 2006).

In order to prevent the algorithm from reaching local minima, we run it 20 times with randomly selected starting weights and terminate the algorithm in the first 100 iterations, or if the

criterion falls below 1e-3, then select the outcome with the highest likelihood value and continue until 1000 iterations or reaching criterion 1e-5. To speed up the fitting process we optimizing the procedure using Rcpp (Eddelbuettel et al., 2011).

P-curve

P-curve is an alternative method for the estimation of unconditional mean power after selection for significance, which we call the Expected Replication Rate (Simonsohn, Nelson, & Simmons, 2014). As it was published several years before z-curve.1.0, it has become a popular tool to examine the credibility of meta-analytic results (P-Curve has more than 1,000 citations according to Google Scholar, at the time of writing).

Brunner and Schimmack (2020) compared p-curve and z-curve.1.0 and found that p-curve provides systematically inflated point estimates of the ERR when power varies across studies. Moreover, Brunner (2018) demonstrated that the latest version of p-curve (p-curve 4.06 that was implemented in 20zz) even produces inflated estimates when effect sizes are homogeneous, but studies vary only in sample sizes. Brunner and Schimmack (2020) focused on point-estimates. Here we extend their investigation of p-curve by comparing the coverage of 95% confidence intervals for z-curve.2.0 and p-curve. Our results provide valuable information for meta-analysts because p-curve 4.06 has been adopted as a method of choice without proper evaluation of the coverage of its confidence intervals.

Simulations

In a previous simulation study, Brunner and Schimmack (2020) compared several methods for the estimation of mean power and found that the original z-curve performed best under realistic conditions; that is, when effect sizes varied across studies and the distribution of effect sizes was unknown. Here, we conducted two new simulation studies that examined the

performance of z-curve and *p*-curve across a wider range of conditions. The most important new feature was that we simulated mixtures of false positives (the null-hypothesis is true) and real effects.

Since it is impossible to examine the performance of z-curve for all possible scenarios when studies are heterogeneous in power. To demonstrate the robustness of z-curve estimates, and to guard against simulation-hacking, each author independently created a simulation study. František Bartoš created Simulation F and Ulrich Schimmack created Simulation U.

Simulation F

The first simulation scenario used beta distributions scaled and shifted to range from .05 to 1 to create true power distributions of studies with mean power μ_p and sd σ_p (possible shapes of these distributions in Figure 2). True power of individual studies p_k was randomly sampled from the true power distribution. The true power values were transformed into corresponding μ_{zk} using Equation (1). Observed z-statistics were obtained by randomly sampling from the normal distributions centered at μ_{zk} with standard deviation of 1. We also added proportion of false-positives results (FDR) with z-statistics sampled from a standard normal distribution. Finally, we took absolute value of the observed z-statistics before passing them to our estimation methods.

We sampled the μ_p from uniform distribution (0.10, 0.95), FDR from uniform distribution (0, 1) and introduced heterogeneity in power by setting σ_p to 0.05 or 0.10. For each number of statistically significant studies K = (100, 300, 1000) we generated 25000 datasets.

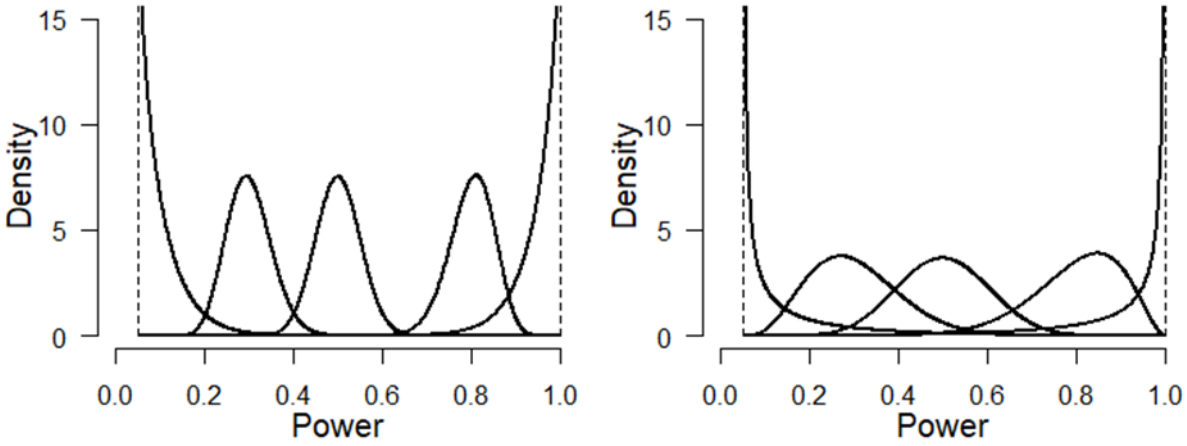

Figure 2. Different distributions for power of studies with μ_p = {0.10, 0.30, 0.50, 0.80, 0.95) with σ_p = 0.05 on left and σ_p = 0.10 on right.

Simulation U

The first set of simulations simulated sampling error with a standard normal distribution. This may advantage z-curve because the p-values originated from a standard normal distribution. However, most test statistics in psychology are t-tests or F-tests that do not have symmetrical distributions when the null-hypothesis is false. In this case, the use of z-statistics is an approximation that may introduce some systematic bias. Brunner and Schimmack (2020) found that this did not influence the estimated replication rate. However, it could have a stronger influence on estimates of the estimated discovery rate. Thus, simulation U simulated t-tests. The simulation also relied on knowledge about typical effect sizes and sample sizes in psychology to test z-curve 2.0 under realistic conditions.

The mean effect sizes, Cohen's ds, ranged from 0 to .6 (0, .2, .4., .6). Heterogeneity in effect sizes was simulated with a normal distribution around the mean effect size with SDs ranging from 0 to .6 (0, .2, .4., .6). Sample sizes for a between-subject design were N = 50, 100, and 200. In addition to this manipulation of power, the simulations included some studies with

true null-hypotheses. The percentage of true null-hypotheses ranged from 0 to 60% (0%, 20%, 40%, 60%). Each cell of this design was tested for sets of 100, 300, and 1,000 statistically significant studies. This 4 x 4 x 4 x 3 x 3 design has 576 cells. To obtain reliable estimates, we ran 100 simulations for each of the 576 cells. The simulations and model fitting functions are accessible at https://osf.io/r6ewt/.

Evaluation

We evaluated the algorithms using bias, mean distance between estimated and true values, root mean square error (RMSE) and confidence interval coverage. To check the performance of the z-curve across different simulation settings (Appendix A and B), we plotted bias across true FDR and power. Because the simulations F sampled parameters uniformly from the parameter space, we fitted generalized additive models (GAM) with s-spline over true FDR, power and standard deviation of the power distribution using mgcv package in R (Wood, 2012). We used the estimated model to display variation in bias and CI coverage as a function of true parameter values. On the other hand, the simulations U use discrete parameter values and fully factorial design, therefore, we analyzed them using analyses of variance (ANOVAs) and logistic regressions. The analysis scripts and results are accessible at https://osf.io/r6ewt/.

Results

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 in simulations F (Figure 3a) nor simulation U (Figure 3b). However, the *p*-curve ERR estimates showed systematic overestimation for high values and underestimation for low true ERR values. Furthermore, whereas z-curves estimates were converging to the true values with increasing

number of studies, the same was not true for *p*-curve which particularly struggled in simulation U.

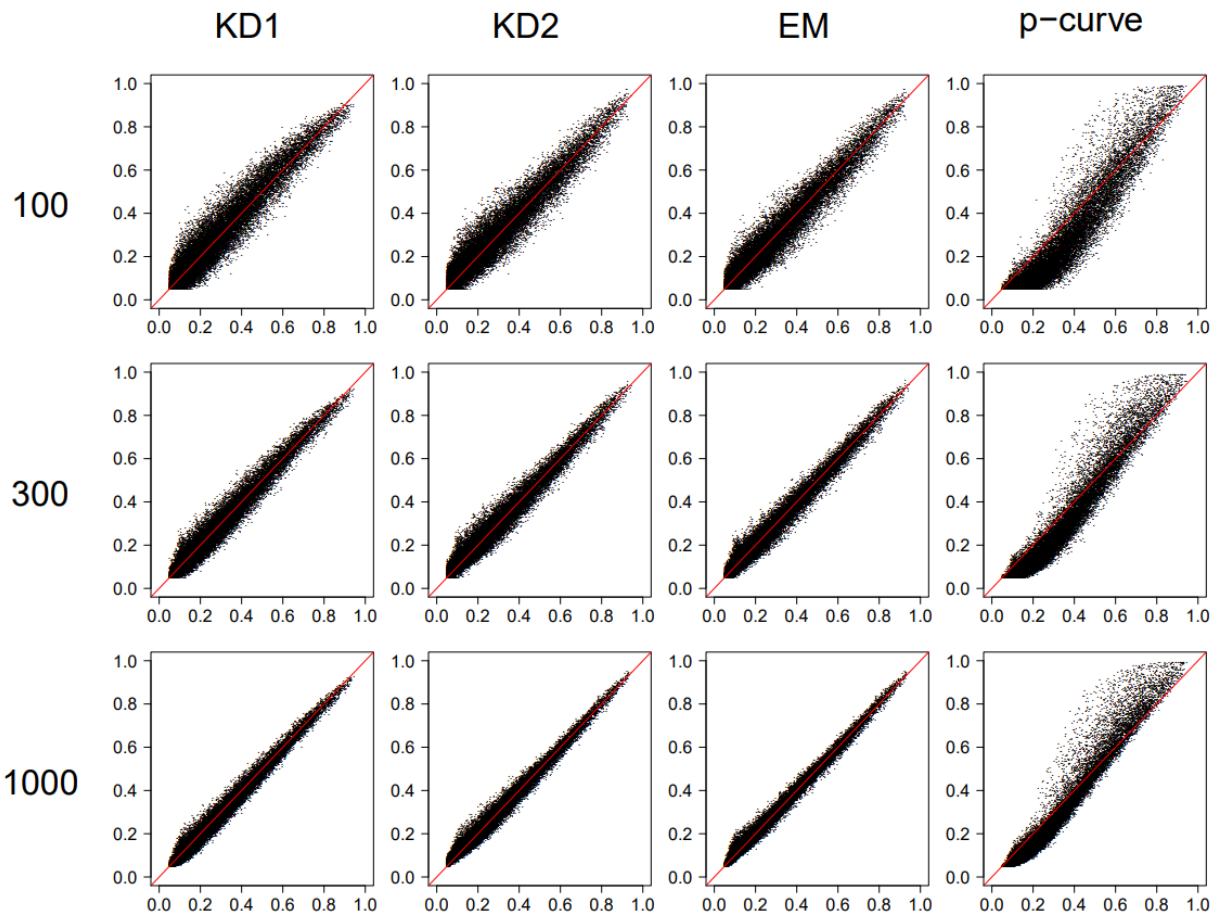

Figure 3a. Estimated (y-axis) vs true (x-axis) ERR in simulation F across a different number of studies.

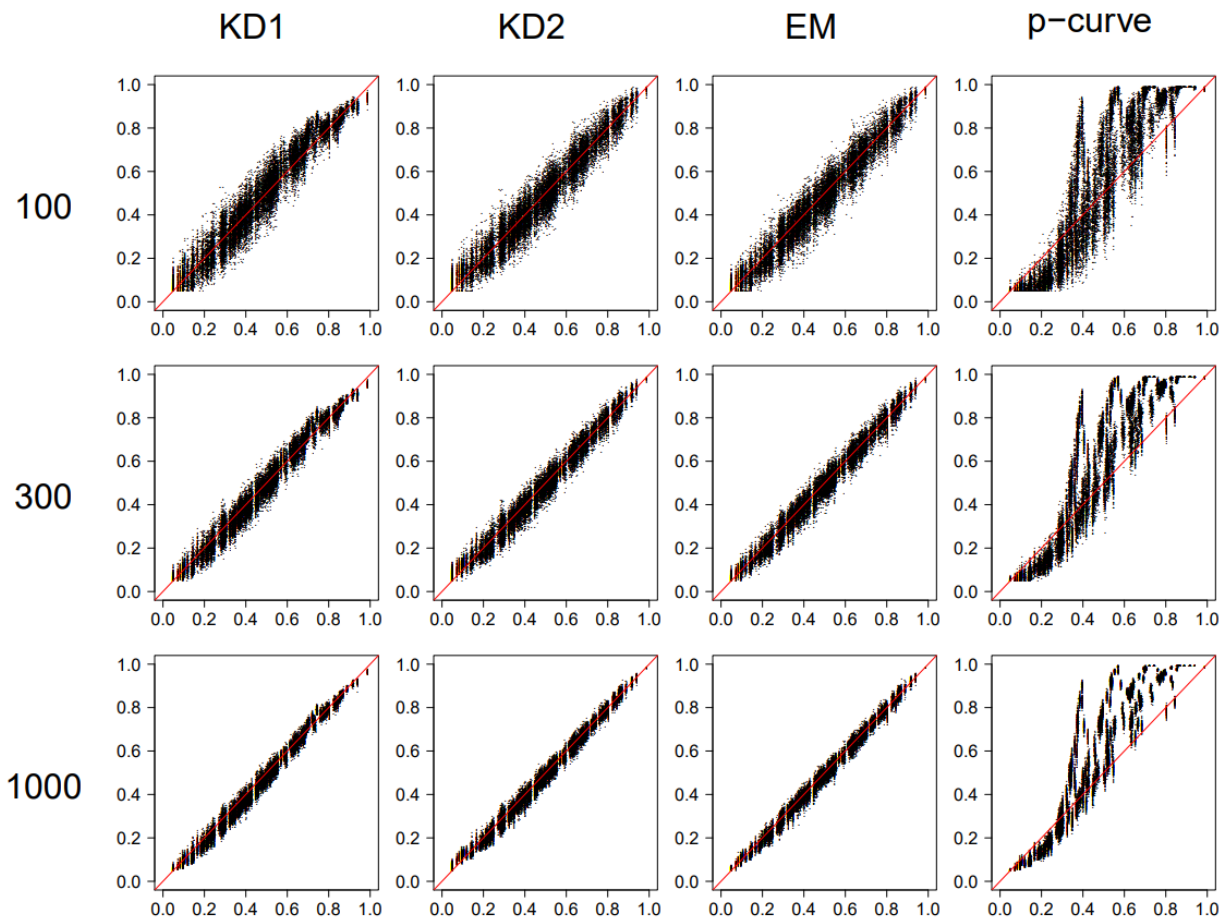

Figure 3b. Estimated (y-axis) vs true (x-axis) ERR in simulation U across a different number of studies.

Tables 1 and 2 confirmed the visual inspection. The bias (Table 1) and RMSE (Table 2) of ERR was decreasing with sample size for all z-curves, with the EM-algorithm slightly outperforming the kernel-density method in all condition. Furthermore, all z-curves noticeably outperformed *p*-curve and the bias and RMSE of *p*-curve was increasing with increasing number of studies in simulation U.

	Simulation F				Simulation U			
	KD1	KD2	EM	*p*-curve	KD1	KD2	EM	*p*-curve
100	0.007	0.008	-0.002	-0.066	0.006	-0.004	-0.009	0.049
300	-0.003	0.004	-0.001	-0.043	-0.007	-0.007	-0.009	0.078
1000	-0.008	0.002	0.001	-0.024	-0.015	-0.009	-0.008	0.096

Table 1. The bias of estimated ERR in each simulation across a different number of studies.

	Simulation F				Simulation U			
	KD1	KD2	EM	*p*-curve	KD1	KD2	EM	*p*-curve
100	0.058	0.056	0.050	0.099	0.057	0.053	0.051	0.164
300	0.038	0.037	0.032	0.075	0.038	0.035	0.033	0.165
1000	0.028	0.026	0.021	0.063	0.029	0.026	0.022	0.170

Table 2. RMSE of estimated ERR in each simulation across a different number of studies.

The CI coverage of ERR (Table 3) showed problems for all algorithms and was not improving with increasing number of studies. The CI coverage issues were more pronounced in simulation U. The EM-algorithm performed better for simulation F and the KD2 algorithm performed better for simulation U, but none of the methods produced acceptable coverage across all conditions. Nevertheless, all z-curves notably outperformed *p*-curve which produced coverage as low as 17.4%., and never exceeded 50% coverage for a 95% confidence interval.

	simulation F				simulation U			
	KD1	KD2	EM	*p*-curve	KD1	KD2	EM	*p*-curve
100	0.947	0.935	0.933	0.448	0.894	0.913	0.885	0.214
300	0.935	0.901	0.924	0.405	0.884	0.902	0.872	0.202
1000	0.893	0.839	0.891	0.272	0.803	0.853	0.830	0.174

Table 3. CI coverage of ERR in each simulation across a different number of studies.

The main reason for low coverage was systematic bias. When the estimate is systematically biased, the CIs are centered over the wrong value. To address this problem, we

constructed conservative confidence intervals (CCI) by extending the CI to account for systematic bias. We found that adding three percentage points on each side dramatically improved coverage for z-curves across all sample sizes (Table 4) and provided sufficient improvement to produce at least nominal coverage for all z-curves across most simulated scenarios (Appendix A). Even though the *p*-curve CI coverage also improved, it still failed to reach acceptable levels of coverage.

	simulation F				simulation U			
	KD1	KD2	EM	*p*-curve	KD1	KD2	EM	*p*-curve
100	0.986	0.984	0.984	0.681	0.985	0.990	0.984	0.335
300	0.991	0.986	0.992	0.731	0.991	0.995	0.990	0.353
1000	0.993	0.985	0.991	0.726	0.995	0.998	0.992	0.362

Table 4. Adjusted CI coverage of ERR in each simulation across a different number of studies.

EDR

Visual inspection of EDRs plotted against the true discovery rates (Figure 4) showed a noticeable increase in uncertainty around the true values as a result of extrapolating from the observed statistically significant results towards the unobserved statistically non-significant results. In addition, there was a systematic overestimation over the whole range of true discovery rates for both newly proposed z-curve methods in simulation F (Figure 4 left) and simulation U (Figure 4 right). Nevertheless, the overestimation of EDR was decreasing with increasing number of studies, and the estimates were converging to their true values.

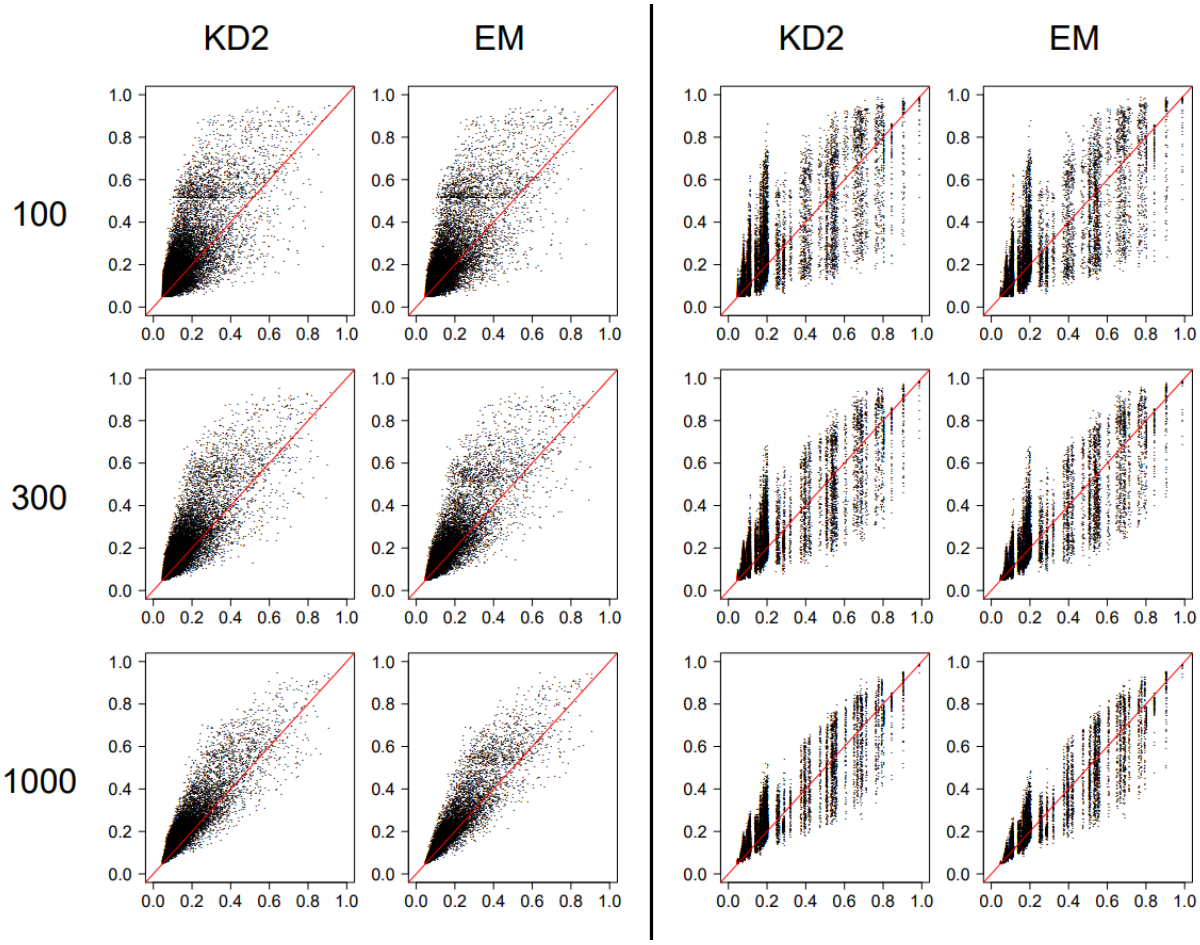

Figure 4. Estimated (y-axis) vs true (x-axis) EDR in simulation F (left panel) and simulation U (right panel) across a different number of studies.

Overestimation was also clearly visible in the deviations of estimated EDRs from true EDR (Table 5). The problem was more pronounced in simulation F and was decreasing with increasing sample size. The EM algorithm outperformed KD2 in both simulations and across all sample sizes. Furthermore, EM estimates were almost unbiased (lower than 0.01) in simulation U with higher samples sizes (N = 300 and 1000).

	Simulation F		Simulation U	
	KD2	EM	KD2	EM
100	0.041	0.030	0.029	0.014

300	0.034	0.023	0.021	0.006
1000	0.030	0.016	0.020	0.002

Table 5. The bias of estimated EDR in each simulation across a different number of studies

The much higher RMSE of EDRs (Table 6) than ERRs (Table 2) confirms that it is much harder to predict discovery rates than to predict replication rates due to the missing information about statistically non-significant results. The RMSE of the EDRs was higher in simulation U than in simulation F. The EM algorithm outperformed KD2 across both simulation and sample sizes.

	Simulation F		Simulation U	
	KD2	EM	KD2	EM
100	0.102	0.092	0.127	0.117
300	0.077	0.070	0.097	0.089
1000	0.058	0.050	0.074	0.065

Table 6. RMSE of estimated EDR in each simulation across a different number of studies

The CI coverage of EDR did not reach the nominal level of 95% and it decreased with increasing number of studies (Table 7). This time a bigger adjustment by five percentage points in each direction was needed to obtain acceptable overall coverage (Table 8) and a good coverage across a wide range of scenarios (Appendix B).

	simulation F		simulation U	
	KD2	EM	KD2	EM
100	0.914	0.903	0.883	0.857
300	0.876	0.890	0.855	0.833
1000	0.745	0.847	0.777	0.791

Table 7. CI coverage of EDR in each simulation across a different number of studies

	simulation F		simulation U	
	KD2	EM	KD2	EM
100	0.986	0.983	0.983	0.980
300	0.988	0.986	0.986	0.979
1000	0.992	0.985	0.989	0.975

Table 8. CI coverage of EDR in each simulation across a different number of studies

Application to Real Data

A large team of psychology researchers replicated 100 studies from three psychology journals to estimate the replicability of published results (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 significant results, including marginally significant ones, only 36 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), but a recent study replicated 10 of the failed replication studies with much larger samples and was able to get significant results in only 2 out of the 10 studies. Not a single study would have produced significant results with the effect size of the replication studies and the sample size of the original studies (Ebersole et al., 2020) These results affirm that replication failures are due to problems with the original studies and not the replication studies.

Z-curve can provide valuable objective information about the replicability of the published results because it relies on the test statistics that were reported in the published articles. Therefore, it avoids the problem of actual replication studies that it is often difficult to reproduce

the original conditions of an experiment. Furthermore, it does not assume that the studies correspond to a single phenomenon, thus allowing to estimate ERR and EDR of highly heterogenous and unrelated studies. We used newly created z-curve package (Bartoš & Schimmack, 2020; accessible at https://cran.r-project.org/web/packages/zcurve/) to fitted z-curve 2.0 to 90 statistically significant results. We excluded 7 studies where the replication results were ambiguous. This did not affect the actual replication rate, 35/90 (38.9%). The EM algorithm results were ERR = .62, 95% CCI [.46, .75] and EDR = .39, 95% CCI [.08, .70]. Figure 5 shows the package generated visualization of the z-curve EM model fit to the observed z-scores. The KD2 results were ERR = .61, 95% CCI [.44, .77], and EDR = .51, 95% CCI [.12 to .73]. Although the EDR confidence intervals are wide, the discovery rate is well below the actual discovery rate even if marginally significant results are treated as statistically non-significant results, ODR = .98, 95% CI [.91, 1]. Thus, z-curve 2.0 provides clear evidence that the original results were selected for statistical significance. As a result, regression to the mean inevitably would make it impossible to replicate all results. It is impossible to draw strong conclusions about the replicability of published results because the sample size is modest. The ERR estimates of 61% or 62% are higher than the actual discovery rate of 39%, but the CCIs are wide and the lower bounds of the CCIs match the upper bound of the CI for the actual discovery rate, 95% CI [.29, .49]. Moreover, ERR assumes that the replication studies are exact copies of the original studies. If, however, replication studies are drawn from a sample of similar studies with varying effect sizes, EDR is a more reasonable estimate of the actual success rate in a set of similar studies. The reason is that the replication studies are actually new studies that are drawn from a population of studies with varying population effect sizes. As the results of these replication studies are not selected for statistical significance, the unconditional mean power of these studies

is lower than the unconditional mean power of studies that were published because they achieved a statistically significant result. Consistent with this logic, EDR estimates (EM-EDR = 39%, KD2-EDR = 51%) are closer to the observed replication rate. To further explore this issue, larger samples of actual replication studies are needed.

Our EDR estimates also provide new insights about the percentage of false positive results that are published in psychology journals. Although the exact rate of false positives cannot be estimated, it is possible to estimate the maximum percentage of false positives on the basis of the discovery rate, using the formula max.FPR = (1/EDR - 1)*(.05/.95) (Sorić, 1989). This formula gives a maximum FDR of 8% for EM and 5% for KD2. However, these estimates are based on the point estimates of the EDR. A true maximum with a 5% error probability is obtained by using the lower bound of the CCIs to compute max.FDR. This yields a max.FDR of 65% for EM and 40% for KD2. While the KD2-based estimate is below 50%, the EM is above the threshold does not allowing to reject the Ioannidis's (2005) claim that most published results are false positives. However, the results still indicate that at least for the studies in the OSC replication project, low power might be a bigger problem than false positives. That being said, the maximum FDR only applies to the definition of false rejections of the null-hypothesis that there is absolutely no effect. It would be much higher if even small and practically non-significant effect sizes would be included in the definition of the null-hypothesis. In this regard, even significant results in replication studies with small effect sizes should not be considered successful replications of original studies with vastly inflated effect sizes (Ebersole et al., 2020).

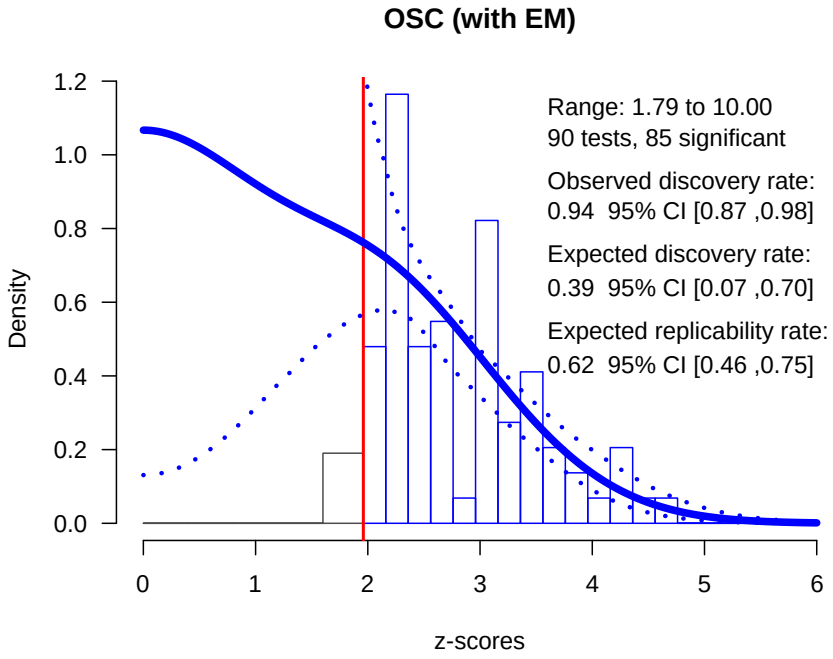

Figure 5. The estimated EM z-curve for the OSC data. The histogram shows the distribution of observed z-statistics with the vertical red line showing the statistical significance criterion. The full blue line displays the density of the estimated model with the dotted line standing for (uncorrected) piece-wise confidence intervals.

Discussion

We extended z-curve 1.0 in multiple ways. First, we examined the performance of z-curve 1.0 in a new set of challenging simulations and compared it to two new estimation methods. We found that all z-curve methods produced robust estimates of the expected replication rate in contrast to the estimates provided by *p*-curve, which produced estimates with high bias and variance, inadequate CI coverage, and was not converging to the true values with increasing number of studies. Second, we evaluated the performance of bootstrapped confidence intervals to provide users with information about the uncertainty of z-curve estimates. We found that even small systematic bias reduced coverage below the nominal level of 95% under some

conditions. We were able to remedy this problem by extending the confidence interval of the ERR by three percentage points. Z-curve 2.0 will use these conservative confidence intervals as the default method. Third, we compared fitting z-curve to a kernel-density distribution of observed z-scores with an expectation-maximization algorithm that is fitted directly to the observed z-scores. Both methods produced similar results, but the results were not identical. Overall, the EM algorithm produced estimates with less bias and lower RMSE. However, the kernel-density methods produced estimates with slightly better coverage across a wider range of scenarios. With small sets of studies, the EM method is faster, especially when confidence intervals are requested. However, computation can take a long time for large sets of studies. In this case, the kernel-density approach reduces the number of data points that need to be fitted and is considerably faster. We provide users with the option to use either method. Typically, both methods yield similar results. When results diverge, the data should be carefully examined and results from both methods should be reported to alert readers that model specifications matter.

The theoretically more important contribution was the extension of z-curve to estimate the expected discovery rate solely based on statistically significant results. To estimate the densities for statistically non-significant values, z-cuve.2.0 used Brunner and Schimmack's (2020) third theorem that relates the population of studies after selection for statistical significance to the population of studies before selection for statistical significance. Our simulation studies showed that z-curve 2.0 is able to provide useful estimated of the expected discovery rate. Although these estimates are sometimes biased, the amount of bias is typically small enough to provide useful information about the discovery rate in a set of published studies. Moreover, across a wide range of scenarios z-curve 2.0 tends to overestimate the discovery rate, which makes it a conservative tool for the assessment of publication bias. On the other hand, it creates an optimistic bias when the EDR is used to estimate the false discovery rate (Sorić, 1989).

We were able to create a conservative confidence interval for the EDR by widening the bootstrapped confidence interval by five percentage-points on each side. These conservative

confidence intervals had good coverage across a wide range of scenarios. Thus, users should focus more on the range of values indicated by the confidence interval than on the point estimate. We demonstrated the usefulness of z-curve 2.0 estimates with studies from the Reproducibility Project (Open Science Collaboration, 2015). For 90 studies the observed discovery rate was 96%, which is consistent with Sterling's findings (Sterling, 1959; Sterling et al., 1995). Z-curve showed that this high discovery rate is inflated by selection for statistical significance. We were also able to estimate the maximum false discovery rate from the EDRs, using Sorić's (1989) formula and found that the maximum false discovery rate is 40% for KD2 and 65% with EM with an error probability of 5%.

Estimating Replicability on the Basis of Original Test Statistics

The past ten years have seen a crisis of confidence in the replicability of published results in scientific journals. The reason for this crisis was that it was nearly impossible to publish statistically non-significant results that failed to replicate original results. This has changed. The past decade has seen a sharp rise in publications of replication failures. Although we welcome initiatives like registered replication reports, we think that actual replication studies alone are unable to solve the replication crisis in psychology for several reasons. First, actual replication studies are costly and have been limited to paradigms that are relatively cheap. The costs to conduct actual replication studies for longitudinal studies or studies with expensive equipment are astronomical and would take away from scientists' ability to investigate new questions.

The second problem is that actual replication studies have failed to create a scientific consensus about the status of important theories. A major concern is that replication studies may have suffered from methodological problems (cf. Schimmack, 2020). We agree that it can be difficult to determine whether inconsistent results are caused by problems in original studies or in replication studies. Z-curve is helpful because it relies on the published results in original studies. Therefore, it avoids the problem of recreating experimental conditions. Using the published results of original articles, we found that only 61% of these studies were expected to produce a statistically significant result again, even if it were possible to recreate the original

studies exactly; with the same sample sizes. Thus, nobody should have been surprised that the actual success rate was not 100%. However, without z-curve estimates no clear predictions could be made about the expected replication rate.

The estimate of 61% makes it possible to ask a new question, namely why the actual success rate was only 39% and not 61%. There are several possible answers to this question. First, there is sampling error in both replication rates and the discrepancy may be smaller than 22 percentage points. Second, it is possible that z-curve estimates are too high because the selection model does not match researchers' practices (John et al., 2012). Examining the performance of z-curve with different types of questionable research practices is one avenue for future research. Third, it is possible that actual replication studies differed in important ways from the original study (Luttrell, Petty, & Xu, 2017).

The most interesting hypothesis is that replication studies differ from the original studies in unknown ways that influence the outcome of similar, yet not identical studies (van Bavel et al., 2016). Z-curves' ERR estimates assume that it is possible to recreate original studies exactly. However, replication studies are never exact copies of original studies. Thus, actual replication studies are more like a random draw from a population of studies. This idea is illustrated by Moorwedge et al.'s description of a research process where researchers try several modifications of a paradigm and publish only the results of studies that produced a statistically significant result (Morewedge, Gilbert, & Wilson, 2014). In this case, a replication study is not an exact replication of the one published study, but another random sample from the population of studies that were attempted. And because the published studies were selected for statistical significance, regression to the mean implies that the power of the replication study will be less than power of the original study. As a result, the success rate of replication studies is better predicted by the expected discovery rate than the expected replication rate. According to this line of reasoning, we would expect as few as 39% statistically significant results based on the EM estimate of the EDR, which is perfectly in line with the actual success rate. Unfortunately, the small sample size of the OSC dataset makes it impossible to test this hypothesis more thoroughly. Once more

results from credible, pre-registered replication studies become available, it will be interesting to compare the ERR and EDR as predictors of the actual success rate in replication studies.

Estimating the Size of the File Drawer

Numerous methods have been developed to examine the presence of publication bias and to correct for the influence of publication bias. A key problem of the existing methods is that their results are not trustworthy under conditions of heterogeneity (Inzlicht, Gervais, & Berkman, 2015; Renkewitz & Keiner, 2019). For example, regression methods cannot distinguish between publication bias and stronger effects in smaller samples. Inzlicht et al.'s critique has raised concerns about the use of bias-correction methods when data are heterogenous (Cunningham & Baumeister, 2016; Inzlicht & Friese; 2019).

With z-curve 2.0, we introduce a method that can provide valid information about the presence and the extent of publication bias under conditions of heterogeneity. The reason is that z-curve 2.0 explicitly models heterogeneity rather than simply assuming that results are robust under conditions of heterogeneity. Furthermore, it does not assume that the conducted studies were studying a single phenomenon, and thus can be used to estimate replicability for whole fields or journals. That opens doors for larger meta scientific questions that can be explored using big datasets, greatly reducing the uncertainty of provided estimates. We demonstrated in extensive simulation studies that z-cuve.2.0 provide useful estimates of the expected discovery rate that can be compared to the observed discovery rate to asses publication bias. We demonstrated the usefulness of these estimates by demonstrating that original studies in the reproducibility project were selected for statistical significance and that for every published study between one or two studies were attempted, but produced statistically non-significant results that were not reported. Given the widespread concerns about other bias-detection methods, we recommend z-curve 2.0 as a valuable tool to ensure that meta-analytic effect size estimates are not inflated by publication bias by comparing the observed discovery rate to the expected discovery rate.

How Many Results are False Positives?

Failed replication studies have sometimes been misinterpreted as evidence that most published results in psychological science are false positives (cf. Brunner & Schimmack, 2020). However, interpreting statistically non-significant results as evidence for the null-hypothesis is a mistake. A statistically non-significant result could also be a type-II error (that is, the effect size is not zero, but the signal-to-noise ratio was too small to be statistically significant). To complicate matters, it is actually impossible to provide positive evidence for the nil-hypothesis (e.g., demonstrate that there are no purple swans on Earth). There is always a small chance that the signal is just so small that we didn't discover it (Cohen, 1994). Thus, trying to quantify the actual percentage of true null-hypothesis is a fools' errand.

However, it is possible to estimate the maximum number of false positives in a set of studies (Sorić, 1989). For example, if we did 100 studies and obtained 100 statistically significant results, the discovery rate (100%) makes it clear that most if not all studies were true positives. After all, an honest test of a true null-hypothesis produces many statistically non-significant results. If, on the other hand, we discover only 5 statistically significant results in 100 tests, the percentage matches the error rate and we would have expected 5 statistically significant results if the null-hypothesis was true in all 100 studies. Thus, all five discoveries could be false positives. Sorić provided a formula that specified the maximum false positive discoveries as a function of the observed discovery rate. This formula only works when we have access to all attempts that were made. However, selection bias inflates the discovery rate and attenuates the estimate of the maximum false discovery rate. As selection bias is pervasive in psychological journals, the observed discovery rate of 90% is misleading. Z-curve 2.0 solves this problem by estimating the actual discovery rate based on the published statistically significant results. The

EDR estimates can then be used with Sorić's formula to get estimates of the maximum false discovery rate. When we applied z-curve 2.0 to the OSC data, we obtained the lower EDR estimate with the EM algorithm. However, even with the EM algorithm and a conservative confidence interval, the lower bound estimate was an EDR of 8%. With this low discovery rate, Sorić's (1989) formula yields a false discovery rate of 65%. Thus, the small sample size and the resulting wide confidence intervals make it impossible to accept or reject Ioannidis's prediction that most published results are false positives. However, with larger sample sizes, z-curve estimates of the discovery rate and Sorić's formula can be used to test this prediction. However, the focus on the point-null hypothesis is a bit misleading. If false positives were defined as a region of population effect sizes close to zero, the percentage of false positives would increase. Thus, we believe that discovery rates and replication rates are ultimately more meaningful than the false discovery rate, especially if it is defined in terms of population effect sizes that are exactly zero.

Conclusion

To summarize, given the widespread practice to select results for statistical significance, statistical significance does not provide information about the replicability of published results. Z-curve was developed to provide this information based on the actual test-statistics published in journals. We show that z-curve 2.0 performs well across a wide range of realistic scenarios and we provide valid bootstrapped confidence intervals, however, we caution against the usage of *p*-curve due to its suboptimal performance. We also validated an estimate of the expected discovery rate that can be used to assess the presence of publication bias and to estimate the maximum false discovery rate. We also created an R-package that makes it possible to conduct z-curve

analyses. We believe that z-curve is a promising method for estimating reproducibility of fields and journals and can provide evidence about the presence and extent of selection bias.

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/.

References

American Psychological Association. (2010). Publication manual of the APA (6th ed.). Washington, DC.

Bartoš, F., Schimmack, U. (2020). "zcurve: An R Package for Fitting Z-curves." R package version 1.0.0

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*, 1884. doi: 10.3389/fpsyg.2019.01884

Brunner, J. (2018). An even better p-curve. https://replicationindex.com/2018/05/10/an-even-better-p-curve/

Brunner, J. & Schimmack, U. (2020). Estimating population mean power under conditions of heterogeneity and selection for significance. *Meta-Psychology, 4*, MP.2018.874, https://doi.org/10.15626/MP.2018.874

Camerer, C. F., Dreber, A., Forsell, E., Ho, T. H., Huber, J., Johannesson, M., ... & Heikensten, E. (2016). Evaluating replicability of laboratory experiments in economics. *Science*, *351*(6280), 1433-1436.

Chang, A. C., & Li, P. (2015). Is economics research replicable? Sixty published papers from thirteen journals say 'usually not'. *Available at SSRN 2669564*.

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

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist, 49*, 997–1003.

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.

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.

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., Lazarević, L. B., Rabagliati, H., Ropovik, I., Aczel, B., Aeschbach, L. F., Andrighetto, L., Arnal, J. D., Arrow, H., Babincak, P., … Nosek, B. A. (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

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.

Eddelbuettel, D., François, R., Allaire, J., Ushey, K., Kou, Q., Russel, N., … Bates, D. (2011). Rcpp: Seamless R and C++ integration. *Journal of Statistical Software, 40(8)*, 1–18.

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

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

Hedges, L. V. (1992). Modeling publication selection effects in meta-analysis. *Statistical Science, 7*(2), 246-255.

Hedges, L. V., & Vevea, J. L. (1996). Estimating effect size under publication bias: Small sample properties and robustness of a random effects selection model. *Journal of Educational and Behavioral Statistics*, *21*(4), 299-332.

Inzlicht, M., & Friese, M. (2019). The past, present, and future of ego depletion. *Social Psychology*.

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. (2005). Why most published research findings are false. *PLoS medicine*, *2*(8), e124. https://doi.org/10.1371/journal.pmed.0020124

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., Lowenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science, 23*, 517–523. 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., Nelson, L., & Nottingham, R. (1961). The folded normal distribution. *Technometrics*, *3*(4), 543–550. https://doi.org/10.1080/00401706.1961.10489974

Luttrell, A., Petty, R. E., & Xu, M. (2017). Replicating and fixing failed replications: The case of need for cognition and argument quality. *Journal of Experimental Social Psychology*, *69*, 178-183. https://doi.org/10.1016/j.jesp.2016.09.006

Maier, M., Bartoš, F., & Wagenmakers, E. (2020). Robust Bayesian meta-analysis: Addressing publication bias with model-averaging. https://doi.org/10.31234/osf.io/u4cns

McShane, B. B., U., and Hansen, K. T. (2016). Adjusting for publication bias in meta-analysis: An evaluation of selection methods and some cautionary notes. *Perspectives on Psychological Science 11*, 730-749. https://doi.org/10.1177/1745691616662243

Morewedge, C. K., Gilbert, D., & Wilson, T. D. (2014). Reply to Francis. Retrieved June 7, 2019, from https://www.semanticscholar.org/paper/REPLY-TO-FRANCIS-Morewedge-Gilbert/019dae0b9cbb3904a671bfb5b2a25521b69ff2cc

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

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

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*. https://doi.org/10.1027/2151-2604/a000386

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin, 86*, 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., & Lakens, D. (2020). An excess of positive results: Comparing the standard Psychology literature with Registered Reports. https://doi.org/10.31234/osf.io/p6e9c

Schimmack, U. (2012). The ironic effect of significant results on the credibility of multiple-study articles. *Psychological Methods, 17*, 551–566. https://doi.org/10.1037/a0029487

Schimmack, U. (2020). A meta-psychological perspective on the decade of replication failures in social psychology. *Canadian Psychology/Psychologie canadienne.* 61 (4), 364-376, DOI:10.1037/cap0000246

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. https://doi.org/10.1037/a0033242

Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2015). Better P-curves: Making P-curve analysis more robust to errors, fraud, and ambitious P-hacking, a Reply to Ulrich and Miller (2015). *Journal of Experimental Psychology: General, 144*(6), 1146–1152. https://doi.org/10.1037/xge0000104

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 decision and the possible effects on inferences drawn from tests of significance – or vice versa. *Journal of the American Statistical Association, 54*, 30–34. https://doi.org/10.2307/2282137

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*, 108–112. https://doi.org/10.2307/2684823

Van Bavel, J. J., Mende-Siedlecki, P., Brady, W. J., & Reinero, D. A. (2016). Contextual sensitivity in scientific reproducibility. PNAS Proceedings of the National Academy of Sciences of the United States of America, 113(23), 6454–6459. https://doi.org/10.1073/pnas.1521897113

Wood, S. (2012). mgcv: Mixed GAM Computation Vehicle with GCV/AIC/REML smoothness estimation.

Wuertz, D. (2014). Rnlminb2: An R Extension Library for Constrained Optimization with nlminb.

Wegner, D. M. (1992). The premature demise of the solo experiment. *Personality and Social Psychology Bulletin*, *18*(4), 504-508. https://doi.org/10.1177/0146167292184017