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 been used for the same purpose. Simulation studies show that both methods perform well when all studies have the same power, but p-curve overestimates power if power varies 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 was typically very high. 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 of 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.

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 noted 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 studies which depends on the actual 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 exactly the same 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 more significant than non-significant results.

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 and studies with low power as well as false positives (i.e., False Positive & True Positive with low power vs. True Positive with High Power). 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 would meet these standards, published true positive results are replicable and false positive results are rare and are much more likely to fail in replication attempts than true positive 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 for 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 psychology has a replication crisis. 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. As a result,

it is difficult to conduct replications for a large representative sample of studies. 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. We examine two methods that aim to correct for the 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 main 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).

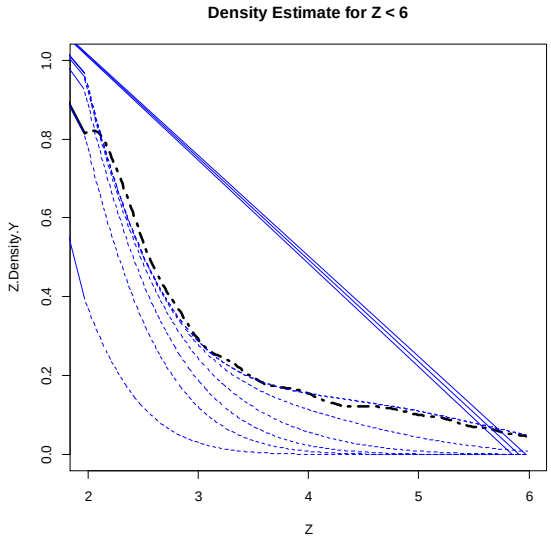
One problem is that P-curve assumes that all studies have the same population effect size. Although this is an unrealistic assumption, Simonsohn et al. (2014) suggest that "p-curve is robust to heterogeneity in effect size across studies" (p. 680). To our knowledge, the robustness of p-curve has not 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 reality, 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. 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 aims to fit the observed distribution with 7 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 actually 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, 29% for m = 1, 14% for m = 2, 12% for m = 3, 14% for m = 4, 14% for m = 5, 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.

Power = 1-pnorm(1.96,m) + pnorm(-1.96,m)

The power values corresponding to the means of the seven standard normal distributions are 5%, 17%, 52%, 85%, 98%, 99%, and 99.99%.

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%.

Thus, the observed distribution of z-scores in Figure 1 implies that the set of studies with significant results has an average power of 50%.

As noted above, the observed distribution of z-scores could have been produced in many ways. It is not possible to interpret the weights assigned to the mixture models as realistic values of the percentage of studies sampled from a particular normal distribution. Hence, it is not possible to infer from the 17% estimate for m = 0 that 17% of the studies were false positives. Although different mixtures are possible, the main assumption of z-curve is that the weighted average of mixture models that fit an observed distribution provides an accurate estimate of the average power of this set of studies. We conducted our simulation studies to examine the ability of z-score to fulfill this promise.

The key difference between p-curve and z-curve is that p-curve assumes that all studies have the same power. In contrast, z-curve allows for heterogeneity in power by using a mixture

model with multiple standard normal distributions that represent different levels of statistical power. Although p-curve may be relatively robust if the assumption of equal power is violated, we predict that z-curve will outperform p-curve with heterogeneous data because it allows for heterogeneity and makes no assumptions about the distribution of true power in the set of studies.

Simulation Study

We used Z values as the observed test statistics for our simulations. 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 lead to the same results. Moreover, Brunner and Schimmack (2016) demonstrated in extensive simulations that the type of test statistic (F, t, or chi-square) does not influence the outcome of simulation studies.

We used a 3 x 3 design for our simulation study. One factor varied the average true power with values of .31, .50, and .80. The other factor varied the distribution of true power. One condition simulated homogeneous data. The second condition simulated heterogeneity with a normal distribution, and the third condition simulated heterogeneity with a skewed distribution. We used a fixed set of k = 100 studies for all simulations. For each condition, we ran 5,000 simulations.

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

Simulation / Method	Mean True Power	Mean Est. Power	SD	% Estimates +/- .10
Homogeneity (M = 1.46, SD = 0)				
P-Curve	.31	.31	.07	.83
Z-Curve	.31	.33	.08	.79
Homogeneity (M = 1.96, SD = 0)				
P-Curve	.50	.50	.07	.84
Z-Curve	.50	.50	.07	.82
Homogeneity (M = 2.80, SD = 0)				
P-Curve	.80	.80	.04	.94
Z-Curve	.80	.78	.05	.92
Heterogeneity, Normal (M = 0, SD = 1)				
P-Curve	.31	.30	.07	.79
Z-Curve	.31	.31	.07	.82
Heterogeneity, Normal (M = 1.20, SD = 1)				
P-Curve	.50	.56	.07	.67
Z-Curve	.50	.51	.07	.82
Heterogeneity, Normal (M = 2.75, SD = 1)				
P-Curve	.80	.89	.03	.56
Z-Curve	.80	.81	.06	.94
Heterogeneity, Skewed (M = 0.75, SD = 0.73)				
P-Curve	.31	.43	.10	.38
Z-Curve	.31	.31	.07	.84
Heterogeneity, Skewed (M = 1.03, SD = 1.10)				
P-Curve	.50	.74	.07	.03
Z-Curve	.50	.51	.07	.83
Heterogeneity, Skewed (M = 2.18, SD = 1.93)				
P-Curve	.80	.97	.01	.00
Z-Curve	.80	.81	.05	.94

Table 1 shows that p-curve performs slightly better than z-curve when all studies have the same power. However, both methods produce a majority of estimates within 10 percentage points of the true power. With normally distributed heterogeneity, p-curve overestimates average power, and with high true power produces only slightly more than 50% estimates that fall within 10 percentage points of true power. With skewed distributions of true power, p-curve

fails to produce reasonable estimates. In contrast, z-curve showed high large-sample accuracy in all conditions. The standard deviations of the two methods are very similar. However, due to the systematic bias in p-curve's estimates, p-curve has a lower percentage of estimates that fall within a +/- 10% interval around true average power. When power was high (80%) and the distribution of true power 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 the dataset is homogeneous (i.e., all studies have identical power), but P-curve estimates can be dramatically inflated when the distribution of true power is skewed. In contrast, z-curve is not affected by heterogeneity or the distribution of true power and produced accurate estimates in all conditions. This is not surprising. Z-curve was developed to fit heterogeneous data, whereas p-curve was developed for the special case of fixed power.

Application to Actual Test Statistics

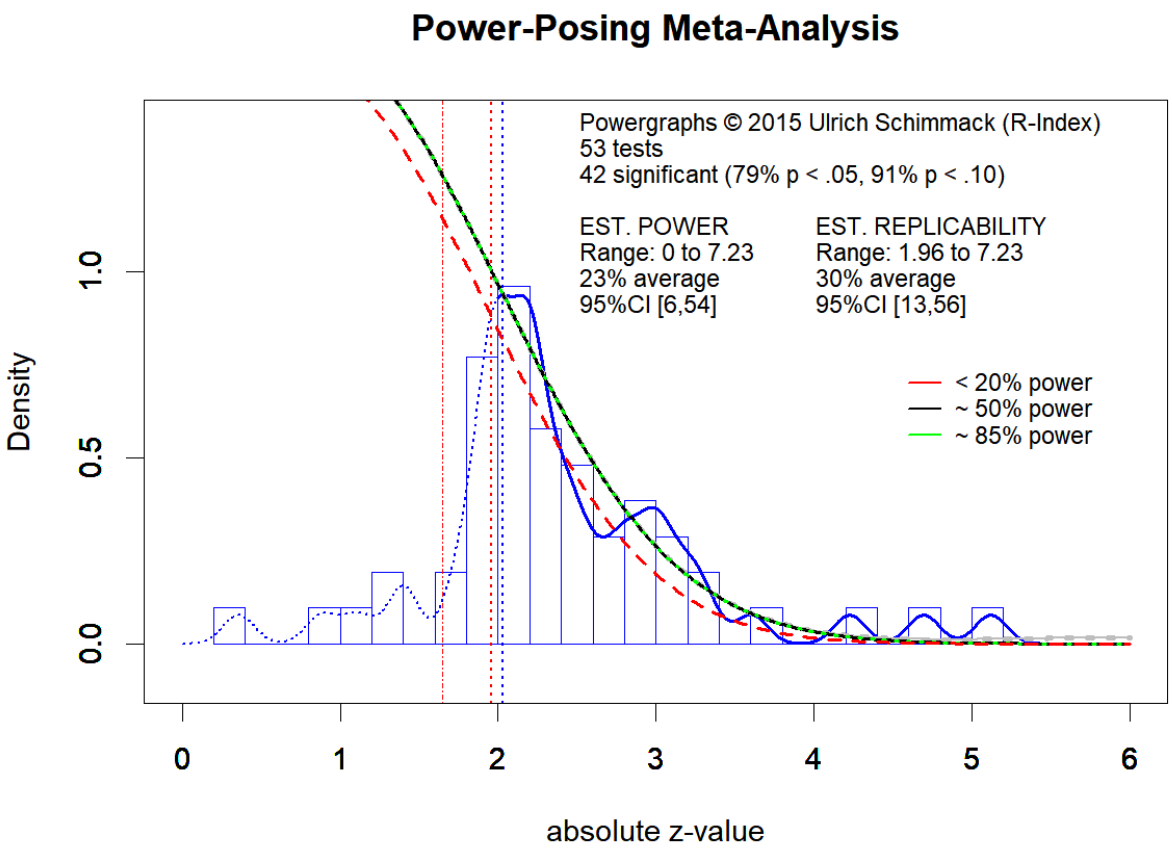
Demonstration 1: A Meta-Analysis of Power Posing Effects

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 2 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 However, the z-curve estimate is only 30%. Figure 2 also shows a 95%CI for the point estimate. This estimate is based on a bootstrap method that has been validated by Brunner and Schimmack (2016). Given the relatively small number of studies, the 95%CI is relatively wide and ranges from 16% to 56%. The z-curve estimate is notably lower than the p-curve estimate of 44%. The reason for the discrepancy is heterogeneity. Figure 2 shows that most studies are just significant, but a few studies reported strong evidence ($z > 4$).

To examine the robustness of estimates against outliers, we also obtained estimates for a subset of studies with z-scores less than 4 ($k = 49$). Excluding the four studies with extreme scores had relatively little effect on z-curve; replicability estimate = 34%. In contrast, the p-curve estimate dropped from 44% to 5%, while the 90%CI of p-curve ranged from 13% to 30% and did not include the point estimate. This suggests further problems with the p-curve method of estimating power.



In conclusion, our first application of z-curve to actual data reveals that the difference between p-curve and z-curve has practical implications. Even studies that investigated a common phenomenon produced sufficient heterogeneity to inflate p-curve estimates of average power. The z-curve estimate of power was only 30%. The lower bound of the 95% confidence interval was above 5%, which makes it possible to reject the null-hypothesis that all studies reported false positive results. However, the low average power also implies that the power-posing hypothesis is supported by underpowered studies. More importantly, the average power of 30% allows for a large subset of studies that reported false positives and it is impossible to distinguish false positives from true positives with low power. Thus, it remains unclear which claims about power posing effects can be replicated. Only new studies with larger samples can answer this question.

In sum, Cuddy et al. (2017) concluded that their 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 is considerably lower and the lower bound of the 95%CI is 16% power. In our opinion, studies with average power of 30% do not constitute robust evidence. Second, 30% power for a homogeneous set of studies may be considered sufficient evidence for an effect. However, for a heterogenous set of studies, 30% average power does not provide information about specific studies that can be replicated. Thus, 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 and which outcomes were affected. To produce robust evidence for an effect, it is necessary to conduct studies with more power.

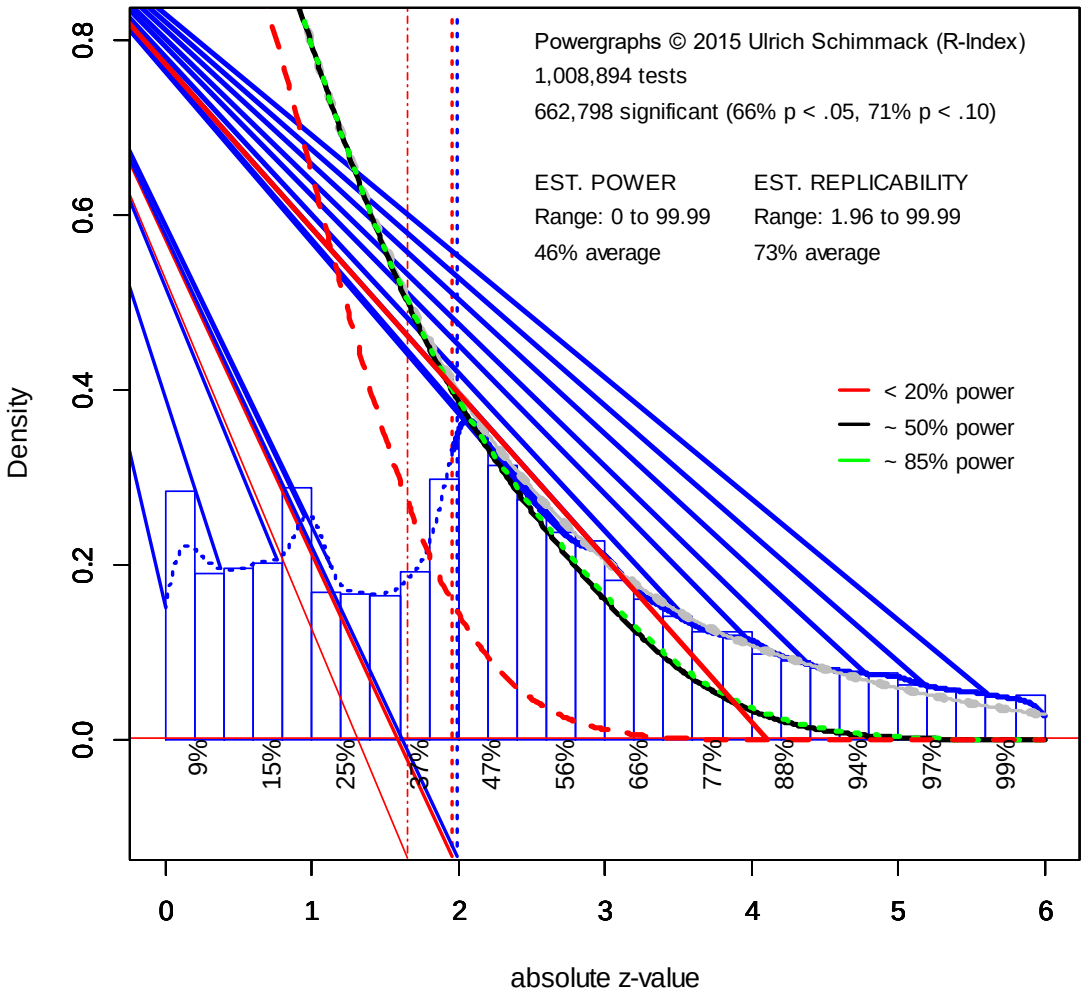
Demonstration 2: Replicability of Psychology

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 104 psychology journals in the years from 2010 to 2016. We downloaded all articles as PDF files and converted them to text files. We wrote a program in R to extract F-tests, t-tests and z-test that were reported in the results section (F(x,xx) = X.XX, t(xx) = X.XX, z = X.XX). The search yielded 1,008,894 test statistics and 66% (k = 662,798) were significant using alpha = .05 (two-tailed). Figure 3 shows the distribution of z-scores. The shape of the distribution shows that there is heterogeneity in power 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.

The z-curve estimate of replicability was 73%. Given the large number of test statistics, the 95% confidence interval around this estimate is very tight and ranged from 71% to 74%. The estimate of 73% is surprisingly high in comparison to the 36% successful actual replications in the OSC reproducibility project and the estimate of 30% replicability for the power-posing meta-analysis. There are a number of factors that can explain this 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. Thus, we used all years to obtain a single replicability estimate for focal hypothesis tests in social psychology. We also downloaded all articles and used the automated extraction method to obtain an estimate based on all test statistics. This way we were able to examine whether focal hypothesis tests are less replicable than other test statistics reported in psychology journals.

Figure 4 shows the results based on the automated extraction of all test statistics. The replicability estimate is 65%, 95%CI = [.62,.68]. This estimate is lower than for the larger set of psychology journals, suggesting that results in social psychology are less replicable than those in other areas of psychology (OSC, 2015; Schimmack, 2017). However, 65% is still considerably higher than the 25% success rate for actual replications of social psychological studies in the OSC project.

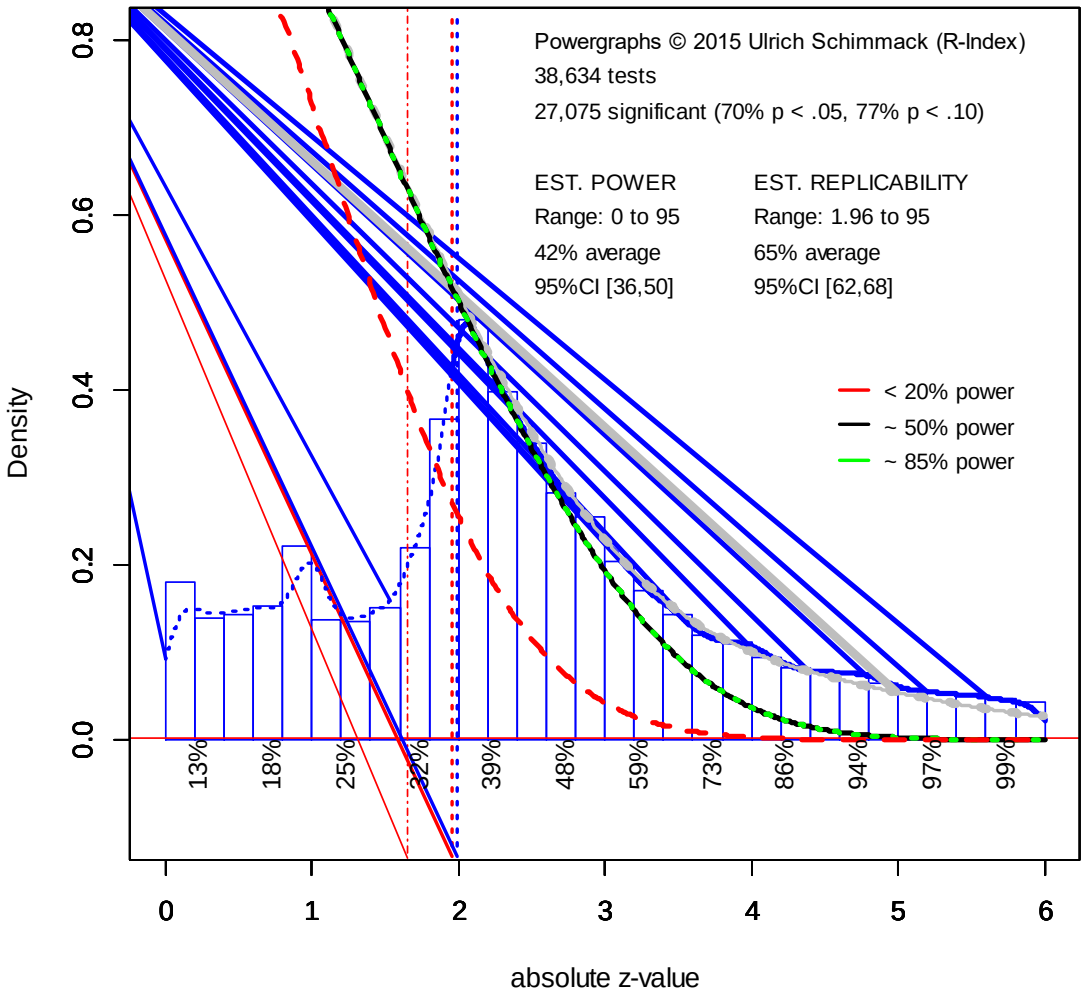
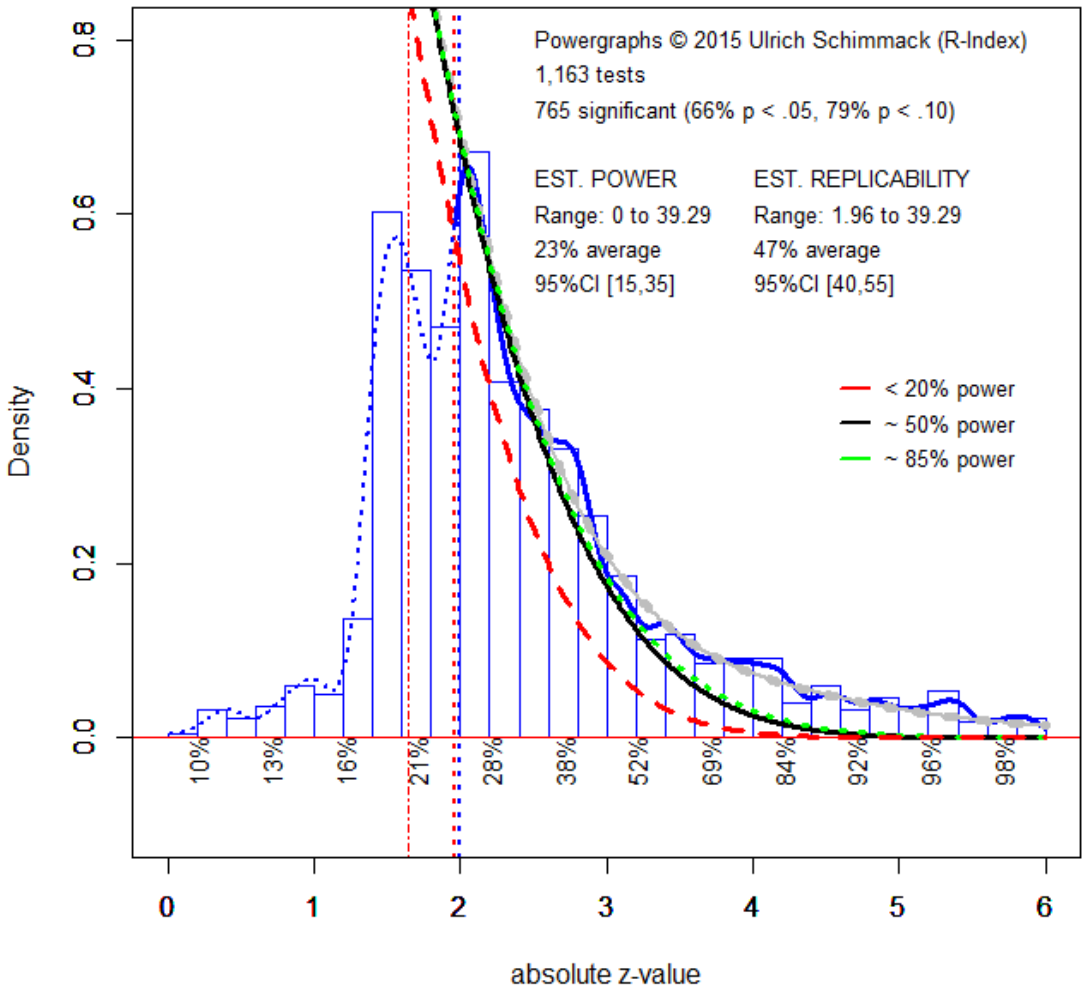

Powergraphs © 2015 Ulrich Schimmack (R-Index)
38,634 tests
27,075 significant (70% p < .05, 77% p < .10)
EST. POWER
Range: 0 to 95
42% average
95%CI [36,50]
EST. REPLICABILITY
Range: 1.96 to 95
65% average
95%CI [62,68]
< 20% power
~ 50% power
~ 85% power
Density
0.0
0.2
0.4
0.6
0.8
13%
18%
25%
32%
39%
48%
59%
73%
86%
94%
97%
99%
0
1
2
3
4
5
6
absolute z-value

Figure 5 shows the results for Motyl et al.'s (2017) focal hypothesis tests. The replicability estimate is only .47, 95%CI = [.40,.55]. This estimate is nearly 20 percentage points lower than the replicability estimate based on the automatic extraction method. This finding shows that the automated extraction method produces overly positive results because it does not distinguish between focal and non-focal tests. To obtain absolute estimates of replicability it is necessary to identify theoretically important test statistics.

The estimate of 47% replicability has several implications. First, the estimate is not as bad as many may have feared. It is unlikely that most published results in social psychology are false positive results. Although we cannot determine the number of false positives an average of 47% 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.


Powergraphs © 2015 Ulrich Schimmack (R-Index)
1,163 tests
765 significant (66% p < .05, 79% p < .10)
EST. POWER
Range: 0 to 39.29
23% average
95%CI [15,35]
EST. REPLICABILITY
Range: 1.96 to 39.29
47% average
95%CI [40,55]
< 20% power
~ 50% power
~ 85% power
Density
0.0
0.2
0.4
0.6
0.8
10%
13%
16%
21%
28%
38%
52%
69%
84%
92%
96%
98%
0
1
2
3
4
5
6
absolute z-value

General Discussion

The main goal of this article was to introduce and evaluate a statistical method, z-curve, that estimates 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 heterogeneity. With heterogeneous and skewed distributions of true power, p-curve overestimates average power.

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%). The extent of bias varies as a function of several unknown factors. Rather than assuming that p-curve provides a robust estimate, we recommend z-curve as the best method to estimate the average the average power of original studies that produced a significant result, which we call replicability. 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 72%. 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. Whereas the estimate for social psychology based on all statistics was 65%, the estimate for focal hypothesis tests was only 47%. This estimate is limited to social psychology and estimates for psychology in general might be somewhat higher. One goal for future research is to conduct replicability analysis for all disciplines in psychology based on representative samples of focal hypothesis tests.

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 a number of limitations that can affect its estimates. Most important, z-curve assumes that all studies used the same criterion for statistical significance. If, for example, a study corrected p-values for multiple comparisons, z-curve will not model the selection process accurately and overestimate replicability. In our experience, this is a minor problem because

most studies use the $p < .05$ criterion to reject the null-hypothesis. Moreover, z-curve could be adjusted to allow for study-specific criterion values.

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 number of 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 make the assumption 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), Hilsdale, New Jersey: Erlbaum.

Cuddy, A. J., Schultz, S. J., & and 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

Hagger M. S., Chatzisarantis N. L., Alberts H., Anggono C. O., Batailler C., Birt A., Zwienenberg 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