How Replicable is Psychology? Estimating Replicability Based on Test Statistics in Original Studies

Ulrich Schimmack and Jerry Brunner

University of Toronto

Author Note

The work reported in this article is a truly collaborative effort. We would like to thank Dr. Jeffrey Graham for providing remote access to the computers in the Psychology Laboratory at the University of Toronto Mississauga. Thanks to Josef Duchesne for technical advice. 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.

Supplemental materials are available online at

http://www.utstat.toronto.edu/~brunner/zcurve

Abstract

In recent years, the replicability of original findings published in psychology journals has been questioned. We show that the replicability of a randomly chosen result cannot exceed population mean power, so that estimates of mean power are optimistic estimates of replicability. We then present two methods that can be used to estimate mean power for a heterogeneous set of studies with significant results: Maximum Likelihood and Z-curve. We present the results of two large-scale simulation studies with heterogeneous effect sizes and sample sizes. Both methods provide robust estimates of replicability, but z-curve has the advantage that it does not require assumptions about the distribution of effect sizes. We show that both methods overestimate replicability in the Open Science Collaborative reproducibility project and we discuss possible reasons for this. Based on the simulation studies, we recommend z-curve as a valid method to estimate replicability based on statistical results of original studies without the need to conduct actual replication studies.

Keywords: Power estimation, Post-hoc power analysis, Publication bias, Maximum likelihood, Z-curve, Effect size, Replicability, Simulation.

How replicable is psychology? Estimating Replicability on the Basis of Test Statistics in Original Studies

Science is built on a mixture of trust and healthy skepticism. Scientists who read and cite published work trust the authors, reviewers, and editors to ensure that most reported results provide sufficient credible support for theoretical conclusions. At the same time, scientists also insist that studies be reported in sufficient detail that other researchers can repeat them and see whether they can replicate the results. Replication studies help to detect false positive results in original studies because a false positive result is unlikely to produce a significant result again in future replication studies. Replicability is acknowledged to be a requirement of good science (Popper 1934, Bunge 1998). According to Fisher, replicability is also a characteristic of a good experiment. If the null hypothesis is false, “a properly designed experiment rarely fails to give ... significance” (Fisher, 1926, p. 504). Therefore, it is not sufficient that a study reports a true positive result. It also should produce true positive results in future replication studies.

Concerns about Replicability

In recent years, psychologists and other scientists have started to realize that published results are far less replicable than one would expect based on the high rate of significant results (Baker 2016; Begley 2013, Begley & Ellis 2012; Chang & Li 2015, Francis, 2012; Hirschhorn, Lohmueller, Byrne, & Hirschhorn 2002, Ioannidis 2008, John, Lowenstein, & Prelec 2012; Schimmack, 2012). In psychology, the Open Science Collaboration (OSC) project estimated the replicability of published results in psychology by replicating 100 primary findings of articles from three influential journals that publish results from social and cognitive psychology (OSC, 2015). The OSC authors used several criteria of replicability. The most widely cited criterion was the percentage of replication studies that reproduce a statistically significant result with the standard criterion of statistical significance (p < .05, two-tailed). The authors advocated this criterion as “a straightforward method for evaluating replication” and suggest that “this dichotomous vote-counting method is intuitively appealing and consistent with common heuristics used to decide whether original studies “worked.” (OSC, 2015, aac4716-4). Whereas 97% of the studies reported a statistically significant result, only 36% of the replication studies were significant. Importantly, it is not clear whether replication studies with non-significant results failed to replicate the original result because the original result was a false positive result or the replication study produced a false negative result. What is important is that many original studies failed to report results that future replication studies can replicate. Thus, original studies in psychology fail to demonstrate a key property of good studies that a good study should have good power to produce a significant result, if the null-hypothesis is false.

The use of actual replication studies to estimate replicability has some practical limitations. First, it is very difficult to conduct actual replications on a large scale, especially for studies that require a long time (longitudinal studies) or are very expensive (fMRI studies), or raise ethical concerns (animal research). Second, replication studies require expertise that only a few experts may have. Third, there are many reasons why a replication study might fail, and replication failures would require additional studies to examine reasons for the failure; that is, was the original result a false positive result or was the replication result a false negative? Thus, it is desirable to have an alternative method of estimating replicability that does not require literal replication studies. We see this method as complementary to actual replication studies. Neither method can be considered a gold standard to assess replicability, but converging evidence from two independent methods can be used to answer the main question of our inquiry: How replicable are published results in psychology journals?

Our approach to the estimation of replicability based on evidence from original studies is based on the concept of statistical power. Power analysis was introduced by Neyman and Pearson (1933) as a formalization of Fisher's (1926) characterization of a good experiment. Most psychologists are familiar with Cohen's (1988) suggestion that good experiments should have 80% power. 80% power implies that an original study has an 80% chance to produce a significant result, if the null-hypothesis is false. It also implies that a replication study has the same 80% chance of obtaining a significant result. For a set of studies, the average power of a set of original studies should match the success rate of replication studies. The use of significance testing for original studies implies that only significant results are interpreted as evidence for an effect. Consequently, non-significant results are either not published or not interpreted as evidence for an effect; and occasionally misinterpreted as evidence for the absence of an effect. Thus, replicability is limited to the subset of studies that produced a significant result in a seminal study. This selection for significance has important implications for the estimation of replicability. For a full set of studies that includes all non-significant and significant results, replicability is simply the percentage of studies that produced a significant result. However, for the subset of studies that produced significant results, replicability is no longer equivalent to the success rate of this subset of studies because they were selected to be significant. One solution to this problem could be to compute observed power for each study and to average observed power estimates. However, this method leads to inflated estimates of replicability because the selection for significance inflates observed power (Schimmack, 2012). Our method corrects for the bias in observed power estimates that is introduced by selecting only significant studies. As the average power of a set of studies determines the success rate for a set of exact replication studies, our method can estimate replicability for a set of studies that were selected for significance; for example, all significant results that were published in a psychology journal as evidence for a new discovery.

Definition of Replicability

There are several ways to define replicability (OSC, 2015). We define replicability as the probability of obtaining the same result in an exact replication study with the same procedure and sample sizes. As most studies focus on rejecting the null hypothesis as support for a theoretical prediction, obtaining the same result typically means obtaining a significant result again. It is important to notice one major difference between our definition of replicability and the use of statistical significance as a criterion for reproducibility in the OSC project. Our definition specifies that the replication study has the same sample size as the original study. The reason is that power changes if sample sizes change. If an original study produced a significant result with N = 30 and 50% power, the chance of replicating this result with N = 30 is 50%. The chance of replicating this result with N = 3,000 is virtually 100%. However, researchers who are trying to build on original research findings are trying to replicate the original results with similar sample sizes and would not have the resources or motivation to invest 10 times as many resources as researchers who published an original study. Thus, it is crucial for future researchers who are planning replication studies to know how likely it is that an original result can be replicated with the same amount of resources. By holding sample size constant, replicability is clearly defined. In contrast, by allowing for variable sample sizes, replication studies may simply fail to replicate an original finding because they used a smaller sample. According to our definition, replication studies with smaller samples are not proper tests of replicability (Schimmack, 2012).

Replicability: Homogeneous versus Heterogenous Sets of Studies

The relationship between statistical power and replicability is simple in the homogenous case, where the same study is repeated with different independent samples from a population. In the homogenous case, replicability is equivalent to statistical power. For a set of studies, each study has the same probability of producing a significant result and the expected value of significant results is power times the number of studies. However, this simple model cannot be used to estimate mean replicability for a heterogenous set of studies where power varies across studies as a function of varying effect size and sample sizes. For example, the 97 studies with significant results in the OSC project varied in their designs (between vs. within-subject designs), sample sizes (N = 8 to > 200,000), as well as effect sizes.

We show (see supplement for proofs) that if a single study is randomly selected for exact replication, the probability of a significant result is exactly the mean power in the population from which the study was selected. If all the studies were replicated exactly (a practical impossibility), the expected proportion of significant results would be the population mean power (see Principle 1 below for details).

For a population of studies with 80% mean power, original studies are expected to produce 80% significant results. As we define replicability as reproducing a significant result, 20% of the original studies are not eligible for a replication attempt. Importantly, our model assumes that the probability of being included in the set of studies to be replicated is also a function of power. To illustrate this, assume that 50 original studies had very low power (20%) and 50 original studies had good power (80%). Most of the low powered studies would fail to produce a significant result and would not be included in the set of studies that are subjected to a replication attempt. Thus, the proportion of high powered studies in the set of studies to be replicated is greater than in the original set of studies (50 * .80)/(50*.20 + 50*.80) = 40/50 = 80%. It follows, that selection for significance increases mean power. Whereas mean power for the original studies was 50% (50*.2 + 50*.8)/100 = .50), mean power after selection for significance is 68% (10 * .20 + 40*.8)/50 = (2 + 32)/50 = 34/50 = 68%. Thus, it is important to distinguish between mean power before selection for significance and mean power after selection for significance. In this article, we focus on mean power after selection for significance as an estimate of replicability, but we return to the estimation of mean power in the Discussion section.

Introduction of Statistical Models

The estimation of power and replicability is a new area of research and we are the first to introduce a method for the estimation of replicability. Most statistical analysis of sets of original studies aim to estimate population effect sizes for a set of conceptually related studies (e.g., a meta-analysis of clinical intervention studies). A major problem of existing meta-analyses methods is that they do not take selection for significance into account and can produce misleading results if non-significant results are not reported. Because selection for significance in journals is very common (Francis, 2012), the results of meta-analysis are typically biased. Three methods exist to estimate effect sizes for a set of studies after selection for significance (Hedges, 1984; Simonsohn, Nelson & Simmons, 2014b; van Assen, van Aert, and Wicherts, 2014). McShane, Böckenholt, and Hansen (2016) evaluated these methods using simulation studies. They found that all three methods performed satisfactory with homogenous sets of studies (i.e, a fixed population effect size), but produced biased estimates with heterogeneous sets of studies in which true effect size varied across studies. Thus, these methods are not useful for our current purpose of estimating replicability for heterogenous sets of studies.

Hedges (1992) developed the only method for effect size estimation for heterogenous sets of studies. Hedges and Vevea (1996) conducted a series of simulation studies to evaluate the method under various conditions of heterogeneity and found that the method considerably reduced bias due to selection for significance and was relatively robust to violation of the model assumptions about the distribution of population effect sizes. This method seems a promising start for our purposes, although Jerry Brunner developed our Maximum Likelihood approach before we learned about Hedge's approach. Thus, our first model uses Maximum Likelihood estimation with assumed effect size distributions to estimate replicability.

The second method uses a different approach. It sidesteps the problem of effect size estimation and uses the strength of evidence against the null-hypothesis of individual studies to estimate replicability. All significance tests use p-values as a common metric to decide whether the evidence is strong enough to reject the null-hypothesis; typically if p < .05 (two-tailed). Our method converts exact p-values into z-scores, by finding the z-score of a standard normal distribution that corresponds to the exact p-value. The distribution of z-scores is then used to estimate replicability. As the method relies on distributions of z-scores, we call it z-curve.

Notation and statistical background

To present our methods formally, it is necessary to introduce some statistical notation. Rather than using traditional notation from statistics, we use R-code to formally specify our models (R core team, 2012). This approach makes it easier for psychologists without formal training in advanced statistics to follow our methods and reproduce our results (see, Simonsohn, Nelson, & Simmons, 2014a, for a similar approach).

The outcome of an empirical study is partially determined by random sampling error, which implies that statistical results will vary across studies. This variation is expected to follow a random sampling distribution. Each statistical test has its own sampling distribution. We will use the symbol T to denote a general test statistic; it could be a t-statistic, F, chi-squared, Z, or something more obscure.

Assume an upper-tailed test, so that the null hypothesis will be rejected at significance level a (usually $a = 0.05$), when the continuous test statistic T exceeds a critical value c. Typically there is a sample of test statistic values $T_1,¼,T_k$, but when only one is being considered the subscript will be omitted. The notation $\mathrm{p}(t)$ refers to the probability under the null hypothesis that T is less than or equal to the fixed constant t. The symbol p would represent pnorm if the test statistic were standard normal, pf if the test statistic had an F-distribution, and so on. While $\mathrm{p}(t)$ is the area under the curve,

$\mathrm{d}(t)$ is height of the curve above the x-axis, as in dnorm. Following the conventions of the S language, the inverse of p is q, so that $\mathrm{p}(\mathrm{q}(t)) = \mathrm{q}(\mathrm{p}(t)) = t$.

Sampling distributions when the null hypothesis is true are well known to psychologists because they provide the foundation of significance testing. Most psychologists are less familiar with non-central sampling distributions (see Johnson, Kotz, & Balakrishnan, 1995). When the null hypothesis is false, the area under the curve of the test statistic's sampling distribution is $\mathrm{p}(t,\mathrm{ncp})$, representing particular cases like $\mathrm{pf}(t,\mathrm{df1},\mathrm{df2},\mathrm{ncp})$. The initials ncp stand for non-centrality parameter. This notation applies directly when T has one of the common non-central distributions like the non-central t, F or chi-squared under the alternative hypothesis, but it extends to the distribution of any test statistic under any specific alternative, even when the distribution in question is technically not a non-central distribution. The non-centrality parameter is positive when the null hypothesis is false, and statistical power is a monotonically increasing function of the non-centrality parameter. This function is given explicitly by Power = $1 - \mathrm{p}(c,\mathrm{ncp})$.

The non-centrality parameter can be factored into the product of two terms. The first term is an increasing function of sample size (n), and the second term is a function of the unknown parameters that can be considered a standardized effect size (es).

In symbols,

$$\mathrm{ncp} = f_1(n) \times f_2(\mathrm{es}). \tag{1}$$

While sample size is observable, effect size is a function of unknown parameters and can never be known exactly. The quantities that are computed from sample data and commonly called effect size are actually estimates of effect sizes.

As we use the term, effect size refers to any function of the model parameters that equals zero when the null hypothesis is true, and assumes larger positive values as the size of an effect (a mean difference or a covariance) becomes stronger. From this perspective, all reasonable definitions of effect size for a particular statistical model are deterministic monotone functions of one another and so the choice of which one to use is determined by convenience and interpretability. This usage is consistent in spirit with that of Cohen (1988), who freely uses the term effect size to describe various functions of the model parameters, even for the same statistical test (see also Grissom & Kim, 2012).

As an example of Equation (1), consider a standard F-test for difference between the means of two normal populations with a common variance. After some simplification, the non-centrality parameter of the non-central F may be written

$$\mathtt{ncp} = n r (1- r) d^2,$$

where $n = n_1 + n_2$ is the total sample size, $r = \frac{n_1}{n}$ is the proportion of cases allocated to the first treatment, and $d = \frac{|m_1 - m_2|}{s}$ is Cohen's (1988) effect size for the two-sample problem. This expression for the non-centrality parameter can be factored in various ways to match Equation (1); for example, $f_1(n) = n r (1- r)$ and $f_2(\mathtt{es}) = \mathtt{es}^2$. Equation (1) applies to the non-centrality parameters of the non-central Z, t, chi-squared and F distributions in general. Thus for a given sample size and a given effect size, the power of a statistical test is

$$\text{Power} = 1- \mathtt{p}(c, f_1(n) \times f_2(\mathtt{es})). \qquad (2)$$

The function $f_2(\mathrm{es})$ is particularly convenient because it will accommodate any reasonable definition of effect size. Details are given in the technical supplement.

Two Populations of Power

Consider a population of independent statistical tests. Each test has its own power value, a true probability of rejecting the null hypothesis determined by the sample size, procedure, and true parameter values. Once tests are conducted, there are two sets of studies. Some produced significant results and some produced non-significant results. We are only considering studies that produced significant results. This selection for significance does not change the power values of individual studies. However, the population of studies in the set of studies selected for significance differs from the original population of studies without selection for significance. To better understand the implications of selection for significance, it is helpful to think about studies as games of chance. Designing a study and selecting a hypothesis to test corresponds to manufacturing a roulette wheel that may not be perfectly balanced. The numbers on the wheel are p-values, and $p < 0.05$ is a win. Running the study and collecting data corresponds to spinning the wheel. The unique balance and other physical properties of the wheel determine the probability of a win; this corresponds to the power of the test. Performing the statistical analysis corresponds to examining the number that comes up on the wheel and noting whether $p < 0.05$. A large number of wheels are manufactured and spun once. This is the population before selection. The wheels that yield wins are put on display; this is the population after selection. Naturally, there is a tendency for wheels with a higher chance of winning to be put on display. The wheels that yield losing numbers are sent to warehouses.

Spinning all the wheels on display a second time would produce winners and losers. Replicability is the percentage of wheels that produce wins during this second spin. If the number of wheels on display were large, the percentage of wins would give us a close estimate of the true average probability of winning for the set of wheels on display. Spinning all the wheels a third time would yield a similar percentage and repeating this exercise many times and averaging the proportions would give the true probability of a win for the wheels on display. The objective of this paper is to estimate this unknown quantity using only the results of the first spin for the wheels that produced a winning number on the first spin.

We now give a set of fundamental principles connecting the probability distribution power before selection for significance to its distribution after selection for significance. These principles do not depend on the particular population distribution of power, the significance tests involved, or the Type I error probabilities of those tests. They do not even depend on the appropriateness of the tests or the assumptions of the tests being satisfied. The only requirement is that each power value in the population is the probability that the corresponding test will be significant. The supplemental materials contain proofs and a numerical example.

Principle 1 *Population mean power equals the overall probability of a significant result.* Principle 1 applies equally to the population of studies before and after selection. Because it applies after selection, this principle establishes the link between replicability and population mean power. If a single published result is randomly selected and the study is repeated exactly, the probability of obtaining another significant result equals population mean power after selection. In terms of the roulette wheel analogy, this is a two-stage game. The first stage is to select a wheel at random from those on display, and the second stage is to spin the wheel. Principle 1 says that the probability of winning the game is exactly the mean probability of a win for the wheels on display.

Principle 2 *The effect of selection for significance is to multiply the probability of each power value by a quantity equal to the power value itself, divided by population mean power before selection. If the distribution of power is continuous, this statement applies to the probability density function.*

In the technical supplement, Principle 2 is used to derive Principle 3.

Principle 3 *Population mean power after selection for significance equals the population mean of squared power before selection, divided by the population mean of power before selection.*

Maximum likelihood replicability estimation

The method of maximum likelihood (Fisher, 1922; also see the historical account by Aldrich, 1997) is a general method for the estimation of an unknown parameter by finding the parameters value that makes the observed data most probable. For any set of observed data, the statistical assumptions allow calculation of the probability of obtaining the observed the data (or for continuous distributions, the probability of obtaining data in a tiny region surrounding the observed data). The *likelihood function* expresses this probability as a function of the unknown parameter. Geometrically, the likelihood function is a curve, and estimation proceeds by finding the highest point on the curve. The maximum likelihood estimate is the parameter value yielding that maximum. The case of multi-parameter estimation is analogous, with the curve being replaced by a convoluted surface in higher dimension. When data are consistent with the model assumptions, maximum likelihood generally yields more precise parameter estimates than other methods, especially for large samples (Lehmann & Casella, 1998).

For simplicity, first consider the case of heterogeneity in sample size but not effect size. In this case the single unknown parameter is the effect size $\mathtt{es}$, and the likelihood function is based on the conditional probability of observing the data given selection for significance. Denoting the observed test statistic values by $t_1, ¼, t_k$, the likelihood function is a product of k terms of the form

$$\frac{\mathrm{d}(t_j, f_1(n_j) \times f_2(\mathtt{es}))}{1 - \mathrm{p}(c_j, f_1(n_j) \times f_2(\mathtt{es}))}, \tag{1}$$

where because of selection for significance, all the t_j values are greater than their respective critical values c_j. Expression (1) becomes the likelihood of Hedges (1984) for the case of a two-sample t-test; see the technical supplement for an example. In general, the maximum likelihood estimate of $\mathtt{es}$ is the effect size value that makes the likelihood function greatest. Denote it by $\widehat{\mathtt{es}}$. The estimated probability of significance for each study is obtained by

$$\text{Estimated Power} = 1 - \mathrm{p}(c_j, f_1(n_j) \cdot f_2(\widehat{\mathtt{es}})),$$

and then the estimated power values are averaged to produce a single estimate of mean power.

Now include heterogeneity in effect size as well as sample size. If sample size and effect size before selection are independent, selection for significance induces a mild relationship between sample size and effect size, since tests that are low in both sample size and effect size are under-selected, while tests high in both are over-selected. Suppose that the distribution of effect size before selection is continuous with probability density $g_q(\mathrm{es})$. This notation indicates that the distribution of effect size depends on an unknown parameter or parameter vector q. In the technical supplement, it is shown that the likelihood function (a function of q) is a product of k terms of the form

$$\frac{\int_0^\infty \mathrm{d}(t_j, f_1(n_j)\cdot f_2(\mathrm{es}))g_q(\mathrm{es})d\mathrm{es}}{\int_0^\infty \left[1- \mathrm{p}(c_j, f_1(n_j)\cdot f_2(\mathrm{es}))\right]g_q(\mathrm{es})d\mathrm{es}}, \tag{2}$$

where the integrals denote areas under curves that can be computed with R's integrate function. Again, the maximum likelihood estimate is the value of q for which the value of the product is highest. Denote the maximum likelihood estimate by $\hat{q}$. Typically $\hat{q}$ is a single number or a pair of numbers.

As before, an estimate of population mean power is produced by averaging estimated power for the k significance tests. It is shown in the technical supplement that the terms to be averaged are

$$\frac{\int_0^\infty \left[1- \mathrm{p}(c_j, f_1(n_j)\cdot f_2(\mathrm{es}))\right]^2 g_{\hat{q}}(\mathrm{es})d\mathrm{es}}{\int_0^\infty \left[1- \mathrm{p}(c_j, f_1(n_j)\cdot f_2(\mathrm{es}))\right]g_{\hat{q}}(\mathrm{es})d\mathrm{es}},$$

an expression that also follows from an informed application of Principle 3.

Z-curve

Z-curve follows a traditional meta-analysis that converts p-values into Z-scores as a common metric to integrate results from different original studies (Stouffer, Suchman, DeVinney, Star and Williams, 1949; Rosenthal, 1979). The use of Z-scores as a common metric makes it possible to combine results from widely different statistical methods and tests. The method is based on the simplicity and tractability of power analysis for the one-tailed Z-test, in which the distribution of the test statistic under the alternative hypothesis is just a standard normal shifted by a fixed quantity that we will denote by m (Heisey & Hoenig, 2001). As described in the technical supplement, m is the non-centrality parameter for the one-tailed Z-test. Input to the Z-curve is a sample of p-values from two-sided or other non-directional tests, all less than $a = 0.05$. These p-values are processed in several steps to produce an estimate.

1. *Convert p-values to Z-scores*. The first step is to imagine, for simplicity, that all the p-values arose from two-tailed Z-tests in which results were in the predicted direction. This is equivalent to an upper-tailed Z-test with significance level $a / 2 = 0.025$. The conversion to Z-scores (Stouffer et al., 1949) consists of finding the test statistic Z that would have produced that p-value. The formula is $Z = \texttt{qnorm}(1- \ p / 2)$.

2. *Set aside $Z > 6$*. We assume that p-values in this range come from tests with power essentially equal to one. To avoid numerical problems arising from p-values that are approximately zero, we set them aside for now and bring them back in the final step.

3. *Fit a finite mixture model*. Before selecting for significance and setting aside values above six, the distribution of the test statistic Z given a particular non-centrality parameter value m is normal[1] with mean m. Afterwards, it is a normal

[1] This statement would be exactly true if the p-values really came from one-sided Z-tests as suggested in Step 1. In practice it is an approximation.

distribution truncated on the left at the critical value c (usually 1.96) truncated on the right at 6, and re-scaled to have area one under the curve. Because of heterogeneity in sample size and effect size, the full distribution of Z is an average of truncated normals, with potentially a different value of μ for each member of the population. As a simplification, heterogeneity in the distribution of Z is represented as a finite mixture with r components. The model is equivalent to the following two-stage sampling plan. First, select a non-centrality parameter μ from $\mu_1,\ldots,\mu_r$ according to the respective probabilities $w_1,\ldots,w_r$. Then generate Z from a normal distribution with mean μ and standard deviation one. Finally, re-scale so that the area under the curve equals one. Under this approximate model, the probability density function of the test statistic after selection for significance is

$$f(z)=\sum_{j=1}^{r} w_j \frac{\mathtt{dnorm}(z-\mu_j)}{\mathtt{pnorm}(6-\mu_j)-\mathtt{pnorm}(c-\mu_j)}, \qquad (3)$$

for c < z < 6.

For the sake of comparing predicted and observed distributions of z-scores, distributions are fitted using a kernel density estimate (Silverman, 1986) as implemented in R's density function, with the default settings.

Specifically, the fitting step proceeds as follows. First, obtain the kernel density estimate based on the sample of Z values between z = 2 and z = 6 and re-scale it so that the area under the curve between z = 2 and z = 6 equals one. Call this the *conditional density estimate*. Next, calculate the conditional density estimate at a set of equally spaced points ranging from 2 to 6. Then, numerically choose w_j and μ_j

values so as to minimize the sum of absolute differences between the conditional density estimate and Expression (3).

4. *Estimate mean power for* $Z < 6$. The estimate of rejection probability upon replication for $Z < 6$ is the area under the curve above the critical value, with weights and non-centrality values from the curve-fitting step. The estimate is

$$\ell = \sum_{j=1}^{r} \hat{w}_j (1 - \texttt{pnorm}(c - \hat{m}_j)), \tag{4}$$

where $\hat{w}_1, \ldots, \hat{w}_r$ and $\hat{m}_1, \ldots, \hat{m}_r$ are the values located in Step 3. Note that while the input data are censored both on the left and right as represented in Formula (3), there is no truncation in Formula (4) because it represents the distribution of Z upon replication.

5. *Re-weight using* $Z > 6$. Let q denote the proportion of the original set of Z statistics with $Z > 6$. Again, we assume that the probability of significance for those tests is essentially one. Bringing this in as one more component of the mixture estimate, the final estimate of the probability of rejecting the null hypothesis for exact replication of a randomly selected test is

$$\begin{aligned} Z_{est} &= (1 - q)\ell + q \cdot 1 \\ &= q + (1 - q)\sum_{j=1}^{r} \hat{w}_j (1 - \texttt{pnorm}(c - \hat{m}_j)) \end{aligned}$$

By Principle 3, this is the estimate of population mean power after selection for significance.

Simulations

The simulations reported here were carried out using the R programming environment (R Core Team, 2012) distributing the computation among 70 quad core Apple iMac computers. The R code is available in the supplemental materials. In the simulations, the estimation methods were applied to samples of significant chi-squared or F statistics, all with $p < 0.05$. This covers most cases of interest, since t statistics may be squared to yield F statistics, while Z may be squared to yield chi-squared with one degree of freedom.

S1: Heterogeneity in Both Sample Size and Effect Size

To model heterogeneity in effect size, we sampled effect sizes before selection from a gamma distribution (Johnson, Kotz, & Balakrishnan, 1995). Sample sizes before selection were sampled from a Poisson distributed with a population mean of 86. For convenience, sample size and effect size were independent before selection.

The simulation study varied the amount of heterogeneity in effect sizes (standard deviation of effect size after selection 0.10, 0.20 or 0.30), true population mean power (0.25, 0.50 or 0.75), number of test statistics upon which estimates of mean power are based (*k* = 100, 250, 500, 1,000 or 2,000), type of test (F or chi-squared), and experimental degrees of freedom (1, 3 or 5). Within each cell of the design, ten thousand significant test statistics were randomly generated, and population mean power was estimated using all four methods. Results for the manipulation of test statistic and experimental degrees of freedom were very similar (see supplemental material). Thus, we present results for F-tests with numerator $df = 1$, which is the most commonly used test in psychological research. Table 1 shows that both methods produce good estimates of the true parameter used for the simulation and similar variation in these estimates across simulation studies. Table 2 shows the mean absolute error of estimation for a single simulation study. With 1,000 test statistics both methods have a practically negligible absolute error. The results of the first simulation study show that both methods can estimate replicability for heterogeneous sets of studies.

S2: Simulation of Complex Heterogeneity

In the preceding simulation, heterogeneity in effect size before selection was modeled as a gamma distribution, with effect size independent of sample size before selection. The use of a gamma distribution gave Maximum Likelihood (ML) an unfair advantage because the simulated distribution matches the assumed distribution. To examine the robustness of ML, we conducted a second simulation study in which the simulated distribution of effect sizes differed from the gamma distribution that ML uses to estimate replicability. A second goal of Study 2 was to examine how correlation between effect size and sample size might influence replicability estimates. ML assumes that effect sizes and sample sizes are independent. In contrast, z-curve does not make any assumptions about the correlation between effect sizes and sample sizes. We limited this simulation to F-tests with one numerator degree of freedom because the previous simulations showed that that the test-statistic and degrees of freedom had practically no effect on the results.

In this simulation, effect size after selection had a beta distribution rather than a gamma distribution. A beta distribution is limited to values between zero and one and thus lacks the long right tail of a gamma distribution, but a value of one is considerably above Cohen's (1988, p. 287) large effect size of $\mathbf{f} = 0.4$. We made sample size and effect size non-independent by connecting them by a Poisson regression. This created varying population correlations between sample sizes and effect sizes across sets of simulated studies. We believe that a negative correlation between sample size is expected because researchers would naturally tend to use larger samples when they expect smaller effects. This is evident in the OSC (2015) studies, where studies from cognitive psychology had larger effects and smaller samples than studies from social psychology.

In the simulations, the variance of effect size after selection was fixed at 0.30, the high heterogeneity value in the preceding simulation study. Sample size after selection was Poisson distributed with expected value $\exp(b_0 + b_1 es)$. Mean effect size after selection and the parameters b_0 and b_1 were selected to achieve (a) desired population mean power after selection, (b) desired population correlation between effect size and sample size after selection, and (c) population mean sample size of 86 after selection at the mean effect size. Details are given in the technical supplement.

Three values of population mean power (0.25, 0.50 and 0.75), five values of the number of test statistics *k* (100, 250, 500, 1000 and 2000) and five values of the correlation between sample size and effect size (0.0, -0.2, -0.4, -0., -0.8) were varied in a factorial design, with ten thousand simulated data sets in each combination of values. All four estimation methods were applied to each simulated data set, with three random starting values for maximum likelihood.

Table 3 shows means and standard deviations of estimated population mean power as a function of true population mean power and the standard deviation of effect size. We were surprised to see that a correlation between sample sizes and effect sizes had little effect on the ML results. Although unexpected, this result suggests that it is permissible to assume independence between effect sizes and sample sizes and to apply ML to datasets, in which sample sizes and population effect sizes may be correlated. However, changing the simulation of the distribution of effect sizes lead to less accurate estimates for ML, whereas this change did not affect z-curve because it does not make any assumptions about the distribution of effect sizes. Table 4 confirms the differences between the two methods with mean absolute error of estimation. Z-curve produced more accurate estimates than MLRE. In conclusion, our results confirm Hedges and Vevea's (1996) findings for effect size estimation, that ML estimates are relatively robust against violations of distribution assumptions. However, we also show that a new method that does not require this assumption produced more accurate results. Based on these results, we recommend z-curve as a viable method to estimate the average replicability of sets of studies.

A conservative bootstrap confidence interval for z-curve

Point estimates should always be accompanied by information about the precision of the estimate. In order to provide this information, we developed a bootstrap confidence interval (Efron 1981, Efron & Tibshirani, 1993). To create a confidence interval for z-curve estimates, we resampled z-scores 500 times with replacement and computed replicability estimates for each sample. We used the histogram of the resulting values as an approximation to the sampling distribution of the statistic. The 95 percent bootstrap confidence interval ranges from the 2.5 percentile to the 97.5 percentile of the 500 estimates.

Especially when samples are small, it is important to verify that a 95% confidence interval contains the true value 95% of the time. This is called the *coverage* of the confidence interval. A first set of simulation studies showed that the coverage of the 95% bootstrap confidence interval was sometimes less than 95%. To avoid this problem, we created a conservative bootstrap interval by decreasing the lower limit by 0.02 and increasing the upper limit by 0.02. This yields our *conservative bootstrap confidence interval*. We tested the conservative bootstrap confidence interval in the setting of full heterogeneity, with 10,000 simulated datasets in each combination of three values of true population mean power, and seven values of the number of test statistics, ranging from k = 25 to k =2,000. Table 5 shows the coverage values. Except for very small sets of studies (k = 25), coverage exceeds the nominal value of 95%. Coverage is typically much higher than this value, which confirms its conservative nature.

Table 15 shows mean upper and lower confidence limits. The upper limit is the top number in each cell, and the lower limit is the bottom number. For example, when the true population mean power is 0.50 and the z-curve estimate is based on k =100 test statistics, the average confidence interval will range from 0.36 to 0.67. In contrast, a set of 1,000 studies produces a 95% confidence interval that ranges from .42 to .56. For small sets of studies, actual replication studies would produce a narrower confidence interval than our statistical estimation method. However, the advantage of our statistical method is that it is much easier to get a large sample of original test statistics.

Application to the Replication Project

Of the 100 original studies in the OSC (2015) Replication Project, three were null results (failures to reject the null hypothesis), and in an additional four studies the original result was only ``marginally" significant, with p-values ranging from 0.051 to 0.073. These were set aside, because technically these studies did not reject the null hypothesis. Of the remaining 93 studies, five were eliminated because the replication studies were not based on an independent sample or had other unusual characteristics. This left a sample size of k = 88 studies. The success rate of replication studies for our set of 88 studies was 39%, which is close to the 36% success rate for the full set of 97 studies.

Most of the test statistics for the originally reported tests were F or chi-squared. The rest were converted by squaring t statistics to obtain Fs, and squaring Z statistics to obtain chi-squared with one degree of freedom. Input to z-curve was simply the set of p-values. For ML, test statistics were divided into subsets according to the type of test (F or chi-squared) and the (numerator) degrees of freedom. Estimates were calculated for each subset, and then combined as a weighted sum, using the observed proportions of the subsets as weights.

The ML estimate of replicability was 59%. The estimate for z-curve was 66%. However, these point estimates have to be interpreted with caution given the small number of studies. The 95% confidence interval for z-curve ranged from 49% to 79%. Moreover, the 39% of actual successful replications is also influenced by sampling error and the 95% interval around this estimate of replicability ranges from 29% to 49%. Although it is not possible to know replicability precisely based and difficult to quantify the difference between the two estimation methods (actual studies vs. statistical estimation), the results suggest that our statistical approach provides optimistic estimates of replicability. We discuss the reasons for this optimistic bias in the Discussion section.

Discussion

The replicability of psychological research has been questioned. In this article, we introduced two methods that can be used to estimate replicability for a heterogeneous set of studies that reported a significant result. One method relies on Maximum Likelihood estimation. The other method relies on the distribution of z-scores. Although both methods produced reasonable estimates, z-curve performed slightly better because it makes no assumption about the distribution of effect sizes. Based on simulation studies, we showed that z-curve provides accurate estimates of replicability for heterogeneous sets of studies with significant results. It is also the most convenient method because it requires only p-values as input, and p-values are easy to obtain from reported test statistics in published articles. P-values are even available for methods without explicit sampling distributions. We also developed a conservative bootstrap confidence interval that makes it possible to demonstrate reliable differences in replicability across different sets of studies or to test whether a set of studies is consistent with Cohen's recommended power of 80% (Cohen, 1988). We applied these methods to test statistics from original studies that were replicated in the OSC reproducibility project. As replicability predicts the success rate of exact replication studies, we used the success rate in the replication studies to compare our results to the replicability estimate based on actual replication studies. Whereas 39% of 88 original studies were replicated, our statistical methods predicted success rates of 59% for MLRE and 66% with z-curve.

Problem of Conducting Exact Replication Studies

The most obvious explanation for the lower success rate in the OSC replication studies is that it is virtually impossible to exactly replicate original studies in psychology. Not surprisingly, several articles have pointed out that the 36% success rate may severely underestimate replicability of original studies. Gilbert, King, Pettigrew, and Wilson (2016) distinguished studies that were exact replications and studies that differed substantially from the original studies (e.g., a study conducted in the United States was replicated in another country). They found that close replications were four times more likely to replicate than studies with some notable differences. Daniel Gilbert even questioned whether the OSC studies can be considered replication studies. “If you read the reports carefully, as we did, you discover that many of the replication studies differed in truly astounding ways — ways that make it hard to understand how they could even be called replications” (Reull, 2016). Given this serious criticism of the reported 36% success rate, it is extremely valuable to have an alternative estimate of replicability that avoids the problems of conducting actual replication studies by using the reported results of original studies. This method is immune to criticisms about experimental procedures of actual replication studies because the estimate relies on the original studies. Our simulation studies show that our method produces accurate estimates of replicability, and our results suggest that nearly a third of the failed replication studies might be due to problems with conducting exact replication studies (66% - 39% = 27%). Moreover, the upper limit of the 95% confidence interval for this small set of studies is 79%, which is just shy of Cohen’s recommended level of 80%. At the same time, the lower limit of the 95% confidence interval is 49%. This is slightly better than the 36% estimate reported in the OSC article, but it would suggest that psychologists need to increase power to ensure that actual replication studies can successfully reproduce a significant result. In the absence of more precise estimates, it is important to realize that the OSC results and our results are both consistent with Cohen’s seminal work on power, which suggested that the typical published study in psychology has 50% power to discover an average effect size (Cohen, 1962; Sedlmeier & Gigerenzer, 1989). Thus, we think the existing evidence suggests that replicability in psychology is about 50%, but this estimate comes with a wide margin of error. Future research with larger and more representative samples of studies are needed to obtain more precise estimates of replicability. More important, all results confirm that psychological studies often have insufficient power, especially when the population effect size is small. Although the past five years have seen a searching concern about false positive results, we would like to remind readers that low power can also produce false negatives in original studies and in replication studies. Thus, increasing statistical power in original studies remains a priority for psychologists for two reasons. First, high statistical power is needed to avoid false negatives in original studies. Second, high statistical power is needed to avoid false negatives in replication studies.

The File Drawer Problem

The file drawer problem (Rosenthal, 1979) provides an alternative explanation for the discrepancy between our statistical predictions of replicability and the actual success rate in the OSC project. To appreciate this fact, it is important to recognize the distinction between selection for significance in original studies and file drawers. Our model assumes that selection for significance occurs only once when researchers conduct a seminal study of a new prediction (e.g., the first study on the effect of Mozart music on performance on an IQ test). If this study fails, the research question is abandoned and nobody else does a second attempt to test it. This selection model ensures that high powered studies are much more likely to end up in the published literature and are subjected to a replication attempt than low powered studies. The model would also assume that subsequent (conceptual or exact) replication studies report all results whether they are significant or not. However, this assumption is inconsistent with a 95% success rate in published journals that report not only original, new studies, but also conceptual replication studies. Ample evidence indicates that the high success rate of conceptual replication studies in published articles is due to missing studies with non-significant results (Francis, 2012; Schimmack, 2012; Sterling et al., 1995). This has consequences for the replicability of published studies because repeated tests of the same hypothesis increase the probability that low powered studies are published. This can be seen by imagining that researchers continue to repeat the same study until it is significant and then only the significant result is published. In this scenario, high powered studies no longer have an advantage to be selected into the set of published studies with significant results. Moreover, as studies with high power require more resources, low powered studies might even be overrepresented in the set of published studies relative to high powered studies. McShane et al. (2016) point out that the selection model has a strong influence on estimates. It is therefore important to recognize that our model assumes that there is no file-drawer of already conducted replication studies. If such a file drawer exists, our method overestimates replicability.

Some researchers may not be aware that even conceptual replications contribute to the file drawer problem. For example, Gilbert and Wilson (2015) describe their work that led to a published article that reported only significant results. “We did some preliminary studies that used different stimuli and different procedures and that showed no interesting effects. Why didn’t these studies show interesting effects? We’ll never know. Failed studies are often (though not always) inconclusive, which is why they are often (but not always) unpublishable. So yes, we had to mess around for a while to establish a paradigm that was sensitive and powerful enough to observe the effects that we had hypothesized." Gilbert and Wilson recognize that it would be unethical to run exact replication studies until one of them produced a significant result. However, they do not recognize that running conceptual replication studies until one of them works also increases the chances that a low powered study produces a significant result that will be difficult to replicate in future studies. Thus, research practices like the one described by Gilbert and Wilson can also account for the discrepancy between the success rate of OSC replication studies and our statistical estimate of replicability. Our estimates are best considered the best-case scenario, assuming exact replications and no file drawer. Given these idealistic assumptions, it is probably not surprising that the actual success rate was considerably lower.

Replicability and False-Positives

A common mistake is to interpret non-significant results as evidence for the absence of an effect (i.e., the null-hypothesis is true). This mistake is often made in original articles, but it has also occurred in the interpretation of the OSC results. Sixty-four percent of the OSC replication studies produced a non-significant result. It is difficult to interpret these non-significant results because there are two alternative explanations for a non-significant result in a replication study. Either the replication study produced a true-negative result and the original study reported a false positive result, or the replication study produced a false negative result and the original study produced a true positive result. The 64% failure rate in the OSC project might be misinterpreted as evidence that 64% of original significant results were false positive results. This interpretation would be a mistake because it ignores the possibility that original studies correctly rejected the null hypothesis and replication studies produced false negative results. One reason for the misinterpretation of the OSC results could be that the article claims that the replication studies used "high-powered designs" (aac4716-1) and that Table 1 suggests that replication studies had over 90% power. It is tempting to infer from a non-significant result in a study with 90% power that a non-significant result can be interpreted as evidence for H_0. However, this is not a valid inference because statistical power depends on the specification of the alternative hypothesis. The OSC replication studies had over 90% power to produce a significant result if the effect size estimate of the original study matched the population effect size. Many of the effect sizes in the original studies were moderate or large (see Cohen, 1988). A non-significant result in replication studies can be used as evidence that the population effect size is likely to be smaller than these effect sizes, but it cannot be used to infer that the effect size is zero.

Further confusion is created by a statistical test of the distribution of non-significant p-values in the OSC project. If all of these original significant results had been type-I errors, the distribution of p-values would be uniform. The authors tested the distribution of p-values and found that "it deviated slightly from uniform with positive skew" and that this deviation was statistically significant with p = .048. This finding can be used to reject the hypothesis that all non-significant results were false positives. However, the OSC article makes the misleading statement that "nonetheless, the wide distribution of P values suggests against insufficient power as the only explanation for failures to replicate" (p. aac4716-3). This statement implies that at least some of the failed replication studies revealed false positive results in the original studies. However, a non-significant deviation from a uniform distribution cannot be used to infer that all or most of the non-significant results are false positive results. Once more, this interpretation would make the mistake of interpreting a non-significant result as evidence for the absence of an effect.

We recommend distinguishing between replicability and detection of false positives. First, exact replication studies with the same sample size as original studies can be used to examine the replicability of published studies. Significant results strengthen the credibility of a published result and a series of independent, mostly successful replication studies can be used to establish an original finding as a true discovery. If replication studies fail most of the time, the probability that an original result was a false positive result increases. In this case, it may be necessary to conduct studies with high precision to examine whether an estimate of the population effect size is sufficiently close to zero to affirm the null-hypothesis. For example, registered replication reports have sample sizes that are large enough to provide positive evidence for null-effects (Simons, Holcombe, Spellman, 2014).

Other Measures of Replicability

Our method focuses on statistical significance as the criterion concluding that a result has been replicated. Z-curve estimates the probability of replicating a randomly chosen significant result in an exact replication study. The OSC (2015) proposed several additional ways to compare the results of original and replication studies. One method compares replication estimated effect sizes to original estimated effect sizes. The authors propose to examine whether the estimated effect size of an original study is within the 95% confidence interval of a replication study. We think that this definition of replicability has numerous limitations. First, a 95% confidence interval depends on sample size. As the sample size of the replication study increases (while of course the sample size of the original study remains constant), sampling error decreases. The method will eventually show that none of the original results could be replicated because the original effect size is contaminated with sampling error and will never match the population effect size exactly. Conversely, a replication study with a very small sample size will have a very wide confidence interval, one that is likely to include the estimated effect size. Thus, the success rate of replication studies depends on sample size, while perversely encouraging under-powered replication studies to demonstrate high replicability. More important, the comparison of effect sizes has no direct relationship to hypotheses. Two effect sizes can differ significantly from each other without leading to a theoretical contradiction, if both effect sizes are in the same direction and both effect sizes are within a range that is predicted by a theory. In our opinion, a definition of replicability that is not tied to a theoretically meaningful outcome of the original study is not particularly informative.

The second approach compares mean estimated effect sizes of original and replication studies. This criterion has similar problems as the previous criterion. Most important, a significant difference in mean effect sizes may have no theoretical implications. For example, the hypothesis "money buys happiness" is supported by a correlation of $r = .1$ or $r = .2$. At best, the comparison of mean effect sizes may be used to detect publication bias. However, even for this purpose the comparison of observed means is problematic. The reason is that it is important to distinguish between two selection mechanisms that have the same effect on the average effect size in a set of replication studies. The mere selection of significant studies to examine replicability will lead to inflated observed effect sizes in the set of studies with significant results that are replicated. Thus, it is practically guaranteed that the average observed effect size of the replication studies will be lower than the observed mean of the original studies. Another reason for lower means in replication studies is the file drawer problem. However, there exist better methods to test whether publication bias is present that do not confound selection for significance with the file drawer problem (Francis, 2012; Schimmack, 2012).

The third statistical criterion is based on a meta-analysis of the original and replication studies. The main problem with this approach is the file drawer problem. Just like large meta-analysis, even a meta-analysis of two studies is biased if additional tests with non-significant results are missing. So, unless the original study is the only study that has been conducted, a meta-analysis needs to take publication bias into account. However, more than two studies are needed to test for bias in meta-analyses and a failed replication study indicates that more research is needed even if the combined results produce a non-significant result.

In conclusion, the main goal of original research articles is to test theoretical predictions about cause-effect relationships. The most common approach to drawing conclusions from an empirical study is to conduct a hypothesis test and to reject the null-hypothesis. Psychological journals report thousands of these tests each year. Given the importance of statistical inference for theory development, we believe that it is important to examine how replicable a statistically significant result in a published article is. Cohen recommended that researchers should plan studies to have 80% power. We believe that 80% is also a reasonable goal for the success rate of replication studies if original studies were carefully planned studies that tested a theoretically important prediction.

Future Directions

In future research, it will be interesting to develop different methods for the estimation of replicability assuming the presence of file-drawers with failed conceptual replication studies. It will also be interesting to examine how robust our methods are when researchers use a variety of questionable research practices to produce significant results (John et al., 2012). One advantage of our method is that it can be used for large sets of studies from diverse areas of psychology. Thus, it can be used to examine differences in replicability across disciplines. For example, the OSC project provided preliminary evidence that cognitive psychology is more replicable than social psychology. Our method can be used to test this hypothesis with much larger sets of studies. Our method can also be used to examine whether recent efforts to improve replicability of psychology have produced positive results. For example, we can compare a representative sample of studies in 2008 to a representative sample of studies in 2016. Finally, our method can be used to plan sample sizes of replication studies. One main problem of the OSC project was the use of observed power to plan sample sizes of replication studies. Our method can correct for the inflation in effect sizes that is introduced by selection for significance to ensure that replication studies have adequate power to avoid false negative results in replication studies.

References

Aldrich, J. (1997). R. A. Fisher and the Making of Maximum Likelihood 1912-1922. *Statistical Science*, 12, 162-176.

Baker, M. (2016) Is there a reproducibility crisis? *Nature* 533, 452454.

Becker, R. A., Chambers, J. M. and Wilks, A. R. (1988). *The new S Language: A programming environment for data analysis and graphics.* Pacific Grove, California: Wadsworth& Brooks/Cole.

Begley, C.G., (2013) Reproducibility: Six red flags for suspect work. *Nature* 497, 433434.

Begley, C, G. and Ellis, L. M. (2012) Drug development: Raise standards for preclinical cancer research. *Nature* 483, 531-533.

Billingsley, P. (1986). *Probability and measure*. New York: Wiley.

Bishop, Y. M. M., Feinberg, S. E. and Holland, M. M. (1975). *Discrete multivariate analysis.* Cambridge, Mass.: MIT Press.

Bollen, K. A. (1989), *Structural equations with latent variables,* New York: Wiley.

Boos, D. D. and Stefnski, L. A. (2012). P-value precision and reproducibility. *The American Statistician* 65, 213-221.

Chang, A. C. and Li, P. (2015) Is Economics Research Replicable? Sixty Published Papers from Thirteen Journals Say “Usually Not”, *Finance and Economics Discussion Series 2015- 083*. Washington, D.C.: Board of Governors of the Federal Reserve System.

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

Cohen, J. (1988). Statistical power analysis for the behavioral sciences. (2nd Edition), Hilsdale, New Jersey: Erlbaum.

Desu, M. M. and Raghavarao, D. (1990). *Sample size methodology*. New York: Academic Press.

Efron, B. (1981). Nonparametric estimates of standard error: The jackknife, the bootstrap and other methods. *Biometrika* 68, 589599

Efron, R. and Tibshirani, R. (1993) *An introduction to the bootstrap*. New York: Chapman and Hall.

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London. Series A* 222 309-368.

Fisher, R. A. (1926). The arrangement of field experiments. Journal of the Ministry of Agriculture of Great Britain, 33, 503–513.

Fisher, R. A. (1928). The general sampling distribution of the multiple correlation coefficient. *Proceedings of the Royal Society of London. Series A* 121, 654-673.

Gerard, P. D., Smith, D. R. & Weerakkody, G. (1998). Limits of retrospective power analysis. Journal of Wildlife Management, 62, 801 - 807.

Gilbert, D., & Wilson, T. D. (2015). Reply to Francis. Downloaded from http://www2.psych.purdue.edu/~gfrancis/Publications/ConsumingExperience/MOREWEDGEREPLY.pdf

Greenwald, A. G., Gonzalez, R., Harris, R. J., and Guthrie, D. (1996). Effect sizes and *p* values: What should be reported and what should be replicated? *Psychophysiology*, 33, 175183.

Gillett, R. (1994). Post hoc power analysis. *Journal of Applied Psychology* 79, 783785.

Grissom, R. J. and Kim, J. J. (2012). *Effect sizes for research: univariate and multivariate applications*. New York: Routledge.

Hedges, L. V. (1984). Estimation of effect size under nonrandom sampling: The effects of censoring studies yielding statistically insignificant mean differences. *Journal of Educational Statistics*, 9, 6185.

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

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

Hirschhorn, J. N., Lohmueller, K., Byrne, E., Hirschhorn K. (2002) A comprehensive review of genetic association studies. *Genetics in Medicine* 4, 4561.

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.

Ioannidis, J. P. A. (2008). Why Most Discovered True Associations Are Inflated. *Epidemiology*, 19(5), 640-646.

Ioannidis, J. P., and Trikalinos, T. A. (2007). An exploratory test for an excess of significant findings. *Clinical Trials*, 4, 245253

John, L. K., Lowenstein, G. and Prelec, D. (2012) Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science* 23 517-523

Johnson, N. L., Kotz, S. and Balakrishnan, N. (1995). *Continuous univariate distributions* (2nd. Edition). New York: Wiley.

Johnson, N. L., Kemp, A. W. and Kotz, S. (2005). *Univariate discrete distributions*. (3d Edition). Hoboken, N.J.: Wiley.

Kepes, S., Banks, G. C., McDaniel, M., and Whetzel, D. L. (2012). Publication bias in the organizational sciences. *Organizational Research Methods*, 15, 624662

Lehman, E. L. (1959). Testing statistical hypotheses. New York: Wiley.

Lehman, E. L. and Casella, G. (1998) *Theory of point estimation* (2nd. Edition). New York: Springer.

Lehman, E. L. and Romano, J. P. (2010) *Testing statistical hypotheses.* (3d Edition). New York: Wiley.

Lindsay, B. G. and Roeder, K. (2008). Uniqueness of estimation and identifiability in mixture models. *Canadian Journal of Statistics* 21, 139-147.

McCullagh, P. and Nelder, J. A. (1989) *Generalized linear models.* (2nd Edition). New York: Chapman and Hall.

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.

Patinak, P. B. (1949). The non-central χ^2 and F distributions and their applications. *Biometrika*, 36, 202-232.

Pinsky, M. A. and Karlin S. (2011). *An introduction to stochastic modeling*. San Diego: Academic Press.

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

Posavac, E. J. (2002). Using p values to estimate the probability of a statistically significant replication. *Understanding Statistics*, 1, 101112.

R Core Team (2012). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/

Reuell, P. (March 6, 2016). Study that undercut psych research got it wrong. Harvard gazette. http://news.harvard.edu/gazette/story/2016/03/study-that-undercut-psych-research-got-it-wrong/ (downloaded November 20, 2016)

Rosenthal, R. (1966) *Experimenter effects in behavioral research*. New York: Appleton-Century-Crofts.

Rosenthal, R. (1979). The File Drawer Problem and Tolerance for Null Results. *Psychological Bulletin*, 86(3), 638.

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

Silverman, B. W. (1986) *Density Estimation*. London: Chapman and Hall.

Simons, D. J., Holcombe, A. O., & Spellman, B. A. (2014). An introduction to registered replication reports at Perspectives on Psychological Science. *Perspectives on Psychological Science, 9*, 552-555.

Simmons, J. P., Nelson, L. D. and Simonsohn, U. (2011) False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science* 22 1359-1366.

Simonsohn, U., Nelson, L. D. and Simmons, J. P. (2014a). *P*-curve: A key to the file drawer. *Journal of experimental psychology: General*, 143, 534-547.

Simonsohn, U., Nelson, L. D. and Simmons, J. P. (2014b). *p*-Curve and Effect Size: Correcting for Publication Bias Using Only Significant Results. *Perspectives on Psychological Science*, 9, 666-681.

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. and Weinkam, J. J. (1995). Publication decisions revisited: The effect of the outcome of statistical tests on the decision to publish and vice versa. *The American Statistician* 49, 108-112.

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.

Stuart, A. and Ord, J. K. (1999). *Kendall's Advanced Theory of Statistics, Vol. 2: Classical Inference & the Linear Model* (5th ed.). New York: Oxford University Press.

Thomas, L. (1997) Retrospective Power Analysis. *Conservation Biology* 11, 276-280.

van Assen, M. A. L. M., van Aert, R. C. M. and Wicherts, J. M. (2014) Meta-analysis using effect size distributions of only statistically significant studies. *Psychological Methods* 1-18.

Yuan, K. H. and Maxwell, S. (2005) On the post hoc power in testing mean differences. *Journal of educational and behavioral statistics* 30, 141-167.

Table 1: Means and standard deviations of estimated power for heterogeneity in sample size and effect size based on 1,000 *F*-tests with numerator *df* = 1

	Mean			*Standard Deviation*		
Population Mean Power = 0.25						
	SD of Effect Size			*SD* of Effect Size		
	0.1	0.2	0.3	0.1	0.2	0.3
MLRE	0.230	0.269	0.283	0.069	0.016	0.015
Z-curve	0.233	0.225	0.226	0.027	0.026	0.024
Population Mean Power = 0.50						
	SD of Effect Size			*SD* of Effect Size		
	0.1	0.2	0.3	0.1	0.2	0.3
MLRE	0.501	0.502	0.506	0.025	0.019	0.019
Z-curve	0.504	0.492	0.487	0.026	0.026	0.025
Population Mean Power = 0.75						
	SD of Effect Size			*SD* of Effect Size		
	0.1	0.2	0.3	0.1	0.2	0.3
MLRE	0.752	0.750	0.750	0.022	0.017	0.014
Z-curve	0.746	0.755	0.760	0.021	0.017	0.016

Table 2: Mean Absolute Error of estimation for heterogeneity in sample size and effect size based on 1, 000 *F*-tests with numerator *df* = 1

	SD of Effect size		
	0.1	0.2	0.3
Population Mean Power = 0.25			
MaxLike	3.55	2.06	3.34
Z-curve	2.59	3.08	2.90
Population Mean Power = 0.50			
MaxLike	1.80	1.49	1.50
Z-curve	2.12	2.19	2.23
Population Mean Power = 0.75			
MaxLike	1.42	1.18	1.16
Z-curve	1.69	1.42	1.55

Table 3: Means and standard deviations of estimated power with beta effect size and correlated sample size and effect size: k = 1, 000 F-tests with numerator df = 1

	Mean					*Standard Deviation*				
Population Mean Power = 0.25										
	Correlation					Correlation				
	-0.8	-0.6	-0.4	-0.2	0.0	-0.8	-0.6	-0.4	-0.2	0.0
MaxLike	0.302	0.301	0.300	0.300	0.300	0.015	0.015	0.015	0.015	0.015
Z-curve	0.232	0.231	0.230	0.231	0.230	0.015	0.015	0.015	0.015	0.015
Population Mean Power = 0.50										
	Correlation					Correlation				
	-0.8	-0.6	-0.4	-0.2	0.0	-0.8	-0.6	-0.4	-0.2	0.0
MaxLike	0.532	0.533	0.533	0.534	0.534	0.018	0.018	0.019	0.019	0.019
Z-curve	0.493	0.494	0.495	0.495	0.495	0.023	0.023	0.023	0.023	0.023
Population Mean Power = 0.75										
	Correlation					Correlation				
	-0.8	-0.6	-0.4	-0.2	0.0	-0.8	-0.6	-0.4	-0.2	0.0
MaxLike	0.826	0.832	0.836	0.838	0.840	0.016	0.016	0.015	0.015	0.015
Z-curve	0.785	0.790	0.793	0.794	0.796	0.013	0.013	0.013	0.012	0.012

Table 4: Mean Absolute Error of estimation with beta effect size and correlated sample size and effect size: *k* = 1, 000 *F*-tests with numerator *df* = 1

	Correlation				
	-0.8	-0.6	-0.4	-0.2	0.0
Population Mean Power = 0.05					
MaxLike	5.17	5.11	5.05	5.05	5.01
Z-curve	2.37	2.41	2.47	2.48	2.50
Population Mean Power = 0.05					
MaxLike	3.25	3.34	3.42	3.43	3.46
Z-curve	1.92	1.91	1.89	1.90	1.89
Population Mean Power = 0.05					
MaxLike	7.62	8.23	8.56	8.76	8.97
Z-curve	3.51	4.01	4.27	4.43	4.59

Table 5: Coverage of the 95% conservative bootstrap confidence interval

Population	Number of Tests						
Mean Power	25	50	100	250	500	1000	2000
0.25	95.78	97.13	98.02	98.69	98.76	98.35	97.95
0.50	94.58	95.51	96.79	98.27	99.11	99.28	99.15
0.75	93.21	94.81	96.83	98.85	99.37	99.73	99.58

Table 6: Average Upper and Lower Confidence limits

Population	Number of Tests						
Mean Power	25	50	100	250	500	1000	2000
0.25	0.54	0.46	0.40	0.35	0.32	0.30	0.29
	0.06	0.09	0.11	0.14	0.16	0.17	0.17
0.50	0.76	0.71	0.67	0.62	0.58	0.56	0.55
	0.26	0.32	0.36	0.39	0.41	0.42	0.43
0.75	0.89	0.87	0.85	0.83	0.81	0.80	0.79
	0.55	0.61	0.65	0.67	0.68	0.69	0.69