

Z-Curve:

A Method for the Estimating Replicability Based on Test Statistics in Original Studies

Ulrich Schimmack and Jerry Brunner

University of Toronto Mississauga

Author Note

The work reported in this article is a truly collaborative effort with equal contribution by both authors. This work was supported by a standard research grant of the Canadian Social Sciences and Humanities Research Council (SSHRC) to Ulrich Schimmack. Correspondence should be sent to Ulrich Schimmack, Department of Psychology, University of Toronto Mississauga, email: ulrich.schimmack@utoronto.ca.

Abstract

In recent years, the replicability of original findings published in psychology journals has been questioned. A key concern is that selection for significance inflates observed effect sizes and observed power. If selection bias is severe, replication studies are unlikely to reproduce a significant result. We introduce z-curve as a new method that can estimate the average true power for sets of studies that are selected for significance. We compare this method with p-curve, which has the same aim. Simulation studies show that both methods perform well when all studies have the same power, but p-curve overestimates power if power varies notably across studies. Based on these findings, we recommend z-curve to estimate power for sets of studies that are heterogeneous and selected for significance. Application of z-curve to various datasets suggests that the average replicability of published results in psychology is approximately 50%, but there is substantial heterogeneity and many psychological studies remain underpowered and are likely to produce false negative results. To increase replicability and credibility of published results it is important to reduce selection bias and to increase statistical power.

Keywords: Power estimation, Post-hoc power analysis, Publication bias, P-Curve, Z-curve, Replicability, Simulation, Meta-Analysis.

Z-Curve:

A Method for the Estimating Replicability Based on Test Statistics in Original Studies

Until recently, psychologists were confident that published results are replicable. Meta-analyses typically concluded that sets of studies supported empirical hypotheses. Multiple-study articles often reported three or more successful replication studies (Schimmack, 2012). The success rate of published replication studies is typically very high (Sterling, 1959). In fact, the modal success rate in multiple study articles is 100%. These results gave the impression that psychological theories rest on a foundation of strong empirical evidence.

This impression changed when Bem (2011) published 9 incredible demonstrations that extraverts, but not introverts, can predict random future events above chance levels. Rather than revealing a surprising new human ability, Bem's article unveiled questionable research practices that can produce misleading results (Francis, 2012; Schimmack, 2012). In response to Bem's controversial article, psychologists have become more aware that publication bias undermines the ability of multiple-study articles and meta-analyses to guard against false positive results.

In our opinion, the main problem that plagues psychological science is the selective publishing of significant results based on studies with low statistical power. Methodologists have long known about the negative effects of publication bias (Sterling, 1959). The main problem is that publication bias renders nominal error probabilities (e.g, $p < .05$) meaningless. Rosenthal (1979) pointed out that in the worst-case scenario, the nominal type-I error rate of 5% that applies to all studies that were conducted is consistent with 100% type-I errors in the subset of

studies selected for significance. Another problem is that publication bias inflates observed effect sizes. Thus, even if the original finding was not a false positive result, replication studies may produce much smaller and practically insignificant effect sizes.

We emphasize the importance of low power because publication bias is less of a concern if studies have adequate power. A common recommendation is to plan for 80% power (Cohen, 1988); that is 8 out of ten replication studies would produce a significant result, if the original study produced a true positive result. Even if there were selection bias, replication studies would, on average, still produce 80% significant results. Thus, the actual power of psychological studies is important to evaluate the credibility of published results and to verify original results with actual replication studies.

Cohen (1962) made a first attempt to estimate the average power of studies reported in the *Journal of Abnormal and Social Psychology*. His method yielded a median power of 50% to detect a medium effect size. Power to detect small effect sizes was very low and only large effect sizes could be detected with high probability. In the following decades psychologists have noticed no improvement in statistical power or evidence that psychologists use a priori power analysis to plan sample sizes (Sedlmeier & Giegerenzer, 1989; Schimmack, 2012).

The problem with Cohen's method of examining power is that estimates are based on a priori effect sizes. This method does not provide a direct estimate of the typical power of published studies, which depends on the unknown population effect sizes of these studies. The goal of this article is to introduce a statistical method that can estimate the average power of a set of studies under the most extreme conditions; that is, (a) population effect sizes are unknown, (b) population effect sizes are heterogeneous, (c) the distribution of population effect sizes is unknown, and (d) studies are selected for significance.

Power and Replicability

Replicability is acknowledged to be a requirement of good science (Popper 1934), but it is less clear how replicability should be defined and measured. Replicating something means to copy or reproduce something. In the context of psychological research, replicating a study means to copy or reproduce a previous study. When a replication study is carried out, the study can produce the same result, or it may produce a different result. A replication study that produces the same result is considered a successful replication study. We define replicability as the probability of carrying out a successful replication study.

We can distinguish two factors that influence replicability. One factor concerns the ability to reproduce identical conditions as in an original study. The second factor is sampling error. Even if conditions are identical and samples are drawn from the same population, sampling error will produce different results. This is the main reason, why it is necessary to use sampling distributions and statistics to draw inferences from samples about populations. Without sampling error, results of identical studies would be identical.

Sampling error creates problems for the definition of replicability because no two studies will produce identical results. Thus, some other criterion needs to be used to define a successful replication. The most widely used criterion for a successful replication is statistical significance (Killeen, 2005). This definition goes back to Fisher, who stated that “a properly designed experiment rarely fails to give ... significance” (Fisher, 1926, p. 504). Therefore, it is not sufficient that an original study produced a significant result. Exact replications of the original study should also produce significant results; at least we should observe more successful than failed replications if the hypothesis is true.

Neyman and Pearson (1933) formalized this requirement in their model of inference that distinguishes type-I and type-II errors. The failure to reject a false null-hypothesis (or to accept a true alternative hypothesis) is called a type-II error and the probability of avoiding a type-II error is called statistical power. Thus, a properly designed experiment should have high statistical power because high statistical power ensures that future replication studies will produce a high rate of significant results. Most psychologists have learned that a good experiment should have 80% power (Cohen, 1988). A study with 80% power is expected to produce 4 out of 5 significant results in the long run. If psychological studies had 80% power, it would also justify that up to 80% of published results in psychology journals are successful. Although it is well-known that a priori power should be 80%, the actual power of psychological studies is unknown, although it is unlikely to be 80% (Sterling et al., 1995). The aim of z-curve is to estimate the actual power of psychological studies and to use this estimate to predict the outcome of replication studies.

False Positives and Replicability

It is important to distinguish two reasons for a replication failure. One possible reason is that the original study reported a true positive result and the replication study produced a type-II error (a false negative result). Another reason could be that the original result was a false positive result. Discussions of replication failures often do not clearly distinguish between these two possibilities and create unnecessary confusion. In our opinion it is very difficult and not very productive to estimate the percentage of false positive results in psychology.

One problem is that it is difficult to demonstrate the absence of an effect and attempts to do so require large samples. Another problem is that the distinction has no practical consequences if studies with true positives have very low power. Type-I errors are expected to

produce a significant result with the probability set by the criterion for significance, typically 5%. A true positive result with very low power could have a probability of 6% to produce a significant result. Both studies are likely to produce much more non-significant results than significant ones (94/100 vs. 95/100), and the observed success rates make it impossible to distinguish between false positive and true positive results.

Once we take replicability into account, the distinction between false positives and true positives with low power becomes meaningless, and it is more important to distinguish between studies with good power that are replicable and studies with low power or false positives that are difficult to replicate. A minimum standard for good power is 50% (Tversky & Kahneman, 1971). If power is greater than 50%, a study is more likely to produce a correct result (a true positive result) than an incorrect result (a false negative result).

In conclusion, we agree with Fisher, Tversky and Kahneman, and Cohen that good studies should have high power and we consider 50% power a minimum standard and 80% power a desirable goal for the average power of psychological studies. If studies in psychology met these standards, published results would be replicable, and psychology would not be in a replication crisis that casts doubt on the credibility of all published results.

An Empirical Approach to Estimating Replicability

One way to estimate replicability is to conduct actual replication studies. In response to the replication crisis, several initiatives have pursued this approach. The Many-Labs approach focuses on a single original study that is replicated as closely as possible across several labs (Klein et al., 2014). Ignoring slight variations in sample sizes for the moment, the average success rate across the many labs provides an estimate of replicability because power determines the long-run success rate of exact replication studies. A superior approach would be to conduct a

meta-analysis of the replication studies, use the average effect sizes as an estimate of the population effect size, and use this population effect size and the sample size of the original study to determine its replicability. The main drawback of this approach is that it can only be applied to a limited set of studies and does not provide an estimate of replicability for larger sets of original studies.

A second approach is to pick a set of original studies and conduct one replication study of each study (Open Science Collaboration, 2015). This approach does not provide accurate estimates of replicability for single studies, but the average success rate provides an estimate of the average true power of the original studies. The OSC reproducibility project found that only 36% (35 out of 97) replication studies produced a significant result. This finding raised concerns that the average power of original studies is well below 50%. The study also suggested differences between disciplines. Whereas 50% of results from cognitive psychology could be replicated, the success rate for social psychology was only 25%. This abysmal outcome casts doubt about the replicability of social psychological findings that are used to support social psychological theories and are presented as facts in social psychology textbooks. The low replicability of social psychology may explain why even replication studies with large samples have failed to provide evidence for classic findings like ego-depletion (Hagger et al., 2016), facial feedback effects (Wagenmakers et al., 2016), and social priming effects (Cheung et al., 2016; O'Donnell, Nelson, McLatchie, & Lynott, 2017).

The use of actual replication studies has advantages and disadvantages. The advantage is that it takes sampling error and practical problems of recreating identical conditions into account. A result that can be replicated with high frequency in actual replication studies even under slightly different conditions can be considered robust. The disadvantage of this approach is that

actual replication studies are expensive, time-consuming, and sometimes impossible. Not surprisingly, replication studies have focused on relatively simple paradigms in cognitive and social psychology and the replicability of results in other disciplines is lacking (Tackett et al., 2017).

The use of statistical estimates based on original test results has the advantage that it is relatively inexpensive and can be applied to studies that are difficult to recreate. Thus, it is easy to estimate replicability for large and representative samples of studies. In fact, text scrapping technology makes it possible to obtain estimates from the population of all published articles. Thus, a statistical approach based on published test statistics can complement recent initiatives to estimate replicability with actual replication studies.

Statistical Approaches

Statistical methods for the estimation of replicability are essentially meta-analyses of observed power (Schimmack, 2012, 2015). Statisticians have warned against the use of observed power for a single study because observed power estimates are highly sensitive to sampling error, which makes these estimates essentially meaningless (Hoenig & Heisey, 2001; Schimmack, 2015). However, sampling error decreases as the number of cases increases and meta-analyses of observed power can produce informative estimates of true power (Schimmack, 2015). The main problem for meta-analyses of observed power is that selection for significance inflates observed effect sizes. As observed power is based on observed effect sizes, meta-analyses of studies selected for significance produce inflated estimates of power (Schimmack, 2012). We examine two methods that aim to correct for this selection effect.

P-Curve

Simonsohn, Nelson, and Simmons (2014) developed a statistical method to adjust observed effect sizes for the inflation introduced by selection for significance. Although their focus was on effect sizes, the article also mentions that the method could be used to estimate power. “As with effect sizes, p-curve’s estimate of power will correct for the inflated estimates that arise from the privileged publication of significant results” (p. 676). P-curve estimates true power by predicting observed p-values for all possible true power values and picking the true power that produces the closest fit to the data. This makes p-curve a one-parameter model. The problem of one parameter models is that they have problems when the true power is heterogeneous (Brunner & Schimmack, 2016). However, Simonsohn et al. (2014) suggest that “p-curve is robust to heterogeneity in effect size across studies” (p. 680). To our knowledge, the robustness of p-curve has not been tested. For this reason, we included p-curve in our simulation studies. We used the r-code posted on the p-curve website for our simulations and validated our results against results provided by the online app on the p-curve website (Simonsohn, 2017).

Z-Curve

Z-curve follows traditional meta-analyses by converting all statistical tests into z-scores (Stouffer, Suchman, DeVinney, Star & Williams, 1949; Rosenthal, 1979). The only difference to a traditional meta-analysis is that the sign of z-scores is not meaningful for sets of studies with different research hypotheses. Thus, all z-scores are converted into absolute z-scores. Absolute z-scores provide evidence about the strength of evidence against the standard null-hypothesis that the population effect size is zero. We use z-scores because they can be easily converted into power estimates and because all observed test results can be modeled as a function of a single sampling distribution, namely the standard normal distribution

Z-curve allows for heterogeneity in power by assuming that observed z-scores are obtained from multiple sampling distributions with different means. A standard normal distribution with a mean of 1, which corresponds to 17% power, will mostly produce low z-scores, whereas a standard normal distribution with a mean of 3, corresponding to 85% power, will produce higher z-scores. In real datasets, there may be as many normal distributions as observed z-scores (each study has a different power), but it is possible to approximate the distribution of observed z-scores with a finite number of standard normal distributions. To fit the model to observed z-scores, the model gives different weights to each normal distribution (see R-Code in Supplement for details; also, see Brunner & Schimmack, 2016, for a more technical explanation of z-curve).

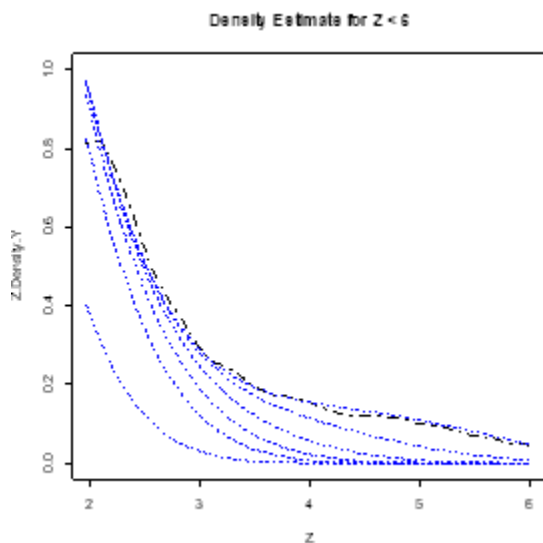


Figure 1 illustrates how z-curve models an observed distribution of absolute z-scores.

The dotted black line in Figure 1 shows the density distribution of observed z-scores between 1.96 ($p < .05$, two-tailed) and 6. The value of 6 is arbitrary, but it is unnecessary to fit the distribution to z-scores greater than 6 because power for these z-scores is essentially 1. Z-curve only fits z-values below 6 and then adjusted the power estimate by the proportion of z-scores greater than 6. The density distribution is composed of seven normal distributions with means ranging from 0 to 6. The bottom blue line shows the contribution of the normal distribution centered over 0. Because there are no negative values, this is a half-normal distribution. The second line from the bottom shows the contribution of the normal distributions for means 0 and 1. The additional area not covered by the area for a mean of 0 shows the contribution of the normal distribution centered at 1. The size of each area is determined by the weight given to each of the seven standard normal distributions. The weights for the model in Figure 1 are 17% for $m = 0$ (5% power), 29% for $m = 1$ (17% power), 14% for $m = 2$ (52% power), 12% for $m = 3$ (85% power), 14% for $m = 4$ (98% power), 14% for $m = 5$ (99% power), and 0% for $m = 6$. The true power for the seven normal distributions is a simple function of the area under the curve in the tails of the criterion value that corresponds to a two-tailed test with $\alpha = .05$.

$$\text{Power} = 1 - \text{pnorm}(1.96, m) + \text{pnorm}(-1.96, m)$$

The average power implied by the observed density distribution is the weighted average of the seven power values

$$100 * (.17*.05 + .29*.17 + .14*.52 + .12*.85 + .14*.98 + .14*.99 + 0*.99.99) = 50\%.$$

We used z-curve to estimate average power for this example ($k = 10,000$). It produced the following estimated weights: 6% for $m = 0$ (5% power), 51% for $m = 1$ (17% power), 0% for $m = 2$ (52% power), 17% for $m = 3$ (85% power), 14% for $m = 4$ (98% power), 12% for $m = 5$

(99% power), and 0% for $m = 6$. A comparison of the actual and estimated weights shows some notable differences. For example, it is not possible to infer from the 6% estimate for $m = 0$ that 6% of the studies were false positives because the true percentage of false positives in this demonstration was 17%. However, the average power estimate was 50%, demonstrating large sample accuracy of z-curve for this example. In short, the example demonstrates how z-curve estimates average power. It approximates an observed density distribution of significant z-scores (less than 6) with a finite set of standard normal distributions. Although the estimated weights of each component do not correspond to the actual distribution of power, the weighted average of the components can be a good estimate of average true power. We conducted our simulation studies to examine the robustness of zcurve estimates under different conditions. In contrast to the good zcurve estimate, pcurve overestimates power in this example by 25% (estimated average power = 76%). Although the amount of bias may be less in our simulations, we predict that zcurve will perform better than pcurve when there is substantial heterogeneity (see also Brunner & Schimmack, 2016).

Simulation Study 1

Our first simulation closely followed Simonsohn et al.'s (2014) simulation of heterogeneity. Standardized population effect sizes varied in steps of $d = .2$ from 0 to .8. The standard deviation of population effect sizes was .2. Sample sizes ranged from $N = 10$ to $N = 70$. The article did not mention the distribution of sample sizes. We decided to use a uniform distribution. Simonsohn et al. (2014) showed that pcurve produces accurate estimates of effect sizes after selection for significance, even when there is variability in population effect sizes. They implied that this finding can be generalized to power. However, Brunner and Schimmack (2016) found that pcurve overestimates power when there is heterogeneity. To reconcile these

conflicting views, we reproduced Simonsohn et al.'s simulations, but estimated average power instead of effect sizes. Based on Brunner and Schimmack's simulations we expected that pcurve would overestimate power, especially for larger effect sizes because larger effect sizes produce more heterogeneity in power.

To simulate power, we computed non-central t-values for each pair of randomly generated effect sizes and sampling size (see Supplement for R-Code). We then simulated observed t-values by randomly sampling from the corresponding non-central t-distribution. To compute true average power, we converted the non-central t-values into power (using $p < .05$, two-tailed as criterion) and averaged power of cases with significant observed t-values. We used sets of 100 studies ($k = 100$) for our simulations. We converted t-values into F-values to estimate power with pcurve. We converted t-values into p-values and then into z-scores to estimate power with z-curve. We ran 5,000 simulations and computed the average estimates of p-curve and z-curve.

Figure 2 shows the results. As predicted, pcurve produced good estimates for small effects, but overestimated power for larger effect sizes. This simulation shows that we cannot generalize from effect size estimation to power estimation. Whereas pcurve provides good estimates of effect sizes, it tends to produce inflated estimates of power in simulations with larger effect sizes and higher average power. Both pcurve and zcurve overestimated power in the simulation with very low power.

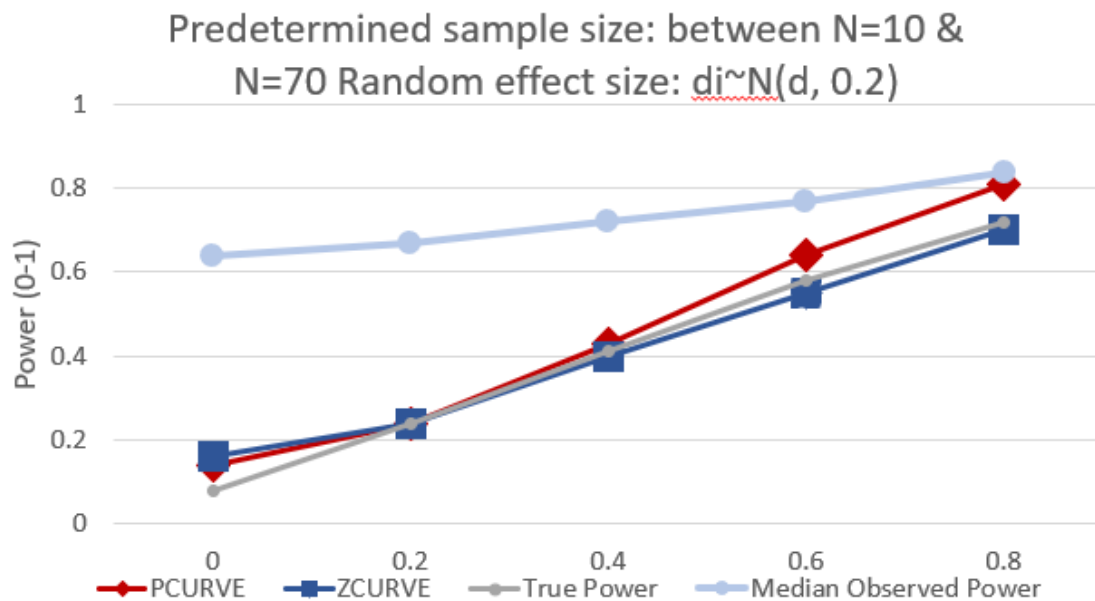


Figure 2. Results of Simulation Study 1.

Simulation Study 2

The main goal of our second simulation study was to manipulate power and heterogeneity more systematically. A second goal was to increase the amount of variability. As Simonsohn et al. (2014), we think it is useful to estimate the average power of studies published in a journal or in a diverse set of studies (Cohen, 1992; Giegerenzer et al., 1989). We would expect considerable heterogeneity in studies of different topics published in the same journal (Schimmack, 2017). In comparison, the amount of heterogeneity in the first simulation study was relatively small.

We used a 3 x 3 design for our simulation study. One factor varied the average true power with three levels representing low (.30), medium (.50) and high (.80) power. The other

factor varied the distribution of true power. One condition simulated data without heterogeneity (i.e., all studies had the same power). The second condition simulated heterogeneity with a normal distribution of z-scores, and the third condition simulated heterogeneity with a skewed distribution of z-scores. As in Simulation Study 1, we used a fixed set of $k = 100$ studies for all simulations. For each condition, we ran 5,000 simulations. In our second simulation study, we used Z values as the observed test statistics. The use of Z-scores does not give z-curve an advantage because p-curve also allows Z values as test statistics and extensive simulations with a variety of test statistics (F-values, chi-square) have shown that simulations with different test statistics typically produce similar results (Brunner & Schimmack, 2016).

Table 1. Results of simulation studies for P-Curve and Z-Curve

Power/ Distribution	Variance (Sig.Obs.Z)	ZCURVE Mean Est.	PCURVE Mean Est	Mean Difference
30%		.33 (.08)	.30 (.07)	-.03
Fixed	0.26	.33 (.08)	.30 (.07)	-.03
Normal	0.32	.31 (.07)	.30 (.07)	-.01
Skewed	0.90	.31 (.07)	.43 (.10)	+.12
Fixed	0.36	.50 (.07)	.50 (.07)	.00
Normal	0.52	.51 (.07)	.56 (.07)	+.05
Skewed	1.83	.51 (.07)	.74 (.07)	+.23
Fixed	0.58	.78 (.05)	.80 (.05)	-.02
Normal	1.00	.80 (.05)	.89 (.03)	+.09
Skewed	2.30	.80 (.05)	.97 (.01)	+.17

Table 1 shows that p-curve performs slightly better than z-curve when all studies have the same power. However, with heterogeneous data, z-curve produces estimates close to the true value. In contrast, p-curve overestimates true power, and this bias increases as the variance in observed z-scores increases. The standard deviations of the two methods are similar, but zcurve

estimates are a bit more variable because a model with multiple components increases variability in estimates.

Small differences in power are not very meaningful, but we think estimates that are more than 10 percentage points different from the true score are undesirable. Thus, we compared the percentages of estimates that are within 10 percentage points of true power (Figure 3).

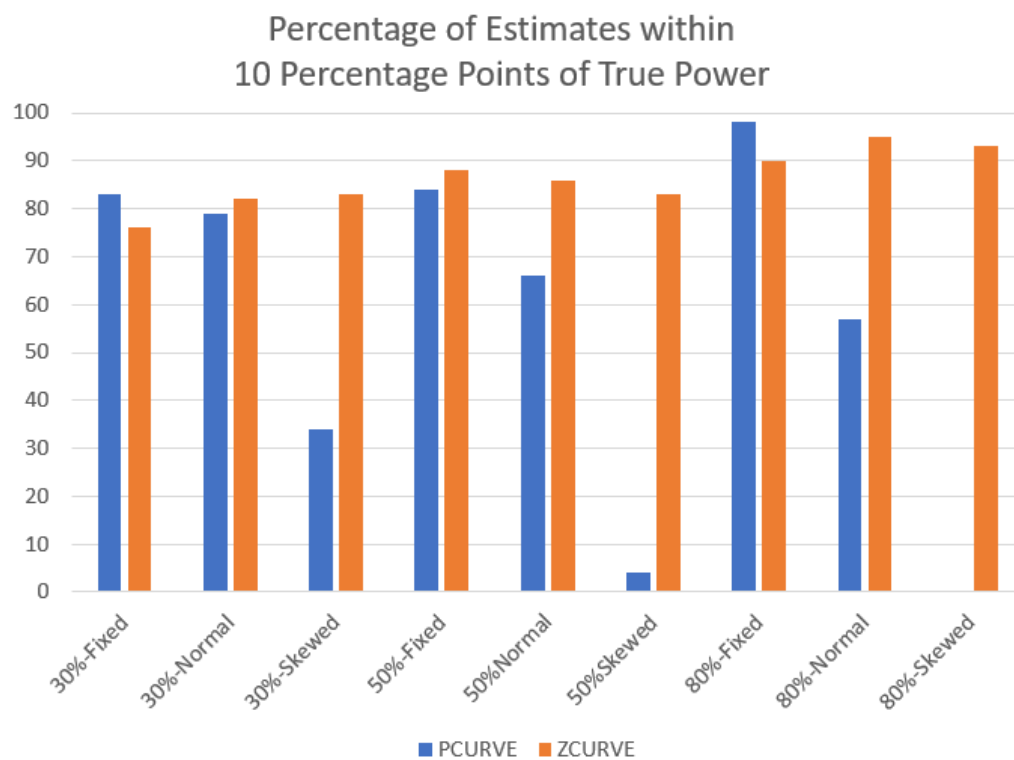


Figure 3. Percentage of Estimates within 10 Percentage Points of True Power

Z-curve did reasonably well in all conditions. P-curve did well with little heterogeneity, but not so well with high heterogeneity. When power was high (80%) and the distribution of non-centrality parameters was skewed, the success rate of P-Curve to produce estimates between 70% and 90% was zero and the average estimate was 97%.

In conclusion, the simulation results show that p-curve produces accurate estimates when variability is low, but P-curve estimates can be severely inflated when there is substantial variability in the observed z-scores, which reflects heterogeneity in actual power. In contrast, z-curve is not affected by heterogeneity or the distribution of true power and produced accurate estimates even when heterogeneity is high, and the distribution is skewed. However, pcurve performs better than zcurve when the data are generated by a fixed value of power. Thus, if variability in observed z-scores is low, pcurve may produce better estimates, although it is difficult to determine heterogeneity when there is selection bias and the number of studies is small.

Application to Actual Test Statistics

Demonstration 1: A Meta-Analysis of Power Posing Effects

Several published articles have used the results of p-curve to draw inferences about replicability. Simmons and Simonsohn (2017) used p-curve to question the credibility of studies that demonstrate an effect of power-posing (i.e., posing in a powerful stance for a brief time can instill feelings of confidence & improve performance). Simmons and Simonsohn's p-curve analysis suggested that published studies provide no evidence for this hypothesis after taking selection bias into account. In response, Cuddy, Schultz, and Fosse (2017) reported the results of a more extensive p-curve analysis. They reported a power estimate of 44% with a 90% confidence interval ranging from 23% to 63%. We retrieved the data from the OSF depository to reproduce the p-curve result and to obtain an estimate using z-curve. We reproduced the 44% estimate with the online app and the p-curve r-code. Next, we converted the test statistics into absolute z-scores and modeled the absolute z-scores with z-curve. Figure 4 shows the distribution of z-scores and the result.

Unlike plots of p-values, the histogram of z-scores makes it easy to see the presence of publication bias or the use of questionable research practices (John, Loewenstein, & Prelec, 2012), which both produce unrealistic sampling distributions. The histogram of absolute z-scores shows a steep drop of observed z-scores around the criterion for statistical significance ($z = 1.96$, $p < .05$, two-tailed). Random sampling error cannot produce this drop. Based on the distribution of significant z-scores ($z > 1.96$), z-curve produced an estimate of 30% replicability; that is, average power of significant results. The z-curve estimate is notably lower than the p-curve estimate of 44%. The large amount of heterogeneity explains this finding; the variance of the 44 significant z-scores was 1.02. Figure 2 shows that most studies are just significant, but a few studies reported strong evidence ($z > 4$). Excluding the four studies with extreme z-scores reduces the variance to 0.20. With low power and low variability, p-curve is likely to produce an accurate estimate. Excluding the four studies with extreme scores had relatively little effect on z-curve (Replicability estimate = 32%). In contrast, the p-curve estimate dropped from 44% (90%CI = 23% to 63%) to 13% (90%CI 5% to 30%). This large drop further suggests that the pcurve estimate of 44% was inflated because pcurve overestimates average power when there is substantial heterogeneity.

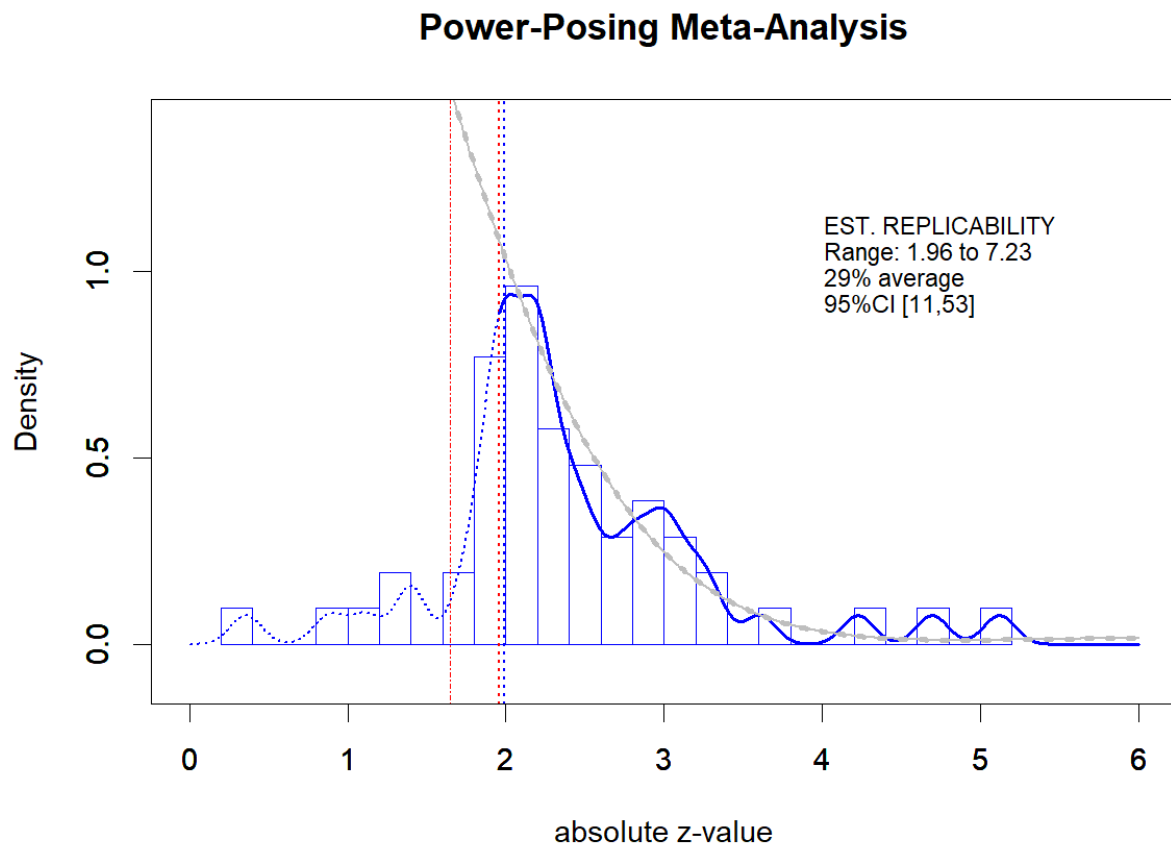


Figure 4. Z-Curve of Power-Posing Studies ($k = 53$)

In conclusion, we do not agree with Cuddy et al. (2017) that p-curve results “reveal strong evidential value for postural feedback effects (i.e., “power posing”). We raise two concerns about this conclusion. First, p-curve produces inflated estimates of power when heterogeneity is present. Z-curve does not have this problem and our z-curve estimate of 30% is considerably lower than the reported p-curve estimate of 44%. Second, we do not agree that studies with low average power can support theoretical predictions, especially when there is heterogeneity across studies. The problem is that average power of 30% can include several studies with very low power that are difficult to replicate, and the average does not provide information about the replicability of individual studies. Thus, it remains unclear which claims

about power posing are true and which effects may be false. At best, we can say that some power posing studies had effects on some measured outcome, but we do not know how many studies are replicable or which outcomes were affected. In this regard, power posing is no exception. We merely focused on power posing because Cuddy et al. used p-curve to draw inferences about evidence and the reported p-curve estimate of average true power is likely to be inflated by substantial heterogeneity due to four studies with extreme values.

Demonstration 2: Replicability of Psychology

There is great uncertainty about the replicability of psychological results (Motyl et al., 2016). The simulation studies showed that z-curve can produce accurate estimates of replicability, especially if the set of studies is large. To provide an estimate of replicability for psychology in general, we extracted test statistics published in 139 psychology journals in the years from 2010 to 2017. We downloaded all articles as PDF files and converted them to text files. We wrote a program in R to extract F-tests and t-tests that were reported in the results section ($F(x,xx) = X.XX$, $t(xx) = X.XX$). The search yielded 995,654 test statistics and 64% ($k = 639,429$) were significant using $\alpha = .05$ (two-tailed). Figure 5 shows the distribution of z-scores. The shape of the distribution shows that there is substantial heterogeneity ($\text{Var} = 3.33$) with a long tail of highly significant results that exceed the stringent 5-sigma criterion in particle physics (cf. Schimmack, 2012). However, the figure also shows that the mode of the distribution is at the criterion for statistical significance. The distribution of non-significant results is not consistent with a plausible sampling distribution. This pattern reveals publication bias, the use of questionable research practices, or both. Given the clear evidence of publication bias, non-significant results are uninformative. The more important question is how strong the evidence for significant results is that were reported as evidence for theoretical claims.

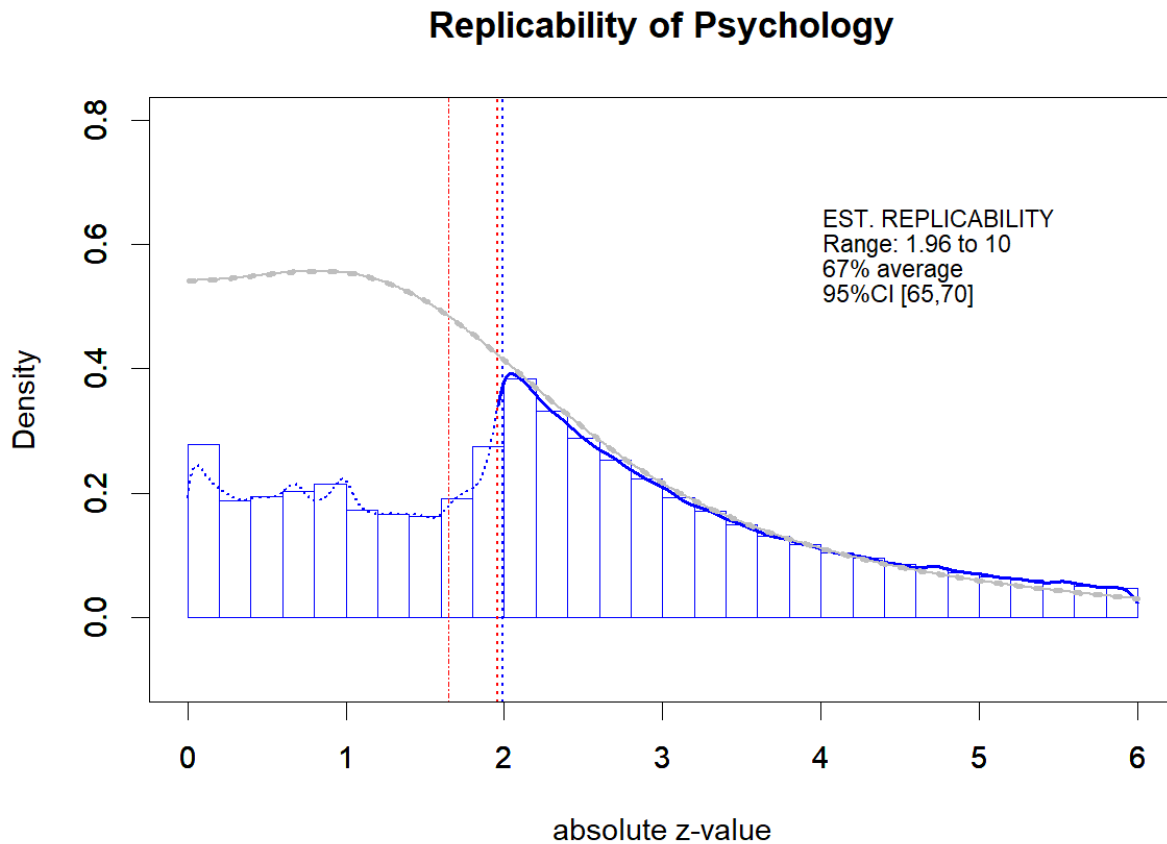


Figure 5. Z-Curve of Test Statistics in 140 Psychology Journals (2010-2017)

The z-curve estimate of replicability of published significant results was 68%. Given the large number of test statistics, the 95% confidence interval around this estimate is very tight and ranged from 65% to 70%. Given the large variability in significant z-scores ($VAR = 3.33$), it is not surprising that the pcurve estimate (83%) was notably higher. Our simulations suggest that this estimate is considerably inflated. Even the zcurve estimate of 67% is surprisingly high in comparison to the 36% successful actual replications in the OSC reproducibility project. Several factors may contribute to this large discrepancy.

In the OSC project, social psychology was overrepresented and social psychology was less replicable than cognitive psychology. According to this hypothesis, the replication crisis is much more severe in social psychology than in other disciplines. A second explanation could be that z-curve assumes exact replication studies and that the actual replication studies in the OSC project failed to reproduce the original conditions exactly. A third hypothesis is that the automated extraction method included test statistics for trivial hypotheses tests such as manipulation checks, whereas the OSC reproducibility project focused on novel theoretical predictions. According to this hypothesis, the replicability of novel and theoretically important hypothesis would be lower. We test this hypothesis in our third demonstration.

Demonstration 3: Replicability of Focal Tests in Social Psychology

Motyl et al. (2017) examined the replicability of social psychology. They randomly sampled articles from major social psychology journals. They focused on the years 2003/04 and 2013/14 to examine possible changes in replicability over time. For each study, they picked a focal hypothesis test and recorded the test statistic. The authors used the R-Index (Schimmack, 2014) to gage the replicability of social psychology. They obtained scores of 62 for the year 2003/2004 and 52 for the years 2013/2014, suggesting no improvement in replicability over time. This finding is consistent with Schimmack's replicability rankings of psychology journals, which show no changes in replicability from 2010 to 2015 (Schimmack, 2017).

For our demonstration, we only used studies that reported t and F tests and converted t-values into F-values. Further, we excluded studies with very large samples ($N > 1000$), very large F-values (> 100), and a large number of experimenter degrees of freedom ($df1 > 10$). This left 974 focal hypotheses. The histogram of z-scores shows clear evidence of publication bias (Figure 6). The variance of significant z-scores was high ($Var = 2.30$). The Pcurve estimate of

average power was 77%, 90CI = [74,80]. Although social psychologists might be thrilled by this finding, our simulation studies suggest that pcurve estimates are inflated with this amount of heterogeneity. Indeed, the zcurve estimate is considerably lower, .46, 95%CI = [.39,.54]. The zcurve estimate of focal tests is also considerably lower than the average power for the automatic extraction method which does not distinguish focal and non-focal tests. This finding suggests that it is necessary to identify focal hypothesis tests to estimate the replicability of novel and theoretically important results in original research articles.

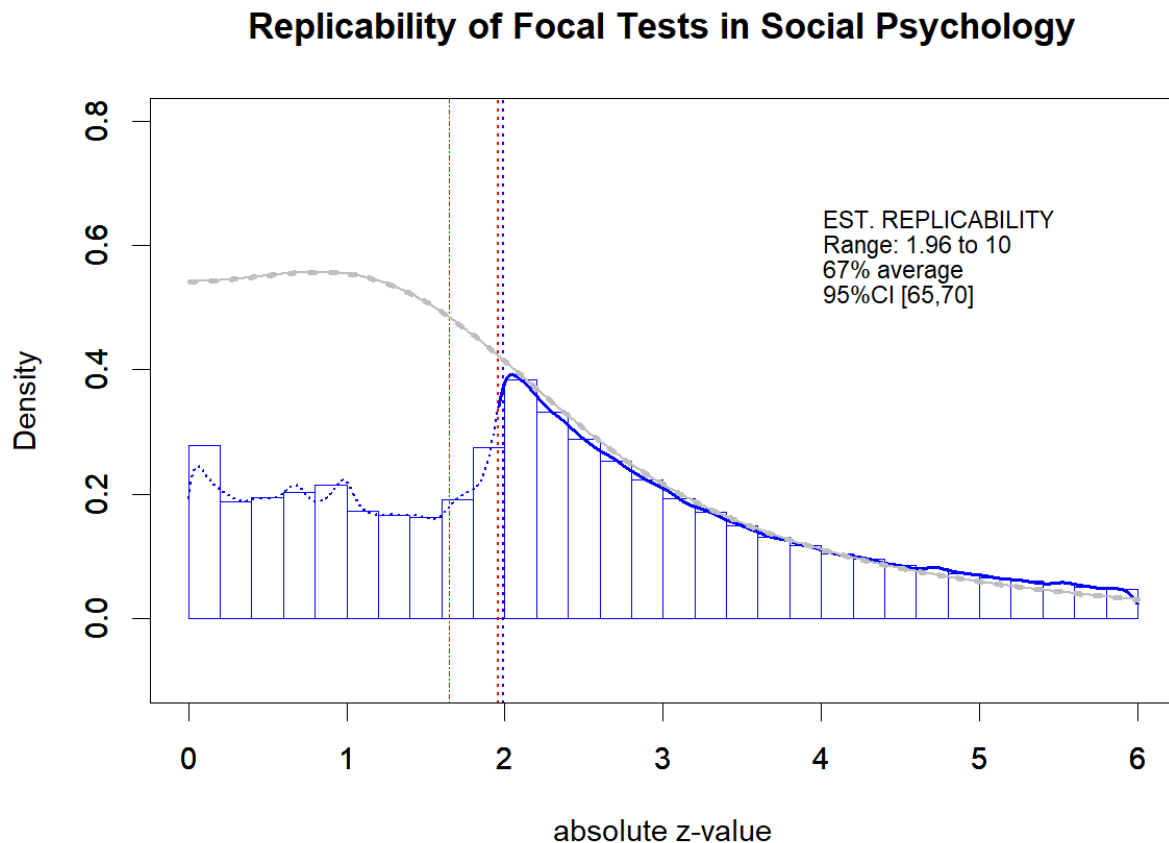


Figure 6. Z-Curve of Motly et al.'s Focal Tests in Social Psychology

The estimate of 46% replicability has several implications. First, the estimate is not as bad as many may have feared (cf. Motyl et al., 2016). It is unlikely that most published results in social psychology are false positive results. Although we cannot determine the number of false positives, 47% average power implies that most published results are not false positives because we would expect 52.5% replicability if 50% of studies were false positives and the other 50% of studies had 100% power. However, the distribution of z-scores in Figure 4 shows that it is unreasonable to assume that half the studies had 100% power. Thus, the false positive rate is likely to be less than 50%.

At the same time, the estimate of 47% implies that the typical study in social psychology falls short of the minimum standard of 50% power and most studies do not meet the textbook standard of 80% power. Based on the present results, social psychologists need to improve the power of their studies to increase replicability and credibility of published findings.

General Discussion

The main goal of this article was to introduce and evaluate a new statistical method, z-curve, to estimate the average replicability of a set of studies. A secondary goal was to compare this method to an existing one, p-curve. Our simulation studies demonstrated that z-curve performs well under many different scenarios, whereas p-curve performs well when the studies are homogeneous, but not when there is substantial heterogeneity. With increasing heterogeneity, p-curve tends to overestimate average power and replicability.

Our first demonstration showed that this bias has practical consequences. A recent meta-analysis of the power-posing effect with p-curve yielded an estimate of 44% power. The z-curve estimate was substantially lower (30%). After removing extreme values that produced substantial heterogeneity, the pcurve estimate decreased by 31 percentage points to 13%. Thus,

it is important to examine the extent of heterogeneity before using p-curve. We suggest to convert p-values into z-scores and to compute the variance of significant z-scores as one method to evaluate heterogeneity. A histogram of z-scores can also provide valuable information about the presence of extreme z-scores that produce substantial heterogeneity and inflate p-curve estimates. When this diagnostic test reveals substantial heterogeneity, zcurve is likely to produce more accurate estimates. When variability is low, pcurve can produce more precise estimates. Thus, it is also a good option to evaluate data with both methods and to carefully examine the data when the methods produce notably divergent results. We also recommend that replicability estimates obtained with z-curve are reported with a 95% confidence interval because point estimates are not very precise unless the set of studies is large (Brunner & Schimmack, 2016).

Our second demonstration applied z-curve to a large set of test statistics reported in 104 psychology journals that cover a broad range of disciplines. We estimated that the average power was 67%. This finding would not justify the notion of a replicability crisis in psychology. However, the estimate is based on all test statistics that are reported in an article, including manipulation checks, and does not provide an estimate of the replicability of theoretically important, novel findings.

Our third demonstration showed that replicability for focal hypothesis tests that are used to support novel and theoretically important predictions is lower. The replicability of focal hypothesis tests was only 47%. This estimate is limited to social psychology and estimates for psychology in general might be somewhat higher. Nevertheless, the discrepancy shows that theoretically new predictions are often tested in studies with low power. When these studies produce a significant result, effect sizes are inflated, and future studies have difficulties

replicating these inflated results. Thus, social psychologists need to increase statistical power so that theories rest on replicable findings.

How Replicable is Psychology?

Our estimates provide valuable information about the extent of the replication crisis in psychology. Based on our results, we think it is unlikely that most published results in psychology are false positives, in the strict sense that the population effect size is zero. At the same time, our results suggest that the majority of studies in psychology fail to meet the minimum standard of a good study; that is, it should have a 50% chance to produce a true positive result when the hypothesis is true (Tvesky & Kahneman, 1971) and even more studies fail to meet the well-known and accepted norm that studies should have 80% power (Cohen, 1988). Our analysis across disciplines suggests that this is not merely a problem of social psychology, but a problem of many areas in psychology. Z-curve can be used to assess the extent of this problem and examine whether recent reforms in psychological publishing are effective in reducing publication bias and increasing replicability.

Limitations

Z-curve has several limitations that can affect its estimates. One problem is that the density distribution is not very robust for small samples. Another problem is that the transformation to z-scores can introduce some bias for studies that have very small sample sizes. Finally, z-curve tends to have some systematic bias when heterogeneity is small. As our simulations showed these limitations are not a problem when 100 or more studies are available and when there is some heterogeneity in the data.

A more important limitation is that our simulations assumed an equal selection criterion for all studies (so does p-curve). This assumption can be violated, if, for example, a study

corrected p-values for multiple comparisons. Future research needs to examine how estimates are affected by varying selection criteria. However, the distribution of z-curve suggests that there is no other sharp selection criterion ($p < .01$). Otherwise, we would see a drop of z-values below this criterion as we see for $p < .05$.

Another concern is that z-curve adjusts estimates for selection effects, but not for the use of questionable research practices. Future research needs to examine how different questionable research practices influence z-curve estimates. Some practices may lead to an underestimation of average power. This could be considered a limitation of z-curve. On the other hand, it can also be considered a conservative bias that is justified because the influence of questionable research practices on replicability is difficult to predict. If questionable research practices lead to lower estimates, it may even act as a deterrent against the use of these practices.

Future Directions

We see a few future directions for the development of z-curve. First, it may be of interest to estimate the average power before the selection for significance. As studies with significant results, on average, have higher statistical power than studies with non-significant results, average power of studies before selection for significance is bound to be lower than the average power of studies selected for significance. However, estimating average power before selection may be a difficult statistical problem because it requires an estimate of the size of the file-drawer (unpublished, non-significant studies).

We are also working on validating estimates for subsets of significant results. For example, it can be of interest to estimate the average power of studies that produced just significant results (e.g., $p < .05$ & $p > .01$). Even with average power of 50%, power for just significant results can be considerably lower and would suggest that these results are difficult to

replicate. Finally, z-curve and p-curve assume that all test statistics are independent. Future research needs to examine how robust z-curve estimates are to violations of this assumption and whether it is possible to develop a method for nested data (multiple test statistics nested within studies).

Conclusion

In conclusion, methodologists have warned about publication bias and low statistical power for decades (Cohen, 1962; Sterling, 1959). However, until recently empirical researchers assumed that these problems were minor and could be ignored. This perception changed and psychologists, at least social psychologists, have wondered about the stability of the empirical foundations of their field. Z-curve provides an opportunity to add some empirical evidence to debates about the replicability of psychological findings. Our statistical approach cannot replace actual replication studies. Actual replication studies are still needed to provide convergent evidence across independent labs and to ensure that published results are not unique to specific historical or situational factors. Our statistical estimates assume that it is possible to replicate original studies exactly. If variation in the historic or situational context changes results, replicability is bound to be lower. This may explain why we obtained an estimate of 47% for social psychology, while the OSC reproducibility project could only replicate 25% of original studies. If this is the case, it is even more important to raise power to 80% to ensure that actual replication studies have a success rate greater than 50%. We are optimistic that recent awareness about the extent of the problem in social psychology will have positive effects on replicability. Our statistical method of estimating replicability makes it possible to examine whether our optimism is warranted.

References

- Bem, D. J. (2011). Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology*, 100(3), 407-425. <http://dx.doi.org/10.1037/a0021524>
- Brunner, J. and Schimmack, U. (2016). How replicable is psychology? A comparison of four methods of estimating replicability on the basis of test statistics in original studies. <http://www.utstat.utoronto.ca/~brunner/zcurve2016/HowReplicable.pdf>
- Cheung et al. (2016). Registered Replication Report: Study 1 From Finkel, Rusbult, Kumashiro, & Hannon (2002). *Perspectives on Psychological Science*, 11, 750-764.
- Cohen J. (1962). The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal and Social Psychology*, 65, 145–152.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. (2nd Edition), Hillsdale, New Jersey: Erlbaum.
- Cuddy, A. J., Schultz, S. J., & Fosse, N. E. (2017). P-curving A More Comprehensive Body of Research on Postural Feedback Reveals Clear Evidential Value For “Power Posing” Effects: Reply to Simmons and Simonsohn. *Psychological Science*. Forthcoming.
- Fisher, R. A. (1926). The arrangement of field experiments. *Journal of the Ministry of Agriculture of Great Britain*, 33, 503–513.
- Francis, G. (2012b). Too good to be true: Publication bias in two prominent studies from experimental psychology. *Psychonomic Bulletin & Review*, 19, 151–156.
[doi:10.3758/s13423-012-0227-9](https://doi.org/10.3758/s13423-012-0227-9)

- Hagger M. S., Chatzisarantis N. L., Alberts H., Anggono C. O., Batailler C., Birt A., Zwieneberg M. (2015). A multi-lab pre-registered replication of the ego-depletion effect. *Perspectives on Psychological Science*, 11, 546–573.
- Hoenig, J. M. and Heisey, D.M (2001). The abuse of power: The pervasive fallacy of power calculations for data analysis. *The American Statistician* 55, 19-24.
- John L. K., Loewenstein G., Prelec D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, 23, 524–532.
doi:10.1177/0956797611430953
- Killeen, P. R. (2005). An alternative to null-hypothesis significance tests. *Psychological Science*, 16, 345-353. <https://doi.org/10.1111/j.0956-7976.2005.01538.x>
- Klein R. A. et al. (2014). Investigating variation in replicability: A “many labs” replication project. *Social Psychology*, 45, 142–152. doi:10.1027/1864-9335/a000178
- Motyl, M. et al. (2016). 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, 34-58. doi: 10.1037/pspa0000084.
- Neyman, J. and Pearson, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society, Series A* 231, 289337.
- Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251), aac4716.
- O'Donnell, M., Nelson, L., McLatchie, N. M., & Lynott, D. J. (2017). Perspectives on Psychological Science.

Popper, K. R. (1959). The logic of scientific discovery. English translation by Popper of Logik der Forschung (1934). London: Hutchinson.

Rosenthal R. (1979). The file drawer problem and tolerance for null results. Psychological Bulletin, 86, 638–641.

Schimmack U. (2012). The ironic effect of significant results on the credibility of multiple-study articles. Psychological Methods, 17, 551–566

Schimmack (2015) Meta-analysis of observed power: Comparison of estimation methods.
<https://replicationindex.wordpress.com/2015/04/01/meta-analysis-of-observed-power-comparison-of-estimation-methods/>

Schimmack, U. (2016). A revised introduction to the R-Index.
<https://replicationindex.wordpress.com/2016/01/31/a-revised-introduction-to-the-r-index/>

Schimmack, U. (2017). Preliminary 2017 replicability rankings of 104 psychology journals.
<https://replicationindex.wordpress.com/2017/10/24/preliminary-2017-replicability-rankings-of-104-psychology-journals/>

Sedlmeier P., Gigerenzer G. (1989). Do studies of statistical power have an effect on the power of studies? Psychological Bulletin, 105, 309–316.

Simonsohn, U. (2017). P-Curve online app code. http://p-curve.com/app4/pcurve_app4.052.r.

Simonsohn, U., Nelson, L. D. and Simmons, J. P. (2014). p-Curve and effect size: Correcting for publication bias using only significant results. Perspectives on Psychological Science, 9, 666-681.

Simmons, J. P. & Simonsohn (2017). Power Posing: P-curving the evidence. Psychological Science, 687-693.

- 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.
- 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. *American Statistician*, 49, 108–112. doi:10.2307/2684823
- Stouffer, S. A., Suchman, E. A , DeVinney, L.C., Star, S.A., & Williams, R.M. Jr. (1949). *The American Soldier, Vol.1: Adjustment during Army Life*. Princeton University Press, Princeton.
- Tackett et al. (2017). It's time to broaden the replicability conversation: Thoughts for and from clinical psychological science. *Perspectives on Psychological Science*, 12, 742-756.
<https://doi.org/10.1177/1745691617690042>
- Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin*, 76(2), 105-110. <http://dx.doi.org/10.1037/h0031322>
- Wagenmakers, E.J. et al. (2016). Registered Replication Report: Strack, Martin, & Stepper (1988). *Perspectives on Psychological Science*, 11, 917-928.
<https://doi.org/10.1177/1745691616674458>

SUPPLEMENT

R-Code

```
rm(list = ls())
```

```
#####
#### PCURVE FUNCTIONS BEGIN
#####
```

```
#### Pcurve.Power.Estimatation (modified code from Uri.Simonsohn)
```

```
#Pcurve.Power.Estimatation(family,value,df1,df2)
```

```
### Input
```

```
# family "f" or "c" for F-test or Chi-Square Test
# value F-value or Chi-Square value
# df1 degrees of freedom for F,chi-square test
# df2 degrees of freedom for F-test
```

```
### Output
```

```
# Power.Estimate
# Lower Bound of 95%CI
# Upper Bound of 95%CI
```

```
#####
# Function to get Non-centrality Parameters
#####
```

```
### Thees functions are needed to find the corresponding non-centrality parameter for a specific power value.
```

```
### This is more difficult for asymmetrical F and chi-square distributions than for symmetrical z-curve
```

```
#F-test
```

```
getncp.f = function(df1,df2, power) {
  criterion.f = qf(p=.95, df1=df1,df2=df2)
  error = function(ncp_est, power, criterion.f, df1,df2) pf(criterion.f, df1 = df1, df2 = df2, ncp =
ncp_est) - (1-power)
  res = uniroot(error, c(0, 1000), power=power, criterion.f = criterion.f, df1 = df1,df2 = df2)$root
  return(res) }
```

```
#chisq-test
```

```
getncp.c = function(df, power) {
  xc=qchisq(p=.95, df=df)
  error = function(ncp_est, power, x, df) pchisq(x, df = df, ncp = ncp_est) - (1-power)
  res = uniroot(error, c(0, 1000), x = xc, df = df, power=power)$root
```

```

    return(res) }

#Combine both in single function
getncp=function(family,df1,df2,power) {
  if (family == "f") ncp=getncp.f(df1=df1,df2=df2,power=power)
  if (family == "c") ncp=getncp.c(df=df1,power=power)
  return(ncp) }

#####
# POWER ESTIMATE
#####

### Powerfit is used to minimize the discrepancy between the observed data and
### predicted values based on a fixed power value and df of data.
### Powerfit is run multiple times with varying power values.

uli.powerfit = function(power_est,family,value,df1,df2) {

  ### get the ncp given power and study parameters
  ncp_est=mapply(getncp,df1=df1,df2=df2,power=power_est,family=family)
  ncp_est

  p = c()
  ### get p-values
  p[family == "f"] = 1-pf(value[family == "f"],df1[family == "f"],df2[family == "f"])
  p[family == "c"] = 1-pchisq(value[family == "c"],df1[family == "c"])

  ### bound p-values
  p = pmin(pmax(p,2.2e-16),1-2.2e-16)
  p

  ### get pp-values for some level of power
  pp_est=ifelse(family=="f" & p<.05,(pf(value,df1=df1,df2=df2,ncp=ncp_est)-(1-power_est))/power_est,NA)
  pp_est=ifelse(family=="c" & p<.05,(pchisq(value,df=df1,ncp=ncp_est)-(1-power_est))/power_est,pp_est)

  ### bound pp-values
  pp_est = pmin(pmax(pp_est,2.2e-16),1-2.2e-16)

  ### Combine pp-values using Stouffer meta-analysis
  res = sum(qnorm(pp_est),na.rm=TRUE)/sqrt(sum(!is.na(pp_est)))

  ### No need to use Stouffer; could just use mean-z-score
  #res = mean(qnorm(pp_est),na.rm=TRUE)
  return(res)
}

### This function simply tries power values from .051 to .99 and then finds the power vlaue with the best
fit

Uli.Estimate.Power = function(family,value,df1,df2) {

  ### Get Fit for different Power Values

```

```

fit=abs(uli.powerfit(.051,family,value,df1,df2))
for (i in 6:99) {
  fit=c(fit,abs(uli.powerfit(i/100,family,value,df1,df2))) #Now do 6% to 99%
}
# plot(c(5:99),fit[]) #plot just for diagnostics

### Find the minimum
mini=match(min(fit,na.rm=TRUE),fit)
mini

#get power value for minimal fit value
hat=(mini+4)/100 # hat = Power.Estimate

return(hat)
}

#Function that finds power that gives p-value=pct for the Stouffer test
#for example, get.power_pct(.5) returns the level of power that leads to p=.5 for the stouffer test.
#half the time we would see p-curves more right skewed than the one we see, and half the time
#less right-skewed, if the true power were that get.power_pct(.5). So it is the median estimate of power
#similarly, get.power_pct(.1) gives the 10th percentile estimate of power...

get.power_pct =function(pct,family,value,df1,df2) {

### Obtain the normalized equivalent of pct, e.g., for 5% it is -1.64, for 95% it is 1.64
z=qnorm(pct) #convert to z because powerfit() outputs a z-score.

### Quantify gap between computed p-value and desired pct
error = function(power_est, z) uli.powerfit(power_est,family,value,df1,df2) - z

###Find the value of power that makes that gap zero, (root)
res = uniroot(error, c(.0501, .99),z)$root

return(res)
}

### not used for simulation ###

# Confidence interval for power estimate

get.power.ci = function(family,value,df1,df2) {

### Boundary conditions
p.power.05 = pnorm(uli.powerfit(.051,family,value,df1,df2)) #Probability p-curve would be at least at right-
skewed if power=.051
p.power.99 = pnorm(uli.powerfit(.99,family,value,df1,df2)) #Probability p-curve would be at least at right-
skewed if power=.99

### Find lower bound of ci

#Low boundary condition? If cannot reject 5% power, don't look for lower levels, use 5% as the end
if (p.power.05 <= .95) power.ci.lb = .05

```

```

# High boundary condition? If we reject 99%, from below dont look for higher power, use 99% as the low
end
if (p.power.99 >= .95) power.ci.lb=.99

# If low bound is higher than 5.1% power and lower than 99% power, estimate it, find interior solution
if (p.power.05 > .95 && p.power.99 < .95) power.ci.lb = get.power_pct(.95,family,value,df1,df2)

### Find higher bound of ci

#If we reject 5% power from below, 5% is above the confidence interval, use 5% as the upper end of the
confidence interval
if (p.power.05 <= .05) power.ci.ub =.05

#If we do not reject that 99% power, don't look higher, use 99% as the higher end
if (p.power.99 >= .05) power.ci.ub =.99

#If the the upper bound is between 5% and 99%, find it
if (p.power.05 >.05 && p.power.99 <.05) power.ci.ub = get.power_pct(.05,family,value,df1,df2)

return(c(power.ci.lb,power.ci.ub))

}

fun.pcurve = function(family,value,df1,df2,CI=FALSE) {

res = c()

### Step 1: Estimate Power
power.estimate = Uli.Estimate.Power(family,value,df1,df2)
res = c(res,power.estimate)

### Step 2: Get Confidence Intervals
if (CI) {
power.ci = get.power.ci(family,value,df1,df2)
res = c(res,power.ci)
}

### Step 3: return results
return(res)

} ### End of Pcurve.Power.Estimation

#####
#### PCURVE FUNCTIONS END
#####

#####
#### ZCURVE FUNCTION BEGIN
#####

### This function fits an observed distribution of z-scores to a multi-model mixture model

```

```

#### z.val.input = observed z-scores,
#### z.crit = criterion value for significance, default = 1.96
#### Int.End End of Interval (do not fit extreme tai (default = 6)
#### bw = bandwidth of kernel density function, default = .05
#Int.End = 6
#bw = .05

fun.zcurve = function(z.val.input, z.crit = 1.96, Int.End=6, bw=.05) {

  Int.Beg = z.crit

  #### resolution of density function (doesn't seem to matter much)
  bars = 500

  #### limit to significant values
  z.val.input = z.val.input[z.val.input > z.crit]

  #### create set with z-scores in the interval used for model fitting
  #### add +.2 to the end of the interval to account for kernel density function decrease to zero
  Z.INT = z.val.input[z.val.input <= Int.End + .2]
  summary(Z.INT)

  #### find the maximum z-score. This is only needed if the maximum z-score is below Int.End
  max.z = Int.End
  if (max(Z.INT) < Int.End) max.z = max(Z.INT)
  max.z

  #### augment z-scores on the left side of Interval to avoid downward trend
  #### of kernal density function (assumes values go to 0)
  Augmentation.Left = TRUE
  if (Augmentation.Left) {
    #### create augmentation z-values to prevent decending to zero
    aug.left = c()
    height = length(Z.INT[Z.INT < Int.Beg + .2])/4
    for (i in 1:8) aug.left = c(aug.left,rep(Int.Beg - .05*i,height))
    #### combine actual data with augmentation data
    Z.INT = c(aug.left,Z.INT)
  }

  #### get the density of the observed data
  Z.Density = density(Z.INT,n=bars,bw=bw)

  #### select the densities in the actual interval (remove augmentation)
  Z.Density.X = Z.Density$x[Z.Density$x > Int.Beg & Z.Density$x < max.z]
  Z.Density.Y = Z.Density$y[Z.Density$x > Int.Beg & Z.Density$x < max.z]

  #### get the actual number of bars for the density distribution
  n.bars = length(Z.Density.X)

  #### get the width of each bar
  bar.width = Z.Density.X[2] - Z.Density.X[1]

  #### rescale the densities so that the area under the curve is 1
  Z.Density.Y = Z.Density.Y/(sum(Z.Density.Y*bar.width))

```

```

####
#### This completes the creation of the observed density distribution to be fitted by z-curve
####

####
#### Fitting the mixture model ####
####

#### define the non-centrality parameters (mean of standard normal) for the mixture model
n.ncz = Int.End+1
ncz = seq(0,n.ncz-1,1)

#### get the densities for each bar and each non-centrality parameter
Dens = c()
for(i in 1:n.bars) {
  for (j in 1:length(ncz)) {
    Dens = c(Dens,dnorm(Z.Density.X[i],ncz[j]))
  }
}
#### Create a matrix of non-centrality parameters by bars
Dens = matrix(Dens,n.ncz,byrow=FALSE)
#### rescale the densities for the interval so that the area under the curve is 1
sum.dens = rowSums(Dens)
Dens = Dens/(sum.dens * bar.width)

#### THIS IS THE FUNCTION THAT FITS THE MIXTURE MODEL TO THE OBSERVED DENSITY
DISTRIBUTION
zcurve.mixture = function(theta,RetEst=FALSE) {

#### get the weights and rescale
weight = theta
weight = weight/sum(weight)

#### compute the new estimated density distribution
z.est = c()
for (i in 1:n.bars) z.est[i] = sum(Dens[,i]*weight)

#### compare the observed and predicted density distributions
#### use the absolute differences as fit criterion
sum.abs.dev = sum(abs(z.est-Z.Density.Y))

#### return value to optimization function
return(sum.abs.dev)

} # End of function comp.zcurves

#### provide some starting values
startval = rep(1,n.ncz)
startval = startval/sum(startval)

#### set lower limit for weights
lowlim = rep(0,n.ncz)

```

```

#### set upper limits for weights
highlim = rep(1,n.ncz)

#### start the estimation process
zcurve.mix.results =
nlminb(startval,zcurve.mixture,lower=lowlim,upper=highlim,control=list(eval.max=100))

#### get the estimated weights
ncz.weight = zcurve.mix.results$par

#### rescale weights to sum to 1
ncz.weight = ncz.weight/sum(ncz.weight)
round(ncz.weight,2)

#### compute power for mixture model
power.x = sum((pnorm(ncz,z.crit) + pnorm(-z.crit,ncz))*ncz.weight)
power.x
extreme = length(z.val.input[z.val.input > Int.End])/length(z.val.input)
extreme
power.est = power.x*(1-extreme) + extreme
power.est

return(power.est)

} # End function zcurve

#####
#### ZCURVE FUNCTION END
#####

#####
#### Run Simulation with t-values as Input
#####

get.estimated.for.t = function(inp,Do.Z=TRUE,Do.P=1) {

ge.res = c()

for (i in 1:dim(inp)[1]) {
  t = inp[i,1:k]
  N = inp[i,(k+1):(2*k)]
  print(paste("zcurve",i))
  z.val.input = qnorm(pt(t,N-2))
  zcurve = fun.zcurve(z.val.input=z.val.input)
  value = t^2
  family = rep("f",length(value))
  df1 = rep(1,length(value))
  df2 = N-2
  print(zcurve)
  print(paste("pcurve",i))
  if (Do.P == 2) {
    pcurve = fun.pcurve(family,value,df1,df2,CI=TRUE)
  } else {

```



```

        pcurve = fun.pcurve(family,value,df1,df2)
      }
      print(pcurve)
      ge.res = rbind(ge.res,c(zcurve,pcurve))
      print(summary(ge.res))
    } #End of For

  return(ge.res)

} # end of get.estimate

#####
#### Begin Simulations for Study 1
#####

### Sim 1.1: d = rnorm(x,d,.2) N = 10-70 ### from Supplement in Simonsohn 2014

k = 100
sim = 5000

n.sim = k*sim*20
n.sim

tp = c()
var.z = c()
med.obs.pow = c()
res = c()

for (d in seq(0,.8,.2)) {

  ES = rnorm(n.sim,d,.2)
  N = round(runif(n.sim,9.5,70.5))
  se = 2/sqrt(N)
  nct = ES/se

  inp = cbind(nct,N)
  obs.t = abs(apply(inp,1,function(x) rt(1,x[2]-2,x[1])))
  obs.t
  hist(obs.t)
  obs.z = qnorm(pt(obs.t,N-2))
  hist(obs.z)
  summary(obs.z)

  obs.t.sig = obs.t[pt(obs.t,N-2) > .975]
  hist(obs.t.sig,xlim=c(0,6))

  power = 1-pt(qt(.975,N-2),N-2,nct)
  hist(power)

  tp = c(tp,mean(power[pt(obs.t,N-2) > .975]))
  print(tp)
}

```

```

med.obs.pow = c(med.obs.pow,pnorm(median(obs.z[pt(obs.t,N-2) > .975]),1.96))
var.z = c(var.z,var(obs.z[obs.z > 1.96]))

inp = cbind(obs.t,N)[pt(obs.t,N-1) > .975,]
inp = inp[1:(k*sim),]
dim(inp)

inp = cbind(matrix(inp[,1],,k),matrix(inp[,2],,k))
dim(inp)

res = cbind(res,get.estimates.for.t(inp))
dim(res)

} #End of Simulations for Study 1

summary(res)

### True Power
round(tp*100,1)

### Median Observed Power
round(med.obs.pow*100,1)

### Variance of significant z-scores
round(var.z,2)

zcurve = colMeans(res[,c(1,3,5,7,9)])
pcurve = colMeans(res[,c(2,4,6,8,10)])

results = cbind(tp,zcurve,pcurve)
round(results*100,1)

#####
### End of Simulation Study 1
#####

#####
### Begin of Simulation Study 2
#####

### criterion value for significance
z.crit = 1.96

### number of test statistics
k = 100

### number of simulations
sim = 5000

###
### Sim 1.1: Normal M=1.46, SD=1 31% Power
###

```

```

power = .30
n.sim = k * sim * 20 #### number of simulated test statistics BEFORE selection
z = rep(qnorm(power,1.96),n.sim) ### create test statistics
length(z)

# randomize order of z-scores and true power
rand = runif(length(z))
z = z[order(rand)]

# create sampling error for observed z-scores
se = rnorm(length(z))
obs.z = abs(z + se)

obs.z.sig = obs.z[obs.z > z.crit]
obs.z.sig = obs.z.sig[1:(k*sim)]
var(obs.z.sig)

summary(obs.z.sig)
length(obs.z.sig)

### get true power
tp = mean(pnorm(z,z.crit)[obs.z > z.crit])
tp

mat.z.val.input = matrix(obs.z.sig,,k)
dim(mat.z.val.input)

res = get.estimates(mat.z.val.input)
dim(res)

summary(res)
sd(res[,1])
sd(res[,3])
sd(res[,9])

length(res[res[,3] > .20 & res[,3] < .40,3])/dim(res)[1]
length(res[res[,9] > .20 & res[,9] < .40,9])/dim(res)[1]

res.sim.1.1 = res
res = res.sim.1.1

summary(res.sim.1.1)

table(res[,8] > 1.96)

###
### Sim 1.2: Normal M=1.96, SD=1 50% Power
###

power = .50
n.sim = k * sim * 20 #### number of simulated test statistics BEFORE selection
z = rep(qnorm(power,1.96),n.sim) ### create test statistics
length(z)

# randomize order of z-scores and true power

```

```

rand = runif(length(z))
z = z[order(rand)]

# create sampling error for observed z-scores
se = rnorm(length(z))
obs.z = abs(z + se)

obs.z.sig = obs.z[obs.z > z.crit]
obs.z.sig = obs.z.sig[1:(k*sim)]
summary(obs.z.sig)
length(obs.z.sig)
var(obs.z.sig)

#### get true power
tp = mean(pnorm(z,z.crit)[obs.z > z.crit])
tp

mat.z.val.input = matrix(obs.z.sig,,k)
dim(mat.z.val.input)

res = get.estimates(mat.z.val.input,Do.P=TRUE)
dim(res)

summary(res)
sd(res[,3])
sd(res[,9])

length(res[res[,3] > .40 & res[,3] < .60,3])/dim(res)[1]
length(res[res[,9] > .40 & res[,9] < .60,9])/dim(res)[1]

res.sim.1.2 = res
res = res.sim.1.2

####
#### Sim 1.3: Normal M=2.80, SD=1 80% Power
####

qnorm(.80,1.96)

power = .80
n.sim = k * sim *20 #### number of simulated test statistics BEFORE selection
z = rep(qnorm(power,1.96),n.sim) #### create test statistics
length(z)

# randomize order of z-scores and true power
rand = runif(length(z))
z = z[order(rand)]

# create sampling error for observed z-scores
se = rnorm(length(z))
obs.z = abs(z + se)

obs.z.sig = obs.z[obs.z > z.crit]

```

```

obs.z.sig = obs.z.sig[1:(k*sim)]
summary(obs.z.sig)
length(obs.z.sig)
var(obs.z.sig)

### get true power
tp = mean(pnorm(z,z.crit)[obs.z > z.crit])
tp

mat.z.val.input = matrix(obs.z.sig,,k)
dim(mat.z.val.input)

res = get.estimates(mat.z.val.input,Do.P=TRUE)
dim(res)

summary(res)
sd(res[,1])
sd(res[,2])

length(res[res[,3] > .70 & res[,3] < .90,3])/dim(res)[1]
length(res[res[,9] > .70 & res[,9] < .90,9])/dim(res)[1]

res.sim.1.3 = res
res = res.sim.1.3

###
### Sim 2.1: Normal M=0, SD=1 31% Power
###

NCZ = 0
n.sim = sim * k * 20
z = abs(rnorm(n.sim,NCZ,1))

# randomize order of z-scores and true power
rand = runif(length(z))
z = z[order(rand)]

# create sampling error for observed z-scores
se = rnorm(length(z))
obs.z = abs(z + se)

obs.z.sig = obs.z[obs.z > z.crit]
obs.z.sig = obs.z.sig[1:(sim*k)]
length(obs.z.sig)
var(obs.z.sig)

### get true power
tp = mean(pnorm(z,z.crit)[obs.z > z.crit])
tp

```

```

mat.z.val.input = matrix(obs.z.sig,,k)
dim(mat.z.val.input)

res = get.estimated(mat.z.val.input)
dim(res)

summary(res)
sd(res[,1])
sd(res[,2])

length(res[res[,1] > .20 & res[,1] < .40,1])/dim(res)[1]
length(res[res[,2] > .20 & res[,2] < .40,2])/dim(res)[1]

res.sim.2.1 = res
summary(res.sim.2.1)

table(res[,8] > 1.96)

####
### Sim 2.2: Normal M=1.2, SD=1 50% Power
###

NCZ = 1.2
n.sim = sim * k * 20
z = abs(rnorm(n.sim,NCZ,1))

# randomize order of z-scores and true power
rand = runif(length(z))
z = z[order(rand)]

# create sampling error for observed z-scores
se = rnorm(length(z))
obs.z = abs(z + se)

obs.z.sig = obs.z[obs.z > z.crit]
obs.z.sig = obs.z.sig[1:(sim*k)]
length(obs.z.sig)

var(obs.z.sig)

### get true power
tp = mean(pnorm(z,z.crit)[obs.z > z.crit])
tp

mat.z.val.input = matrix(obs.z.sig,,k)
dim(mat.z.val.input)

res = get.estimated(mat.z.val.input,Do.P=TRUE)
dim(res)

summary(res)
sd(res[,3])
sd(res[,9])

```

```

res = res.sim.2.2
length(res[res[,3] > .40 & res[,3] < .60,3])/dim(res)[1]
length(res[res[,9] > .40 & res[,9] < .60,9])/dim(res)[1]

res.sim.2.2 = res
summary(res.sim.2.2)
table(res.sim.2.2[,8] > 1.96)

####
#### Sim 2.3: Normal M=2.75, SD=1 80% Power
####

NCZ = 2.75
n.sim = sim * k * 20
z = abs(rnorm(n.sim,NCZ,1))

# randomize order of z-scores and true power
rand = runif(length(z))
z = z[order(rand)]

# create sampling error for observed z-scores
se = rnorm(length(z))
obs.z = abs(z + se)

par(new=TRUE)
hist(obs.z,col="red",density=10,freq=FALSE,xlim=c(0,8),ylim=c(0,.4))

obs.z.sig = obs.z[obs.z > z.crit]
obs.z.sig = obs.z.sig[1:(sim*k)]
length(obs.z.sig)

var(obs.z.sig)

#### get true power
tp = mean(pnorm(z,z.crit)[obs.z > z.crit])
tp

#### get true power
tp = mean(pnorm(z,z.crit)[obs.z > z.crit])
tp

mat.z.val.input = matrix(obs.z.sig,,k)
dim(mat.z.val.input)

res = get.estimates(mat.z.val.input,Do.P=TRUE)
dim(res)

summary(res)
sd(res[,3])
sd(res[,9])

```

```
length(res[res[,3] > .70 & res[,3] < .90,3])/dim(res)[1]
length(res[res[,9] > .70 & res[,9] < .90,9])/dim(res)[1]
```

```
res.sim.2.3 = res
summary(res.sim.2.3)
```

```
### Sim 3.1: Skewed 30% Power #Z-Curve 29% P-Curve 45%
```

```
n.sim = sim*k
#create distribution of non-centrality parameters (mean of standard normal)
z = abs(rnorm(120*n.sim,0,.5))
z = c(z,abs(rnorm(100*n.sim,1,.5)))
z = c(z,abs(rnorm(1*n.sim,2,.5)))
z = c(z,abs(rnorm(1*n.sim,3,.5)))
z = c(z,abs(rnorm(1*n.sim,4,.5)))
z = c(z,abs(rnorm(1*n.sim,5,.5)))
z = c(z,abs(rnorm(1*n.sim,6,.5)))
hist(z)
```

```
# randomize order of z-scores and true power
rand = runif(length(z))
z = z[order(rand)]
```

```
# create sampling error for observed z-scores
se = rnorm(length(z))
obs.z = abs(z + se)
```

```
var(obs.z[obs.z > z.crit])
```

```
tp = mean(pnorm(z,z.crit)[obs.z > z.crit])
tp
```

```
obs.z.sig = obs.z[obs.z > z.crit]
obs.z.sig = obs.z.sig[1:(sim*k)]
length(obs.z.sig)
var(obs.z.sig)
```

```
mat.z.val.input = matrix(obs.z.sig,,k)
dim(mat.z.val.input)
```

```
res = get.estimates(mat.z.val.input)
dim(res)
```

```
summary(res)
sd(res[,1])
sd(res[,2])
```

```
length(res[res[,1] > .20 & res[,1] < .40,1])/dim(res)[1]
length(res[res[,2] > .20 & res[,2] < .40,2])/dim(res)[1]
```



```

res.sim.3.1 = res
summary(res.sim.3.1)

### Sim 3.2: Skewed 50% Power

n.sim = sim*k
#create distribution of non-centrality parameters (mean of standard normal)

z = abs(rnorm(50*n.sim,0,.5))
z = c(z,abs(rnorm(40*n.sim,1,.5)))
z = c(z,abs(rnorm(10*n.sim,2,.5)))
z = c(z,abs(rnorm(2*n.sim,3,.5)))
z = c(z,abs(rnorm(2*n.sim,4,.5)))
z = c(z,abs(rnorm(2*n.sim,5,.5)))
z = c(z,abs(rnorm(1*n.sim,6,.5)))

hist(z)

# randomize order of z-scores and true power
rand = runif(length(z))
z = z[order(rand)]

# create sampling error for observed z-scores
se = rnorm(length(z))
obs.z = abs(z + se)

tp = mean(pnorm(z,z.crit)[obs.z > z.crit])
tp

obs.z.sig = obs.z[obs.z > z.crit]
obs.z.sig = obs.z.sig[1:(sim*k)]
length(obs.z.sig)
var(obs.z.sig)

mat.z.val.input = matrix(obs.z.sig,,k)
dim(mat.z.val.input)

res = get.estimates(mat.z.val.input)
dim(res)

summary(res)
sd(res[,1])
sd(res[,2])

length(res[res[,1] > .40 & res[,1] < .60,1])/dim(res)[1]
length(res[res[,2] > .40 & res[,2] < .60,2])/dim(res)[1]

res.sim.3.2 = res
summary(res.sim.3.2)

###
### Sim 3.3: Skewed 75% Power #Z-Curve 75% P-Curve 75%

```

```
###
```

```
n.sim = sim*k
#create distribution of non-centrality parameters (mean of standard normal)
```

```
z = abs(rnorm(8*n.sim,0,.5))
z = c(z,abs(rnorm(4*n.sim,1,.5)))
z = c(z,abs(rnorm(3*n.sim,2,.5)))
z = c(z,abs(rnorm(3*n.sim,3,.5)))
z = c(z,abs(rnorm(2*n.sim,4,.5)))
z = c(z,abs(rnorm(2*n.sim,5,.5)))
z = c(z,abs(rnorm(2*n.sim,6,.5)))
```

```
hist(z)
```

```
# randomize order of z-scores and true power
rand = runif(length(z))
z = z[order(rand)]
```

```
# create sampling error for observed z-scores
se = rnorm(length(z))
obs.z = abs(z + se)
```

```
tp = mean(pnorm(z,z.crit)[obs.z > z.crit])
tp
```

```
obs.z.sig = obs.z[obs.z > z.crit]
obs.z.sig = obs.z.sig[1:(sim*k)]
length(obs.z.sig)
var(obs.z.sig)
```

```
mat.z.val.input = matrix(obs.z.sig,,k)
dim(mat.z.val.input)
```

```
res = get.estimates(mat.z.val.input)
dim(res)
```

```
summary(res)
sd(res[,3])
sd(res[,9])
```

```
length(res[res[,3] > .70 & res[,3] < .90,3])/dim(res)[1]
length(res[res[,9] > .70 & res[,9] < .90,9])/dim(res)[1]
```

```
res.sim.3.3 = res
summary(res.sim.3.3)
table(res.sim.3.3[,8] > 1.96)
```

```
#####
```

Combine Results

```
res.all.1 = cbind(res.sim.1.1,res.sim.1.2,res.sim.1.3)
res.all.2 = cbind(res.sim.2.1,res.sim.2.2,res.sim.2.3)
res.all.3 = cbind(res.sim.3.1,res.sim.3.2,res.sim.3.3)
res.all = cbind(res.all.1,res.all.2,res.all.3)
dim(res.all)
summary(res.all.2)
colMeans(res.all)
```