How replicable is psychology? A comparison of four methods of estimating replicability on the basis of test statistics in original studies

Jerry Brunner and Ulrich Schimmack

University of Toronto

Author Note

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 Jerry Brunner, Department of Mathematical and Computational Sciences, University of Toronto Mississauga, email: brunner@utstat.toronto.edu. Supplemental materials are available online at http://www.utstat.toronto.edu/~brunner/zcurve2016.

Abstract

In the past five years, the replicability of original findings published in psychology journals has been questioned. We show that replicability can be estimated by computing the average power of studies. We then present four methods that can be used to estimate average power for a set of studies that were selected for significance: p-curve, p-uniform, maximum likelihood, and z-curve. We present the results of large-scale simulation studies with both homogeneous and heterogeneous effect sizes. All methods work well with homogeneous effect sizes, but only maximum likelihood and z-curve produce accurate estimates with heterogeneous effect sizes. All methods overestimate replicability using 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. We also validated a conservative bootstrap confidence interval that makes it possible to use z-curve with small sets of studies.

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

How replicable is psychology? A comparison of four methods of estimating replicability on the basis of test-statistics in original studies

Science is built on a mixture of trust and healthy skepticism. On one hand, scientists who read and cite published work trust the authors, reviewers, and editors to ensure that most reported results provide sufficient credible evidence based on objective empirical studies. On the other hand, scientists also insist that studies be reported with sufficient detail to reproduce them and to see whether other researchers can replicate the results. Replication studies ensure that false positives will be promptly discovered when replication studies fail to confirm the original results. 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; "A properly designed experiment rarely fails to give ... significance" (Fisher, 1926, p. 504).

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 in published articles (Hirschhorn, Lohmueller, Byrne and Hirschhorn 2002, Ioannidis 2008, Simmons, Nelson and Simonsohn 2011, Begley and Ellis 2012, John, Lowenstein and Prelec 2012, Begley 2013, Chang and Li 2015, Baker 2016). In Psychology, the Open Science Collaboration (OSR) project attempted to estimate 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 (OSR, 2015). Ninety seven percent of the replicated studies reported a statistically significant result, but only 37% of the replication studies were able to replicate this outcome. This low success rate has created heated debates, especially in social psychology where the success rate was only 25%.

The use of actual replication studies to estimate replicability has a number of limitations. First, it is practically impossible to conduct actual replications on a large scale, especially for studies that require a long time (longitudinal studies) or are very expensive (MRI studies), or raise ethical concerns (animal research). Second, actual replication studies may require expertise that only a few researchers in the world have. Third, there are many reasons why a particular replication study might fail, and replication failure would call for additional efforts to seek reasons for the failure. For these reasons, it is desirable to have an alternative method of estimating replicability that does not require literal replication. We see this method as complementary to actual replication studies. Actual replication studies are needed because they provide more information than just finding a significant result again. For example, they show that the results can be replicated over time and are not limited to a specific historic, cultural context. They also show that the description of the original study was sufficiently precise to reproduce the study in a way that it successfully replicated the original result. At the same time a statistical estimation method based on the results reported in original articles can provide information that replication studies do not provide. For example, they can show that it was highly probable or improbable that an exact replication study would be successful. This information can be helpful in the evaluation of failed replication studies. If the replicability estimate of the original study is low, it is not surprising that an actual replication study failed to produce a significant result. In contrast, if the estimated replicability was high, it suggests that the replication study was not exact or had some problems. Thus, statistical estimates of replicability and the outcome of replication studies can be seen as two independent methods that are expected to produce convergent evidence of replicability.

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 criterion of a good experiment. According to Fisher (1926, p. 504), a good experiment should rarely produce a non-significant result when the null hypothesis is false. Most psychologists are familiar with Cohen's (1988) suggestion that good experiments should have 80% power; that is 4 out of 5 replications should produce a significant result and only 1 out of 5 studies would fail to reject the false null hypothesis; that is, making a type-II error. However, in actual practice, psychologists have ignored a priori power analysis and typically conduct studies with less power (Schimmack, 2012). A common estimate of power is that average power is about 50% (Cohen 1962, Sedlmeier and Gigerenzer 1989). This means that about half of the studies in psychology have less than 50% power. Power has direct consequences for replicability because power is the long-run probability of obtaining a statistically significant result. Thus, even if a study with 30% power produced a significant result, the chance of obtaining the same result again in a replication study remains 30%.

Methodologists have wondered for a long time why researchers ignore power if power is essential for producing significant results in original studies and in replication studies (Schimmack, 2012; Sedlmeier & Gigerenzer, 1989) without a satisfactory answer. We believe one possible explanation is that researchers confuse the rate of significant results in published articles with the replicability of published findings. As power is the long-run probability of obtaining a significant result, the rate of obtained significant can be used to estimate observed power, and the success rate of original articles is over 90%. (Sterling, 1959; Sterling, Rosenbaum and Weinkam, 1995). This may create the illusion that studies have high power and nearly always produce significant results and the expectation that replication studies will be equally successful. However, Sterling et al. (1995) pointed out that the observed success rate in journals provides an inflated estimate of power, and therefore with replicability, because journals are more likely to publish significant results than non-significant results. It is well known that non-significant results often end up in Rosenthal's (1969) proverbial file drawer, but it is not known how many non-significant results remain unpublished. As a result, neither the true success rate of original studies, nor the replicability of these studies is currently known. In this article, we present four methods that can be used to estimate replicability of published studies even if published studies are selected for significance. These methods can be used to estimate the replicability of psychological research in general or the replicability of results in specific journals.

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, we define “obtaining the same result” as obtaining a significant result again. This definition ignores the sign of an effect or the pattern of a complex interaction effect, which leads to a slight over-estimation of replicability, if one were also taking the sign of an effect into account. However, this bias is small because it is very rare that a replication study produces a significant result in the opposite direction

(OSR, 2015). All four methods of estimating replicability are using the statistical evidence against the null hypothesis in original studies to estimate average power. In a simple scenario, where all studies have the same power (homogeneous case), replicability is power. However, in the more realistic and complex scenario, where studies have different power (heterogeneous case), replicability corresponds to mean power. In our technical description of the statistical methods we focus on the statistically well-defined concept of power. In the end, we use this approach to make predictions about replicability in the OSR reproducibility project.

Introduction of Statistical Methods for Power Estimation

Consider a population of significance tests in which every test has its own probability of being significant; that is, there is a population of power values. Now suppose that one finding is randomly selected from the population. The study is repeated exactly, and the same statistical test is used to test for significance in the replication study. We show that the probability of obtaining a significant result in the replication study equals the mean power of studies in the population from which the study was drawn. It is important to note that this assumes *exact* replication. If replication studies are not exact, it is possible that the population effect sizes change. However, unless there are systematic changes in designs (e.g., replication studies increased sample sizes, which increases power), these changes will average out and power estimates are still good predictors of average power in studies that are not exact replications. As mentioned earlier, estimating average power is trivial when all studies are reported. In this case, the rate of significant results provides a simple and accurate estimate of average power. However, it is more difficult to estimate average power for a set of studies that were selected to produce significant results. In this case, the average rate of significant results and average observed power yield inflated estimates of power. Thus, the main goal of our article is to introduce and evaluate four different methods that can estimate the average power of a set of studies that were selected for significance.

Two of the methods we consider -- Simonsohn, Nelson and Simmons' (2014b) p-curve and van Assen, van Aert, and Wicherts' (2014) p-uniform -- were developed to correct for publication bias in meta-analyses of effect sizes. Both methods assume a fixed population effect size. Simonsohn et al. have extended their method to estimate power in the restricted setting of a single fixed power value for the entire population, implying homogeneity in sample size as well as effect size (`www.p-curve.com`). However, we predict that these methods will not perform well when there is moderate to large heterogeneity in effect sizes.

We introduce two additional methods that are explicitly designed to estimate power for sets of heterogeneous studies such as the set of studies that were replicated in the OSC reproducibility project. We use extensive simulation studies to compare all four methods for a wide variety of scenarios. Finally, we apply all four methods to the studies from the OSC reproducibility project, where the success rate of actual replication studies can be used as a validation criterion for the four statistical estimation methods.

We predicted that each method would do well in simulations that match its assumptions, but would perform worse than other methods when these assumptions are violated. Based on this model, we predicted that the homogeneous methods (p-curve, p-uniform) would do well in simulations with homogeneous effect sizes, but that the other two methods will outperform p-curve and p-uniform in simulations with heterogeneous effect sizes and in predicting the outcome of the OSC replication project. We will introduce more specific predictions of the two methods that do not assume homogeneous effect sizes, after we explain the two methods.

Statistical Power

The power of a statistical test (Neyman and Pearson, 1933; Lehman, 1959; Cohen, 1988) is the probability of correctly rejecting the null hypothesis. Power can be calculated exactly for any chosen set of parameter values, without using sample data in any way. This may be done before data are collected in order to choose sample size (Cohen 1988, Desu and Raghavarao 1990), or to estimate the power of published studies to produce significant results assuming different effect sizes (Cohen 1962, Sedlmeier and Gigerenzer 1989). The problem with this use of power analysis is that it provides power for hypothetical effect sizes, while we are interested in estimating average power of studies for the actual population effect sizes of studies, when the actual effect sizes are unknown.

The formal definition of power is the probability of obtaining a significant result when the null hypothesis is false. This may seem to present a problem for the use of power analysis with actual effect sizes because some studies may have an effect size of zero; that is, when the null hypothesis is true. Although power is not strictly defined in this case, the probability of obtaining a significant result is defined by the significance criterion, usually 0.05. Thus, power can never reach a value of zero and is bound at the level of the significance criterion. It is practically irrelevant whether we assume that effect sizes are never strictly zero and treat .05 as the lower limit of the range of power values or whether we use an extended definition of power to include the probability of obtaining a significant result even when the null hypothesis is true and equate power with alpha (Sterling et al., 1995). In short, our goal is to estimate average power of a set of studies with unknown population effect sizes that can assume any value, including zero. Moreover, we aim to do so for studies that were selected for significance either because journals only published significant results or because the goal of a replication study is to confirm that an original significant result is replicable.

Observed power The difficulty of estimating true power based on a single study is well documented (Boos and Stefanski 2012, Gerard, Smith and Weerakkody 1998, Gillett 1994, Hoenig and Heisey 2001, Thomas 1997, Yuan and Maxwell 2005). One problem is that the observed power method relies on the observed effect size as an estimate of the population effect size to compute power, and if studies are selected for significance, observed effect sizes are inflated. Even without selection bias, observed power estimates based on a single study are too variable to be practically useful (Yuan & Maxwell, 2005). However, low precision in estimates based on a single study does not mean that estimates from individual studies are useless. Each study provides some reliable information about power and when results are aggregated across studies, average observed power can be used to reducing sampling error and to obtain meaningful estimates of true power (Francis, 2012; Schimmack, 2012; Simonsohn et al., 2014b).

Heterogeneity Since power is a function of effect size and sample size, estimates of effect size lead immediately to estimates of power. Furthermore, some methods of estimating effect size explicitly take publication bias into account. In our view, the most promising of these are the p-curve method of Simonsohn, Nelson and Simmons (2014b) and the p-uniform method of van Assen, van Aert, and Wicherts (2014). Once an estimate of the population effect size has been found, it is straightforward to use this quantity to compute an estimated power for each study. Averaging the power estimates produces an estimate of population mean power.

Estimates of effect size generally assume that a single quantity is being estimated. In contrast, our interest is in a setting where not only the sample sizes, but the effect sizes, the topics being investigated and the statistical tests employed are all subject to sampling variation. That is, we wish to estimate population mean power not just assuming selection for significance, but also under *heterogeneity* --- that is, assuming that each test in the population has its own true power (a fixed, unknown number), and they all might be different.

It has been suggested (Ioannidis and Trikalinos 2007, Kepes, Banks, McDaniel and Whetzel 2012, Simonsohn et al 2014b, van Assen et al 2014) that methods developed for homogeneity may be applied to the heterogeneous situation by subsetting the data into tests with approximately the same effect size, or even the same true power. It has never been examined whether this approach can solve the problem of heterogeneous effect sizes. Schimmack (2012) proposed to conduct a meta-analysis of power estimates. This approach does not require homogeneous effect sizes, but it leads to inflated power estimates when studies are selected for significance. We propose two novel methods for estimating mean power in the challenging scenario where effect sizes and sample sizes are heterogeneous and publication bias is present. One method uses maximum likelihood estimation (which we view as the default method of estimation in Statistics). The other method fits predicted standardized test values (z-scores) to observed z-scores and uses the estimated weights and non-centrality parameters to compute average power. We call this method z-curve. We test these methods in a wide range of simulation studies. Furthermore, we compare our methods to p-curve and p-uniform to examine the robustness of these methods for homogeneous data when heterogeneity is present. As previous simulations have focused on effect size estimation, our simulations provide the first test of these methods for the estimation of power and replicability.

Notation and statistical background

To present our methods formally, it is necessary to introduce some statistical notation. Rather than using traditional notation from Statistics that might make it difficult for non-statisticians to understand our method, we use computer syntax as notation (Simonsohn, Nelson, & Simmons, 2014a). We use the S syntax (Becker, Chambers and Wilks, 1988) that is familiar to psychologists who conduct data analysis using the R statistical software (R core team, 2012). It also makes it easier to program our methods in R.

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 α (usually $\alpha = 0.05$), when the continuous test statistic T exceeds a critical value c. Typically there is a sample of test statistic values $T_1, \ldots, T_k$, 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 $\mathtt{p}(t)$ is the area under the curve, $\mathtt{d}(t)$ 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 $\mathtt{p}(\mathtt{q}(t)) = \mathtt{q}(\mathtt{p}(t)) = t$.

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 and Balakrishnan (1995) for a detailed and authoritative treatment. When the null hypothesis is false, the area under the curve of the test statistic's sampling distribution is $\mathtt{p}(t,\mathtt{ncp})$, representing particular cases like $\mathtt{pf}(t,\mathtt{df1},\mathtt{df2},\mathtt{ncp})$. 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 = $1-\mathtt{p}(c,\mathtt{ncp})$.

For the most important non-central distributions (Z, t, chi-squared and F), the non-centrality parameter can be factored into the product of two terms. The first term is an increasing function of sample size, and the second term is a function of the unknown parameters that reflects how wrong the null hypothesis is. In symbols,

$$\mathtt{ncp} = f_1(n) \cdot f_2(\mathtt{es}). \qquad (1)$$

In this equation, *n* is the sample size and $\texttt{es}$ is *effect size*. 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 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 ``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 for example 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

$$\texttt{ncp} = n\rho(1-\rho)d^2,$$

where $n = n_1 + n_2$ is the total sample size, $\rho = \frac{n_1}{n}$ is the proportion of cases allocated to the first treatment, and $d = \frac{|\mu_1 - \mu_2|}{\sigma}$ 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, $f_1(n) = n\rho(1-\rho)$ and $f_2(\texttt{es}) = \texttt{es}^2$. This is just an example; 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

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

The function $f_2(\mathtt{es})$ 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. The tests are conducted. Significant results are published and become available as data. Non-significant results go into the mythical ``file drawer" of Rosenthal (1979). This means that there are two populations of power values: the original population, and the sub-population corresponding to the tests that happened to be statistically significant.

Selection for significance (publication bias) 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. The reason is that selection for significance tends to select studies with higher power. For example, a study with 80% power is more likely to end up in the sample of studies selected for significance than a study with 20% power.

Probability models may often be clarified by thinking of them 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 (the file drawer), or more likely to landfill.

Spinning all the wheels on display a second time would take a great deal of effort, but if we did so we could record the proportion of wins. This would not be the true probability of significance, but if the number of wheels on display is large it would be close. Spinning all the wheels a third time would yield another proportion of wins, presumably close to the first. Repeating this impossibly tedious exercise a large number of 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 important unknown quantity using only the numbers that appeared on first spin, for the wheels on display.

We now give a set of fundamental principles connecting the probability distribution power before selection to its distribution after selection. 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.*

For example, suppose that before selection, 80% of studies have power equal to 0.10 and 20% have power equal to 0.60. Table 1 shows the distribution of power before and after selection. Expected (population mean) power before selection is 0.10*0.8 + 0.60*0.2 = 0.20. After selection there are still the same two power values, but their probabilities change. To obtain the probability that power equals 0.10 after selection, multiply 0.8 by the power value 0.1, and divide by the expected power before selection of 0.20. The resulting probability after selection is 0.8*0.1/0.2 = 0.40.

Insert Table 1 about here

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.*

P-curve and p-uniform estimation of mean power

The p-curve (Simonsohn et al 2014b) and p-uniform (van Assen et al 2014) methods are designed for estimating effect sizes in meta-analyses where there is a single fixed effect size, but possibly varying sample sizes. We adapted them slightly to produce estimates of mean power. Both p-uniform and p-curve are based on the idea that p-values are uniformly distributed when the null hypothesis is true. Originally, the test statistics were used to test the null hypothesis that the population effect size is zero. To use this method for the estimation of population effect sizes, the method is applied repeatedly with *modified* hypotheses that the effect size equals some specified non-zero value. If a hypothesized effect size matches the actual population effect size, the resulting p-values would again have a uniform distribution. To find the best fitting effect size for a set of observed test statistics, p-curve and p-uniform compute p-values for various effect sizes and chose the effect size that yields the best approximation of a uniform distribution.

If the modified null hypothesis that effect size = $\mathtt{es}$ is true, the cumulative distribution function of the test statistic is the conditional probability

$$\begin{aligned} F_0(t) &= Pr\{T \le t \,|\, T > c\} \\ &= \frac{\mathtt{p}(t,\mathtt{ncp}) - \mathtt{p}(c,\mathtt{ncp})}{1 - \mathtt{p}(c,\mathtt{ncp})} \\ &= \frac{\mathtt{p}(t, f_1(n)\cdot f_2(\mathtt{es})) - \mathtt{p}(c, f_1(n)\cdot f_2(\mathtt{es}))}{1 - \mathtt{p}(c, f_1(n_i)\cdot f_2(\mathtt{es}))}, \end{aligned}$$

using $\mathtt{ncp} = f_1(n)\cdot f_2(\mathtt{es})$ as given in Equation (1). The corresponding modified p-value (which Simonsohn et al. call the *pp*-value) is

$$1 - F_0(T) = \frac{1 - \mathtt{p}(T, f_1(n)\cdot f_2(\mathtt{es}))}{1 - \mathtt{p}(c, f_1(n)\cdot f_2(\mathtt{es}))}.$$

Note that since the sample sizes of the tests may differ, the symbols $\mathtt{p}$, n and c as well as T may have different referents for $j = 1,\ldots,k$ test statistics. The subscript j has been omitted to reduce notational clutter.

The only difference between p-curve and p-uniform is the criterion that is used to find the effect size that yields a distribution of modified p-values that best approximates a uniform distribution. P-curve is based on a Kolmogorov-Smirnov test for departure from a uniform distribution, choosing the $\mathtt{es}$ value yielding the smallest value of the test statistic. P-uniform is based on a different criterion. Denoting by P_j the modified p-value associated with test j, calculate $Y = -\sum_{j=1}^{k} \ln(P_j)$, where $\ln$ is the natural logarithm. If the P_j values were uniformly distributed, Y would have a Gamma distribution with expected value k, the number of tests. The P-uniform estimate is the modified null hypothesis effect size $\mathtt{es}$ that makes Y equal to k, its expected value under uniformity.

These technologies are designed for heterogeneity in sample size only, and assume a common effect size for all the tests. Given an estimate $\widehat{\mathtt{es}}$ of the common effect size, estimated power for each test is solely determined by sample size. Using Expression (2), the estimated power of test j is $1-\mathrm{p}(c_j, f_1(n_j)\cdot f_2(\widehat{\mathtt{es}}))$. Population mean power can then be estimated by averaging the k power estimates. This natural way of estimating mean power is merely implicit in the papers by van Assen et al (2014) and Simonsohn et al (2014b).

Maximum likelihood estimation of mean power

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 and 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 $t_1,\ldots,t_k$, the likelihood function is a product of k terms of the form

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

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.

As an example, consider a one-way ANOVA with four treatment groups, equal sample sizes, and a ``medium" value of 0.25 for Cohen's (1988, p. 275) effect size $\mathbf{f}$. As shown in the technical supplement, $\mathtt{ncp} = f_1(n)\cdot f_2(\mathtt{es}) = n\cdot \mathtt{es}^2$ for this problem, where n is the total sample size. Figure 1 shows the likelihood function for a simulated set of $k = 25$ F statistics. In this example, the sample sizes before selection varied about a mean of twenty per treatment. The likelihood function reaches its maximum when effect size equals 0.244; this is the maximum likelihood estimate. It is quite close to the true value of 0.25.

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 as for p-curve and p-uniform, 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 $g_\theta(\mathrm{es})$. This notation indicates that the distribution of effect size depends on an unknown parameter or parameter vector θ. In the technical supplement, it is shown that the likelihood function (a function of θ) is a product of k terms of the form

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

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 θ for which the value of the product is highest. Denote the maximum likelihood estimate by $\hat{\theta}$. Typically $\hat{\theta}$ 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

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

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

Z-curve

Z-curve follows a traditional meta-analyses 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 fit a single function to p-values arising 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 m (Heisey & Hoenig, 2001). As described 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 $\alpha = 0.05$. 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 $\alpha / 2 = 0.025$. 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 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 m 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 m from $m_1,\ldots,m_r$ according to the respective probabilities $w_1,\ldots,w_r$. Then generate Z from a normal distribution with mean m 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

$$f(z)=\sum_{j=1}^{r} w_j \frac{\texttt{dnorm}(z-m_j)}{\texttt{pnorm}(6-m_j)-\texttt{pnorm}(c-m_j)}, \tag{3}$$

for c < z < 6.

[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.

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 m_j values so as to minimize the sum of absolute differences between the conditional density estimate and (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 four estimation methods (p-curve, p-uniform, maximum likelihood and z-curve) 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.

Heterogeneity in Sample Size Only: Effect size fixed

Sample sizes after selection for significance were randomly generated from a Poisson distribution with mean 86, so that they were approximately normal, with population mean 86 and population standard deviation 9.3 (Johnson, Kemp and Kotz, 2005). Population mean power, number of test statistics on which the estimates were based, type of test (chi-squared or F) and (numerator) degrees of freedom were varied in a complete factorial design. Within each combination, we generated 10,000 samples of significant test statistics and applied the four estimation methods to each sample. In these simulations, it was not necessary to simulate test statistic values and then literally select those that were significant. A great deal of computation was saved by simulating directly from the distribution of the test statistic after selection; details are given in the technical supplement.

Effect sizes were selected to yield population mean power values after selection of 0.05, 0.25, 0.50 or 0.75. For F -tests, we used Cohen's (1988, p.275) effect size metric $\mathbf{f}$. For chi-squared tests, we used $\mathbf{w}$ (Cohen, 1988, p. 216). The number of test statistics k on which estimates were based was 15, 25, 50, 100 or 250. Numerator degrees of freedom (just degrees of freedom for the chi-squared tests) were one, three or five. Because the pattern of results was similar for F and chi-squared tests and for different degrees of freedom, we give details for F -tests with one numerator degree of freedom; preliminary data mining of the psychological literature suggests that this is the case most frequently encountered in practice. Full results are given in the supplemental materials.

Table 2 shows means and standard deviations of estimated population mean power after selection. Differences between the mean estimates and the true values represent bias in estimation. We conclude that all methods performed fairly well, with z-curve showing a bit more bias than the other methods.

Insert Table 2 about here

Absolute error of estimation It is desirable for average estimates to be close to the true values, but still positive and negative errors may cancel. More interesting is how close the estimate is on average to the true value being estimated. Table 3shows mean absolute error of estimation for F-tests with one numerator degree of freedom (full results are given in the supplemental material). As expected, all the methods become more accurate with larger numbers of tests. Though the differences are fairly small, z-curve is least accurate when mean power is low, and most accurate when mean power is high. Maximum likelihood has a slight edge over the other methods under most circumstances, except that z-curve sometimes does better when population mean power is moderate to high and the estimates are based on a small number of tests.

Insert Table 3 about here

Testing differences in accuracy Because results like the ones in in Table 2 are based on random number generation, some of the apparent differences could be due to chance. Thus we find ourselves applying statistical tests to an investigation of statistical tests. Within each of the 20 combinations of power and number of tests, there are six potential pairwise comparisons of mean absolute error. These comparisons were carried out using large-sample two-sided matched Z-tests with a Bonferroni correction, yielding a joint 0.001 significance level for the 120 tests.

Table 4 shows the number of times that the row method was significantly more accurate than the column method by this stringent criterion. There are 6 sub-tables, one for each combination of type of test (chi-squared or *F*) and degrees of freedom. For *F* -tests, *df* refers to the numerator (experimental) degrees of freedom. Note that the Bonferroni correction was applied separately to each sub-table. In all, Table 4 summarizes the results of 720 tests (full details are given in the supplemental material).

Insert Table 4 about here

In each sub-table of Table 4, the most accurate method overall is maximum likelihood, followed by p-uniform. When maximum likelihood lost a comparison it was usually to z-curve. As one would expect from the general theory of maximum likelihood estimation (Lehmann and Casella 1998, Ch 6), maximum likelihood performed particularly well when estimates were based on a large number of tests. It is important to recognize, however, the differences in average estimation error are fairly small. We conclude that although maximum likelihood performs best, all the methods yield reasonable estimates when effect sizes are homogeneous.

In conclusion, our results confirmed that p-curve and p-uniform perform well in simulations with homogeneous effect sizes. Nevertheless, the new maximum likelihood method outperformed the existing methods. This is not surprising because maximum likelihood is hard to beat when data conform to the model underlying maximum likelihood estimation.

Heterogeneity in Both Sample Size and Effect Size

To model heterogeneity in effect size, we let effect sizes before selection vary according to a gamma distribution (Johnson, Kotz and Balakrishnan, 1995), a flexible continuous distribution taking positive values. Sample size before selection remained Poisson distributed with a population mean of 86. For convenience, sample size and effect size were independent before selection. Maximum likelihood also assumed that effect size is gamma distributed, and the likelihood search was over the two parameters of the gamma distribution. The other 3 methods were not modified in any way. P-curve and p-uniform continued to assume a fixed effect size, and z-curve continued to assume heterogeneity in the non-centrality parameter without distinguishing between heterogeneity in sample size and heterogeneity in effect size.

The simulation study followed the format of the previous study with amount of heterogeneity in effect sizes as an additional factor. The factors were true population mean power (0.25, 0.50 or 0.75), standard deviation of effect size after selection (0.10, 0.20 or 0.30), 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. For brevity, we present results for F -tests with numerator $df = 1$. Full results are given in the supplemental material.

Table 5 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. P-uniform broke down completely for higher heterogeneity in effect size, with most estimates close to one regardless of the true value. For moderate to high mean power, the p-curve also produces an over-estimate on average, with the problem becoming most severe when mean power and heterogeneity in effect size are both high. Maximum likelihood and z-curve performed much better.

Insert Table 5 about here

Table 6 shows the mean absolute error of estimation. The results are consistent with those in Table 5 with maximum likelihood and z-curve outperforming p-curve and p-uniform.

Insert Table 6 about here

Table 7 directly pits the methods against each other and counts the wins for all three *df* values, and for chi-squared tests as well as *F*-tests. The clear winner is maximum likelihood, followed by z-curve, p-curve and p-uniform in that order. When other methods beat maximum likelihood, it was almost always when heterogeneity in effect size and true population power were both low. This is consistent with Table 5, in which maximum likelihood performs better when mean power is moderate to high.

Insert Table 7 about here

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. Maximum likelihood had a substantial and arguably unfair advantage, since it assumed exactly the correct distribution for effect size. Also, sample size and effect size before selection were independent in both the simulations and in the assumptions of maximum likelihood. It is well known that when its assumptions are correct, maximum likelihood is very accurate compared to other methods (Lehmann and Casella 1998, Ch. 6). When assumptions are incorrect however, there are no general theoretical results and the performance of maximum likelihood must be assessed through simulation studies.

To test the robustness of maximum likelihood to assumptions, we conducted a simulation study with a complex distribution of effect sizes. We limited this simulation to

F-tests with one numerator degrees of freedom because the previous simulations showed that that the test-statistic and degrees of freedom had no effect on the results.

Effect size after selection had a beta distribution rather than a gamma before selection. 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, 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 are using smaller samples when they expect larger effects. This is evident in the OSR (2015), 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 $\exp(\beta_0 + \beta_1 \mathrm{es})$. Mean effect size after selection and the parameters β_0 and β_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 8 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 no effect on the results. Although unexpected, this result suggests that it is permissible to assume independence between effect sizes and sample sizes, which makes the methods more useful for applications to real datasets where this assumption is violated. The results also replicated the finding that the two methods for heterogeneous data outperformed p-curve and p-uniform. The most important result was that maximum likelihood no longer was a clear winner. In fact, z-curve generally produced slightly more accurate estimates than maximum likelihood.

Insert Table 8 about here

Table 9 shows mean absolute error of estimation; the results are consistent with the results in Table 8.

Insert Table 9 about here

Table 10 shows the head-to-head comparison of the four methods. The results provide the strongest evidence that z-curve outperforms maximum likelihood when the distribution of effect sizes does not match a gamma distribution, which is the assumed distribution of the maximum likelihood method. Z-curve has an advantage over maximum likelihood because it does not assume a specific distribution of effect size. This shows the advantage of methods with few assumptions to outperform methods with assumptions when these assumptions are violated.

Insert Table 10 about here

Simulation of Full Heterogeneity

When population mean power in a field of study is being estimated, there will typically be heterogeneity not just in sample size and effect size, but also in the tests on which estimates are based. The distribution of sample size is unlikely to be Poisson, the distribution of effect size will not be gamma and the null hypothesis will be true with non-zero probability. Our full heterogeneity simulation examines the performance of the four methods in this situation. We also changed the simulation of sample sizes. While the Poisson distribution is a widely accepted model for count data (Johnson, Kemp and Koch, 2005), sample size may be more dispersed and skewed than a Poisson distribution. Figure 1 compares the Poisson distribution with mean 86 to a histogram of 7,000 approximate sample sizes based on denominator degrees of freedom in the journal *Psychological Science* since the start of the journal to the end of 2015.

Insert Figure 1 about here

Sample sizes were estimated from degrees of freedom in reported statistical tests, using the formula $n = df_1 + df_2 + 1$. Numerator degrees of freedom were limited to ten or fewer, and the data were edited so that sample size ranged from 20 to 500, with a mean of 86. In this simulation, eighty percent of the tests were F-tests, and twenty percent were chi-squared. For the F-tests, (df_1, df_2) pairs were randomly sampled with replacement from the *Psychological Science* data. The degrees of freedom for the chi-squared tests were randomly sampled with replacement from the df_1 values. Sample size was selected with replacement, independently of degrees of freedom.

Effect size In this set of simulations, effect size has a mixed continuous-discrete distribution. With probability 0.10, effect size equals zero, so that the null hypothesis is exactly true. With probability 0.05, effect size has a standard exponential distribution shifted by one; in this case the minimum effect size is over twice Cohen's (1988) ``high" value, representing manipulation checks and other ``findings" that are too good to be true. The other 0.85 probability is devoted to a beta distribution, with parameters chosen to make population mean power after selection either 0.25, 0.50 or 0.75. No special attempt was made to hold the standard deviation of effect size constant, but all values were above 0.30. Sample size and effect size are independent after selection, so that before selection they are non-independent.

Figure 2 shows the distribution of effect size after selection and the resulting distribution of power after selection. It is evident that the effect of heterogeneity in sample size and effect size is increased heterogeneity in power. Since power is bounded by 0.05 and one, its distribution is forced to the extremes.

Insert Figure 2 about here

Table 11 shows means and standard deviations of the estimates for the four methods. The results mostly replicate previous findings, but this time maximum likelihood performed better than in the previous simulation and often outperformed z-curve.

Insert Table 11 about here

Table 12 shows only small differences between Maximum Likelihood and z-curve and the head-to-head comparison in Table 13 ends in a tie.

Insert Table 12 about here

Insert Table 13 about here

It is not entirely clear why the simulation of full heterogeneity led to better performance of maximum likelihood than the previous simulation with a gamma distribution. One possibility is that combining a beta distribution and a shifted exponential distribution produced a right-skewed distribution that is fairly well approximated by a gamma. An advantage of z-curve is that it does not make assumptions about the distribution of effect sizes. This makes the method more robust when the distribution deviates from a beta distribution that is the assumed distribution underlying the maximum likelihood method. Nevertheless, we recommend both methods and both methods are likely to produce similar estimates when they are applied to real data. Z-curve may be preferable because it runs much faster, at least in or implementation.

A conservative bootstrap confidence interval for z-curve

Estimates should always be accompanied by confidence intervals, to give an idea of their precision. For z-curve, the most natural choice is a bootstrap confidence interval.

The bootstrap (Efron 1981, Efron and Tibshirani 1993) is based on re-sampling from the observed data with replacement, calculating a statistic on each re-sampled data set, and using the histogram of the resulting values as an approximation to the sampling distribution of the statistic. In this case the statistic is the z-curve estimate. Our choice is the percentile confidence interval method, which assumes that the sampling distribution of the estimate is symmetric, and centered on the quantity being estimated. Here, we re-sampled test statistics and computed z-curve estimates *B* = 500 times. The 95 percent bootstrap confidence interval ranges from the 2.5 percentile to the 97.5 percentile of the estimates.

Especially when samples are small, it is important to verify that a proposed 95% confidence interval contains the true value 95% of the time. This is called the *coverage* of the confidence interval. In a pilot study, we found that the coverage of the 95% bootstrap confidence interval was sometimes less than 95%. For example, notice in Table 11 that the mean estimate for power = 0.25 and *k* =2,000 is 0.23 rather than 0.25. The sampling distribution of the z-curve estimate is nicely symmetric as required by the bootstrap method, but it is centered on 0.23 and not 0.25. The resulting coverage of the confidence interval is roughly 84% when it should be 95. With increasing volume of data, the width of the confidence interval would shrink and the coverage would decrease to zero.

Reviewing the average z-curve estimates from all the simulations, we determined that the bias of the z-curve estimate is seldom more than two percentage points, and never more than two percentage points for larger samples. Thus an easy fix of the confidence interval is to decrease the lower limit by 0.02 and increase 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 (again, the distributions in Figure 2), and seven values of the number of test statistics, ranging from *k* =25 to *k* =2,000. Table 14 gives the coverage values. Even for *k* =25 its performance is respectable. The table shows that the conservative bootstrap confidence interval is indeed conservative under most circumstances. When the estimates are based on larger numbers of test statistics, it behaves more like a 99 percent confidence interval. For estimates based on fewer than 25 test statistics, it might be helpful to increase the correction factor from 0.02 to 0.025.

Insert Table 14 about here

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 0.75 and the z-curve estimate is based on *k* =100 test statistics, the average confidence interval will range from 0.65 to 0.85. This may be sufficient precision for some purposes, but it is desirable to base estimates on a larger number of test statistics if possible.

Insert Table 15 about here

Application to the Replication Project data

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 the methods discussed in this paper assume that selected results are a representative sample. That is, there is no additional selection for larger effect sizes. This assumption is unlikely to be true for marginally significant results. Of the remaining 93 studies, five were eliminated because the replication studies were not based on an independent sample or had other unusual characteristics. Thirty-four of the 88 replication studies produced a significant result (39%), which is similar to the replication rate for all 97 significant original studies (37%).

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 t obtain chi-squared with one degree of freedom. Input to z-curve was simply the set of p-values. For the other three methods, 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 estimates of population mean power were 0.68 for p-curve, 0.76 for p-uniform, 0.59 for maximum likelihood and 0.66 for z-curve. The 95% confidence interval for z-curve was from 0.49 to 0.79. Given the small sample size, it is difficult to make strong claims about the relative performance of the four methods, yet the ordering of estimates is consistent with the simulation studies with heterogeneous effect sizes. Accordingly, p-uniform produces the strongest overestimation followed by p-curve, then z-curve and then maximum likelihood. The confidence interval around the z-curve estimate allows for a wide range of values, but even the lower bound is above the actual success rate of 39%. This suggests that all methods overestimate the success rate in actual replication studies. We discuss possible explanations for this discrepancy between statistical predictions and actual success rates in the discussion section.

Discussion

In this paper, we have compared four methods for estimating population mean power after selection for significance: p-curve, p-uniform, maximum likelihood and z-curve. P-curve (Simonsohn et al., 2014b) and p-uniform (van Assen et al., 2014) are slight adaptations of methods for estimating a fixed effect size. Maximum likelihood is a generic approach to estimation for any parametric model, and z-curve is new. Based on a set of large-scale simulation studies, we conclude that z-curve is the most accurate method when there is substantial heterogeneity in effect size and the distribution of effect size is unknown. It is also the most convenient, requiring only a set of *p*-values as input. Estimates should be accompanied by confidence intervals. We describe a conservative bootstrap confidence interval for z-curve and verify by simulation that it has good coverage even for small samples.

In a meta-analysis of studies testing exactly the same hypothesis with very similar subject populations, it is reasonable to assume that effect size is a single fixed constant, while sample size of course may vary. This is the setting for which p-curve and p-uniform were designed. Here, all the methods performed reasonably well in our simulations, but the most accurate method was maximum likelihood. Thus, we recommend maximum likelihood for conducting fixed effects meta-analysis when the set of studies shows evidence of publication bias.

Some important statistical features of z-curve require further investigation. One is the question of independence. In all he simulations, the input *p*-values were independent. While z-curve does not formally assume independent inputs, the bootstrap confidence interval definitely does. Further simulations need to examine how z-curve performs when clusters of *p*-values come from multiple dependent tests of the same data. tests conducted on the same raw data set. Another unresolved issue is how well the method performs for tests that do not have one of the common non-central distributions under the alternative hypothesis (e.g., repeated measures ANOVA).

Another important extension of z-curve and maximum likelihood is to examine the ability of these methods to estimate the average power before selection of significance. These estimates would make it possible to estimate the size of the file-drawer with unpublished non-significant results. This estimate will always be lower than the estimate for the population of studies that were selected for significance because the non-significant studies in the file-drawer have lower power. At present, the mean power of studies selected for significance provides a conservative and optimistic estimate of the average power of studies conducted. The results for the OSG dataset yielded estimates of 60% to 70% power. These estimates are surprisingly high and higher than Cohen's (1962) estimate of 50% power (see also Gigerenzer & Sedlmeier, 1989). There are numerous reasons for this discrepancy and future research with larger samples is needed to obtain more precise estimates of the typical power of psychological studies.

Last but not least, future research needs to examine which factors contribute to the discrepancy between our estimate of replicability of the OSC studies and the actual success rate of replication studies. Before we discuss possible reasons, it is important to recognize that the sample size is small and that sampling error may at least partially explain the discrepancy. However, a number of additional factors are likely to play a role as well.

It is possible that actual replication studies underestimate replicability of original studies, and this may occur for a number of reasons. First, it is possible that the replication studies were not exact replications, and that minor differences between the original and replication studies accounted for a lower success rate. Although this is a possibility, there is no direct evidence to support this claim. Moreover, our results clearly show that the 97% success rate in the original study can only be explained with selection for significance. The 95% confidence interval of our estimate of replicability ranges from 49% to 79%. As power before selection for significance is lower than power after selection for significance, the most optimistic estimate of the true success rate before selection for significance would be 79%; and this is a very optimistic estimate, indeed. Thus, our results provide strong evidence for the presence of publication bias in psychology journals (Sterling et al., 1995). Moreover, our results cannot be attributed to problems with actual replication studies because they are based on the very same significance tests that were used in original articles to claim a significant result.

Another important factor that needs to be considered is the relationship between replicability and mean power after selection for significance. Mean power after selection for significance is only an accurate estimate of replicability if we assume that researchers conduct one study and then select significant results for publication. However, if researchers conduct several similar conceptual replication studies, the chances for a study with low power to be selected into the population of studies with significant results increases. As a result, the power of studies in the population of significant results decreases and becomes more and more similar to the average power of studies without selection. To illustrate this, imagine that researchers would continue conducting exact replication studies until a significant result is found. In this case, even null effects would be selected into the pool of studies with significant results with 100% probability. As a result, there would be no difference between the set of studies before selection and the set of studies after selection. Of course, we are not implying that researchers conduct exact replication studies until they produce a significant result. However, they may engage in research practices that have essentially the same effect on replicability. For example, they may include several dependent variables and publish only the dependent variable that produced a significant result (John et al., 2012; Simmons et al., 2011). The use of these research practices would explain why our estimate of replicability, one that is based on a simple selection process, overestimates the success rate of actual replication studies in the OSC reproducibility project. Future simulation studies needs to examine how z-curve behaves when significant results are obtained with the help of questionable research practices.

At present it is impossible to quantify the replicability of psychological research precisely. However, it is encouraging that three different methods produce converging evidence that replicability in psychology is neither 100% nor 0%. Based on the OSC project we get an estimate of 40%. Cohen's approach yields estimates around 50% (Sedlmeier & Gigerenzer, 1989), and our new method suggested an estimate around 60%. Until further evidence provides more precise information we can expect that about 50% +/- 10% of published significant results will replicate. The advantage of our method is that it can be used for large sets of studies, making it possible to examine predictors of replicability. The OSC project already identified discipline, designs, and effect sizes as predictors of replicability, but the evidence was weak due to the small sample size. Our method can be applied to all published results and estimate replicability for the whole population of published results. Our method can also be used quickly, or even in real time, as new results become available. This makes it possible to track replicability and to evaluate whether various initiatives to improve replicability are actually paying off. Nearly 30 years ago, Gigerenzer and Sedlmeier observed that studies of power fail to increase the power of studies and there is no evidence that power of studies in psychology has increased over the past 50 years since Cohen published his seminal article. We hope that our method will be able to show an increase in power in response to the replicability crisis.

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.

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.

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/

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.

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: Illustration of Principle 2

	Probability	
Power value	Before selection	After selection
0.10	0.80	0.40
0.60	0.20	0.60

Table 2: Means and standard deviation of estimated population mean power for heterogeneity in sample size only: *F*-tests with numerator *df* = 1

	Mean					*Standard Deviation*				
Population Mean Power = 0.05										
	Number of Tests					Number of Tests				
	15	25	50	100	250	15	25	50	100	250
P-curve	0.083	0.073	0.064	0.059	0.055	0.059	0.039	0.024	0.015	0.007
P-uniform	0.076	0.067	0.061	0.058	0.054	0.050	0.032	0.019	0.012	0.006
MaxLike	0.076	0.067	0.061	0.057	0.054	0.050	0.033	0.020	0.012	0.006
Z-curve	0.086	0.071	0.058	0.049	0.040	0.088	0.065	0.044	0.031	0.019
Population Mean Power = 0.25										
	Number of Tests					Number of Tests				
	15	25	50	100	250	15	25	50	100	250
P-curve	0.269	0.261	0.256	0.253	0.251	0.156	0.128	0.095	0.069	0.046
P-uniform	0.256	0.253	0.252	0.251	0.251	0.147	0.121	0.089	0.065	0.042
MaxLike	0.260	0.255	0.253	0.251	0.251	0.146	0.120	0.087	0.064	0.042
Z-curve	0.314	0.305	0.293	0.280	0.268	0.155	0.127	0.093	0.068	0.045
Population Mean Power = 0.50										
	Number of Tests					Number of Tests				
	15	25	50	100	250	15	25	50	100	250
P-curve	0.484	0.491	0.496	0.497	0.499	0.175	0.139	0.102	0.073	0.046
P-uniform	0.473	0.485	0.493	0.496	0.499	0.170	0.132	0.097	0.070	0.044
MaxLike	0.479	0.489	0.495	0.497	0.499	0.166	0.130	0.095	0.068	0.043
Z-curve	0.513	0.516	0.513	0.508	0.502	0.151	0.121	0.091	0.068	0.045
Population Mean Power = 0.75										
	Number of Tests					Number of Tests				
	15	25	50	100	250	15	25	50	100	250
P-curve	0.728	0.736	0.742	0.747	0.749	0.128	0.098	0.069	0.048	0.030
P-uniform	0.721	0.732	0.740	0.746	0.748	0.126	0.097	0.067	0.047	0.029
MaxLike	0.728	0.736	0.742	0.747	0.749	0.121	0.093	0.065	0.045	0.028
Z-curve	0.704	0.712	0.717	0.723	0.728	0.105	0.084	0.064	0.048	0.033

Table 3: Mean absolute error of estimation for heterogeneity in sample size only: *F*-tests with numerator *df* = 1

	Number of Tests				
	15	25	50	100	250
Population Mean Power = 0.05					
P-curve	3.32	2.25	1.41	0.93	0.52
P-uniform	2.57	1.75	1.11	0.76	0.43
MaxLike	2.59	1.74	1.09	0.73	0.39
Z-curve	6.53	4.90	3.38	2.44	1.79
Population Mean Power = 0.25					
P-curve	12.94	10.49	7.69	5.53	3.64
P-uniform	12.11	9.87	7.17	5.18	3.38
MaxLike	12.07	9.76	7.05	5.10	3.32
Z-curve	13.55	11.09	8.21	5.96	3.87
Population Mean Power = 0.50					
P-curve	14.32	11.20	8.14	5.80	3.67
P-uniform	13.93	10.68	7.80	5.56	3.51
MaxLike	13.61	10.41	7.60	5.39	3.41
Z-curve	12.42	9.91	7.44	5.48	3.59
Population Mean Power = 0.75					
P-curve	9.77	7.59	5.38	3.72	2.35
P-uniform	9.79	7.59	5.34	3.71	2.32
MaxLike	9.33	7.23	5.11	3.53	2.21
Z-curve	8.34	6.96	5.56	4.30	3.13

Table 4: Number of times row method is significantly more accurate than column method: Heterogeneity in sample size only

		Chi-squared tests						***F*-tests**				
					df = 1							
		PC	PU	ML	ZC	Total		PC	PU	ML	ZC	Total
P-curve	(PC)	0	0	0	14	14	(PC)	0	0	0	13	13
P-uniform	(PU)	15	0	0	14	29	(PU)	15	0	0	13	28
MaxLike	(ML)	20	16	0	16	52	(ML)	20	17	0	14	51
Z-curve	(ZC)	5	4	3	0	12	(ZC)	7	5	4	0	16
					df = 3							
		PC	PU	ML	ZC	Total		PC	PU	ML	ZC	Total
P-curve	(PC)	0	0	0	15	15	(PC)	0	0	0	13	13
P-uniform	(PU)	16	0	0	15	31	(PU)	15	0	0	14	29
MaxLike	(ML)	20	15	0	16	51	(ML)	20	16	0	15	51
Z-curve	(ZC)	5	2	2	0	9	(ZC)	6	4	3	0	13
					df = 5							
		PC	PU	ML	ZC	Total		PC	PU	ML	ZC	Total
P-curve	(PC)	0	0	0	15	15	(PC)	0	0	0	13	13
P-uniform	(PU)	15	0	1	16	32	(PU)	14	0	0	14	28
MaxLike	(ML)	20	15	0	17	52	(ML)	20	16	0	15	51
Z-curve	(ZC)	3	2	2	0	7	(ZC)	6	4	3	0	13

Table 5: 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
P-curve	0.225	0.272	0.320	0.024	0.033	0.039
P-uniform	0.294	0.694	0.949	0.029	0.056	0.028
MaxLike	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
P-curve	0.549	0.679	0.757	0.024	0.027	0.026
P-uniform	0.602	0.913	0.995	0.024	0.019	0.003
MaxLike	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
P-curve	0.824	0.928	0.962	0.013	0.009	0.006
P-uniform	0.861	0.992	1.000	0.012	0.003	0.000
MaxLike	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 6: 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			
P-curve	2.87	3.16	7.08
P-uniform	4.50	44.38	69.90
MaxLike	3.55	2.06	3.34
Z-curve	2.59	3.08	2.90
Population Mean Power = 0.50			
P-curve	4.93	17.86	25.70
P-uniform	10.21	41.28	49.54
MaxLike	1.80	1.49	1.50
Z-curve	2.12	2.19	2.23
Population Mean Power = 0.75			
P-curve	7.45	17.75	21.23
P-uniform	11.08	24.17	24.99
MaxLike	1.42	1.18	1.16
Z-curve	1.69	1.42	1.55

Table 7: Number of times row method is significantly more accurate than column method: Heterogeneity in sample size and effect size

		Chi-squared tests						***F*-tests**				
					df = 1							
		PC	PU	ML	ZC	Total		PC	PU	ML	ZC	Total
P-curve	(PC)	0	45	0	0	45	(PC)	0	45	4	0	49
P-uniform	(PU)	0	0	0	0	0	(PU)	0	0	0	0	0
MaxLike	(ML)	41	45	0	33	119	(ML)	40	45	0	31	116
Z-curve	(ZC)	45	45	8	0	98	(ZC)	42	45	10	0	97
					df = 3							
		PC	PU	ML	ZC	Total		PC	PU	ML	ZC	Total
P-curve	(PC)	0	45	4	1	50	(PC)	0	45	5	4	54
P-uniform	(PU)	0	0	0	0	0	(PU)	0	0	5	0	5
MaxLike	(ML)	40	44	0	34	118	(ML)	40	40	0	34	114
Z-curve	(ZC)	40	45	7	0	92	(ZC)	39	45	7	0	91
					df = 5							
		PC	PU	ML	ZC	Total		PC	PU	ML	ZC	Total
P-curve	(PC)	0	45	5	4	54	(PC)	0	45	5	6	56
P-uniform	(PU)	0	0	0	0	0	(PU)	0	0	5	1	6
MaxLike	(ML)	40	45	0	36	121	(ML)	40	40	0	34	114
Z-curve	(ZC)	38	45	5	0	88	(ZC)	38	42	8	0	88

Table 8: 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
P-curve	0.407	0.405	0.403	0.403	0.402	0.043	0.044	0.043	0.044	0.044
P-uniform	0.853	0.852	0.852	0.852	0.852	0.003	0.004	0.003	0.004	0.004
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
P-curve	0.839	0.840	0.841	0.841	0.841	0.022	0.022	0.022	0.022	0.022
P-uniform	0.906	0.906	0.906	0.906	0.906	0.004	0.004	0.004	0.004	0.004
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
P-curve	0.990	0.991	0.992	0.992	0.992	0.002	0.002	0.002	0.002	0.002
P-uniform	0.964	0.966	0.966	0.967	0.967	0.003	0.003	0.003	0.003	0.003
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 9: 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					
P-curve	15.67	15.49	15.33	15.30	15.24
P-uniform	60.26	60.24	60.23	60.22	60.22
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					
P-curve	33.88	33.99	34.07	34.09	34.11
P-uniform	40.59	40.61	40.63	40.63	40.64
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					
P-curve	24.04	24.13	24.18	24.21	24.24
P-uniform	21.43	21.56	21.63	21.67	21.72
MaxLike	7.62	8.23	8.56	8.76	8.97
Z-curve	3.51	4.01	4.27	4.43	4.59

Table 10: Number of times row method is significantly more accurate than column method with beta effect size and correlated sample size and effect size: *F*-tests with numerator *df* = 1

	P-curve	P-uniform	MaxLike	Z-curve	Total
P-curve	0	50	0	0	50
P-uniform	25	0	0	0	25
MaxLike	75	75	0	5	155
Z-curve	75	75	69	0	219

Table 11: Means and standard deviations of estimated population mean power under full heterogeneity

	Mean					*Standard Deviation*				
Population Mean Power = 0.25										
	Number of Tests					Number of Tests				
	100	250	500	1000	2000	100	250	500	1000	2000
P-curve	0.280	0.280	0.283	0.288	0.292	0.072	0.051	0.037	0.027	0.020
P-uniform	0.691	0.776	0.823	0.856	0.877	0.155	0.107	0.077	0.054	0.039
MaxLike	0.267	0.267	0.268	0.269	0.269	0.046	0.029	0.020	0.015	0.012
Z-curve	0.251	0.240	0.234	0.232	0.230	0.064	0.042	0.032	0.025	0.020
Population Mean Power = 0.50										
	Number of Tests					Number of Tests				
	100	250	500	1000	2000	100	250	500	1000	2000
P-curve	0.561	0.571	0.577	0.581	0.585	0.063	0.040	0.029	0.020	0.015
P-uniform	0.807	0.861	0.891	0.911	0.923	0.090	0.060	0.042	0.030	0.022
MaxLike	0.473	0.468	0.465	0.463	0.462	0.054	0.035	0.025	0.019	0.015
Z-curve	0.517	0.505	0.497	0.491	0.487	0.071	0.047	0.035	0.026	0.020
Population Mean Power = 0.75										
	Number of Tests					Number of Tests				
	100	250	500	1000	2000	100	250	500	1000	2000
P-curve	0.828	0.836	0.840	0.842	0.844	0.034	0.020	0.014	0.010	0.007
P-uniform	0.921	0.945	0.956	0.964	0.968	0.035	0.022	0.015	0.011	0.008
MaxLike	0.740	0.736	0.734	0.731	0.730	0.045	0.030	0.022	0.016	0.012
Z-curve	0.764	0.756	0.750	0.745	0.740	0.042	0.030	0.023	0.018	0.014

Table 12: Mean Absolute Error of Estimation under Full Heterogeneity

	Number of Tests				
	100	250	500	1000	2000
Population Mean Power = 0.25					
P-curve	6.27	4.68	4.05	4.00	4.25
P-uniform	44.14	52.57	57.35	60.56	62.67
MaxLike	3.87	2.66	2.23	2.03	1.99
Z-curve	5.13	3.53	2.95	2.60	2.43
Population Mean Power = 0.50					
P-curve	7.39	7.21	7.67	8.10	8.50
P-uniform	30.67	36.14	39.13	41.06	42.30
MaxLike	4.81	3.84	3.67	3.74	3.79
Z-curve	5.93	3.78	2.81	2.23	1.98
Population Mean Power = 0.75					
P-curve	7.88	8.62	8.99	9.24	9.41
P-uniform	17.11	19.48	20.61	21.36	21.84
MaxLike	3.67	2.61	2.16	2.03	2.07
Z-curve	3.64	2.45	1.81	1.48	1.38

Table 13: Number of times row method is significantly more accurate than column method under full heterogeneity

	P-curve	P-uniform	MaxLike	Z-curve	Total
P-curve	0	15	0	0	15
P-uniform	0	0	0	0	0
MaxLike	15	15	0	6	36
Z-curve	15	15	7	0	37

Table 14: 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 15: 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

Figure 1: Likelihood Function for 25 F-tests With True Effect Size = 0.25

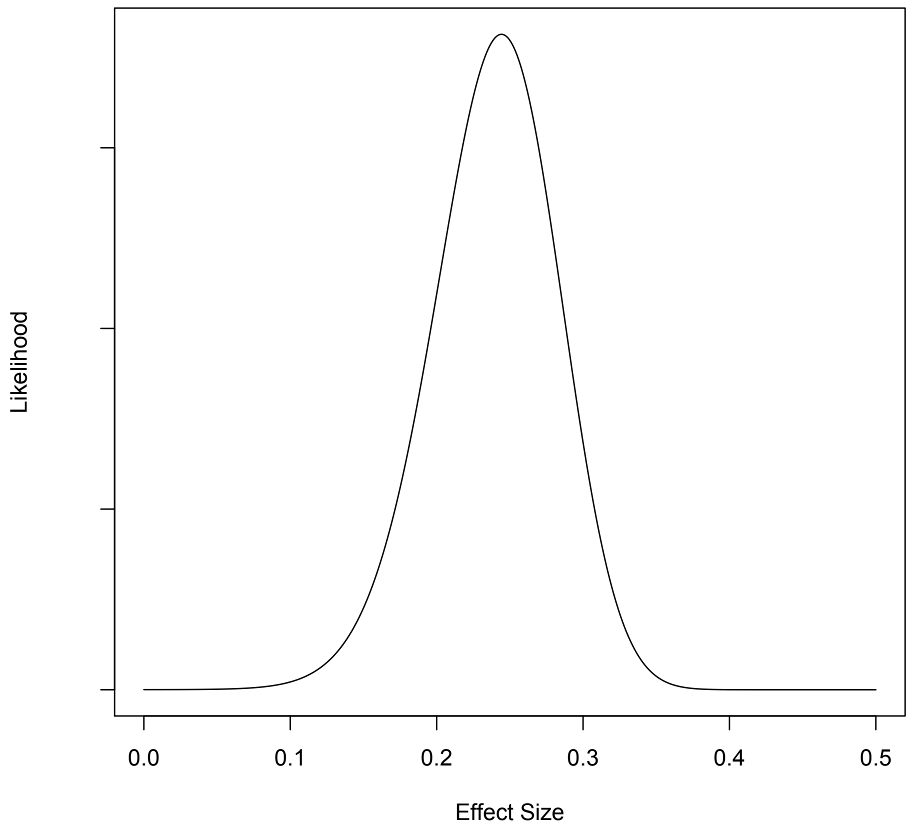


Figure 2: Poisson versus *Psychological Science* Sample Sizes

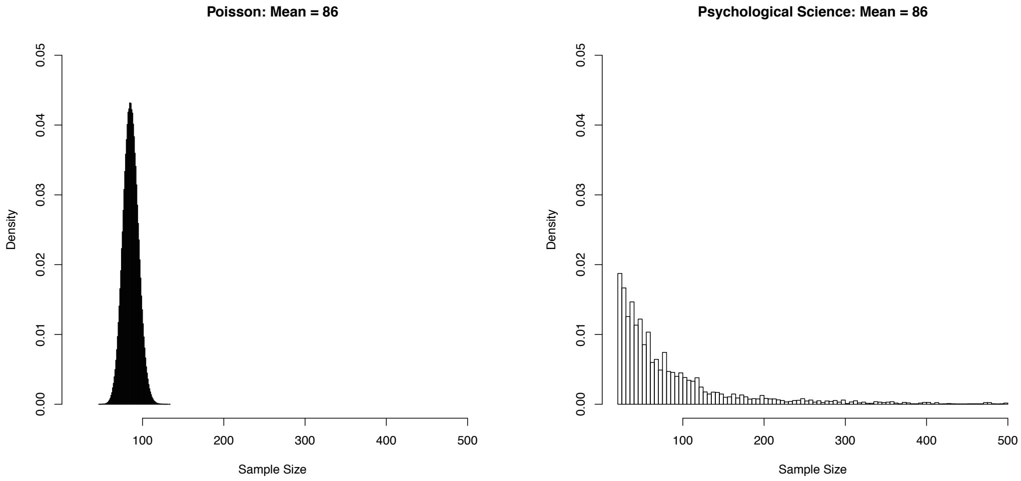
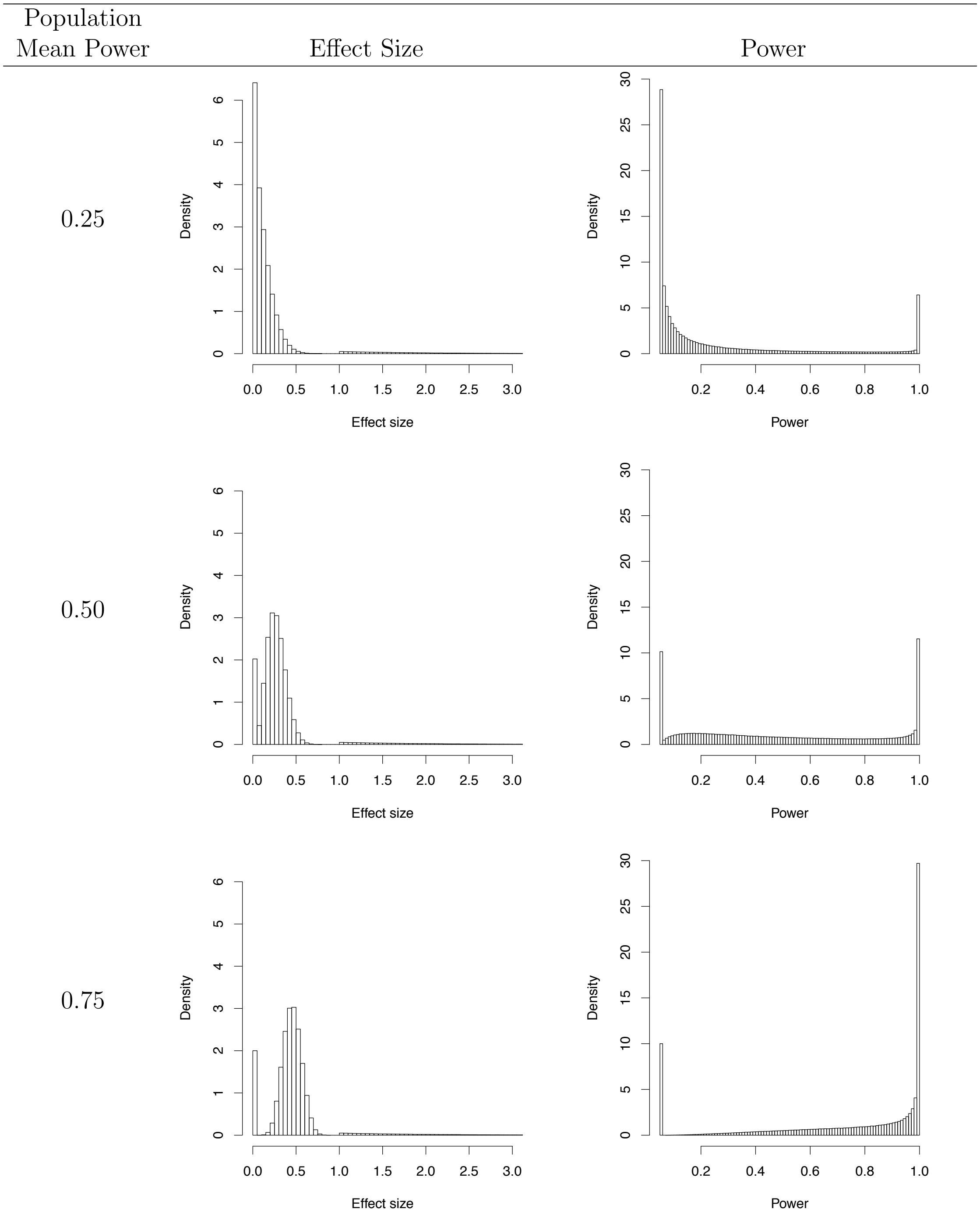


Figure 3: Distributions of effect size and power after selection under full heterogeneity


Population
Mean Power
Effect Size
Power
Density
Effect size
Power