A New Standard for the Analysis and Design of Replication Studies

Leonhard Held

Epidemiology, Biostatistics and Prevention Institute (EBPI)

Center for Reproducible Science

University of Zurich

Hirschengraben 84, 8001 Zurich, Switzerland

Email: leonhard.held@uzh.ch

27th November 2018

Abstract: A new standard is proposed for the evidential assessment of replication studies. The approach combines a specific reverse-Bayes technique with prior-predictive tail probabilities to define replication success. The method gives rise to a quantitative measure for replication success, called the sceptical p-value. The sceptical p-value integrates traditional significance of both the original and replication study with a comparison of the respective effect sizes. It incorporates the uncertainty of both the original and replication effect estimates and reduces to the ordinary p-value of the replication study if the uncertainty of the original effect estimate is ignored. The proposed framework can also be used to determine the power or the required sample size to achieve replication success. Numerical calculations highlight the difficulty to achieve replication success if the evidence from the original study is only suggestive. An application to data from the Open Science Collaboration project on the replicability of psychological science illustrates the proposed methodology.

Key Words: Confidence Interval; Power; Replication Studies; Replication Success; Sample Size; Sceptical p-value; Significance Test

1. Introduction

Replicability of research findings is crucial to the credibility of all empirical domains of science. As a consequence of the so-called replication crisis (Ioannidis, 2005; Begley and Ioannidis, 2015), the past years have witnessed increasing interest in large-scale replication projects, *e. g.* Open Science Collaboration (2015); Camerer et al. (2016, 2018). Such efforts help assess to what extent claims of new discoveries can be confirmed in independent replication studies whose procedures are as closely matched to the original studies as possible.

However, there is no established standard for the statistical evaluation of replication success. Standard significance of the replication study is often used but can easily lead to conclusions opposite to what the evidence warrants (Simonsohn, 2015). A comparison of the effect sizes of the original and replication study is also common, where a smaller replication effect estimate decreases the credibility of the original study result. A modification of this is to investigate whether the replication effect estimate is compatible with the original effect estimate (Bayarri and Mayoral, 2002; Patil et al., 2016). Meta-analytic combined effect estimates can also be computed, which, however, treat the original and replication study as exchangeable. This is often a rather unrealistic assumption in the presence of publication or reporting bias of the original study, leading to overly optimistic effect estimates and p-values.

Recently the lack of a single accepted definition of replicability has been emphasized by Goodman et al. (2016) who call for a better understanding of the relationship between reproducibility and the truth of scientific claims. Researchers have started to develop Bayesian methods to analyse and design replication studies (Verhagen and Wagenmakers, 2014; van Aert and van Assen, 2017; Schönbrodt and Wagenmakers, 2018), but there is a lack of appropriate methodology based on traditional metrics (effect estimates, confidence intervals and p-values). To address this deficiency, I propose a principled approach, combining the analysis of credibility (Matthews, 2001a,b) with the Box (1980) prior criticism approach to define *replication success* (Section 2). This gives rise to a new quantitative measure of replication success, the *sceptical p-value* (Section 3).

The sceptical p-value has attractive properties. It takes into account the results from both the original and the replication study and is always larger than the ordinary p-values from the original and the replication study. If the uncertainty of the original effect estimate is ignored, the sceptical p-value reduces to the ordinary p-value from the replication study. Moreover, the sceptical p-value considers replication stud-

ies with relatively small effect estimates (compared to the original estimates) as less successful. To avoid the so-called replication paradox (Ly et al., 2017), a one-sided sceptical p-value is derived within the proposed framework, ensuring that replication success can only occur if the original and replication effect estimates have the same sign.

Statistical power is of central importance in assessing the reliability of science (Button et al., 2013). Appropriate design of a replication study is key to tackling the replication crisis as many such studies are currently severely under-powered, even by traditional standards (Anderson and Maxwell, 2017). Methods to calculate the power for replication success are proposed in Section 4. Numerical calculations highlight the difficulty to achieve replication success if the evidence from the original study is only suggestive. The framework is also used to determine the required sample size to achieve replication success with appropriate power. Section 5 presents a reanalysis of data from the Open Science Collaboration (2015) project on the replicability of psychological science to illustrate the usefulness of the proposed methodology. I close with some comments in Section 6.

2. Assessment of Replication Success

Analysis of credibility (Matthews, 2001a,b) is a reverse-Bayes procedure originally designed to assess the credibility of significant findings in the light of existing evidence. The idea to use Bayes's theorem in reverse originates in the work of I.J. Good (Good, 1950, 1983) and is increasingly used to assess the plausibility of scientific findings (Greenland, 2006, 2011; Held, 2013; Colquhoun, 2017). A significant effect estimate from the original study is combined with a sceptical prior (Spiegelhalter et al., 1994, Section 4.1.3) centered around zero to represent doubts about large effect estimates. A sceptical prior shrinks the original effect estimate towards zero, where the amount of

shrinkage depends on the sceptical prior variance. Fletcher et al. (1993) have argued for the use of sceptical priors for original clinical study results, which often show a tendency for overoptimism.

In order to challenge the original study it is natural to ask how sceptical we would have to be not to find its apparently positive effect estimate convincing. This leads to a reverse-Bayes approach, where the 'posterior' is fixed to have a lower (or upper) credible limit exactly equal to zero and the sceptical prior variance is chosen accordingly. The approach thus represents the objection by a sceptic who argues that the original result would no longer be 'significant' if combined with a sufficiently sceptical prior. The goal is now to persuade the sceptic by showing that this prior is unrealistic. To do so, a replication study is conducted. If the data from the replication study are in conflict with the sufficiently sceptical prior, the original study result is confirmed.

Suppose the original study gives rise to a conventional confidence interval for the unknown effect size θ at level $1-\alpha$ with lower limit L and upper limit U. Assume that L and U are symmetric around the original point estimate $\hat{\theta}_o$ (assumed to be normally distributed) and that both are either positive or negative, *i. e.* the original effect is significant at significance level α. After a suitable transformation this framework covers a large number of commonly used effect measures such as differences in means, odds ratios, relative risks and correlations.

We first need to compute the variance of the sufficiently sceptical prior. Matthews (2001a) has shown that the equi-tailed credible interval of the sufficiently sceptical prior at level $1-\alpha$ has limits $\pm S$ where

$$S = \frac{(U-L)^2}{4\sqrt{UL}} \tag{1}$$

is the *scepticism limit* (Matthews, 2018). Note that (1) holds for any value of α, not just for the traditional 5% level. The sufficiently sceptical prior variance τ^2 can be derived

from (1) and expressed as a function of the variance σ_o^2 (the squared standard error, assumed to be known) of the estimate $\hat{\theta}_o$, the corresponding test statistic $t_o = \hat{\theta}_o/\sigma_o$ and $z_{\alpha/2}$, the $1 - \alpha/2$ quantile of the standard normal distribution:

$$\tau^2 = \frac{\sigma_o^2}{t_o^2/z_{\alpha/2}^2 - 1}, \tag{2}$$

where $t_o^2 > z_{\alpha/2}^2$ holds due to significance of the original study at level α. See Appendix A for a derivation.

Equation (2) shows that the sufficiently sceptical prior variance τ^2 can be both smaller or larger than σ_o^2, depending on the value of t_o^2. For a "borderline" significant result where t_o^2 is close to $z_{\alpha/2}^2$, the sufficiently sceptical prior variance will be relatively large. If t_o^2 is substantially larger than $z_{\alpha/2}^2$, then the sufficiently sceptical prior variance will be relatively small.

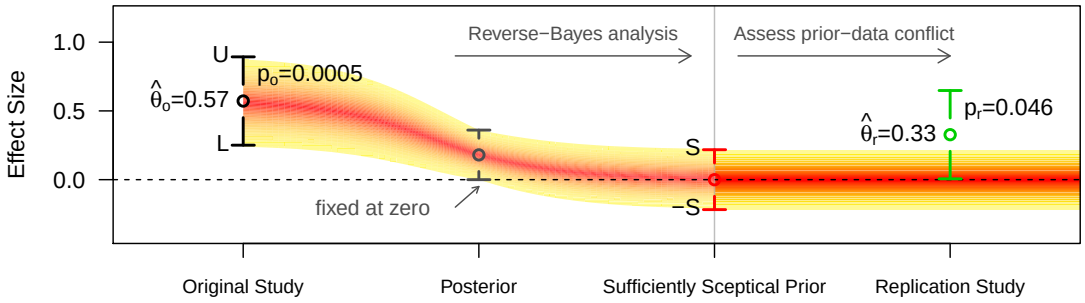

Figure 1: Example of the assessment of replication success. The original study has effect estimate $\hat{\theta}_o = 0.57$ (95% CI from 0.25 to 0.89) and two-sided p-value $p_o = 0.0005$. The reverse-Bayes approach gives the scepticism limit $S = 0.22$. The fan chart in the left part of the figure illustrates the combination of the original study result and the sufficiently sceptical prior into the posterior with lower credible limit fixed at zero. The fan chart in the right part represents the sufficiently sceptical prior and can be used to visually assess potential conflict with the replication study result ($\hat{\theta}_r = 0.33$, 95% CI from 0.01 to 0.65, $p_r = 0.046$).

Figure 1 shows an example of this procedure. The original study has effect estimate

0.57 (95% CI from 0.25 to 0.89) and two-sided p-value $p_o = 0.0005$. The scepticism limit, calculated from (1), turns out to be $S = 0.22$, as illustrated in the left part of Figure 1. The replication study gave an estimated effect of 0.33 (95% CI from 0.01 to 0.65, $p_r = 0.046$). If the replication result is in conflict with the sufficiently sceptical prior, the original result is deemed credible. A visual comparison of the fan chart (displaying the sufficiently sceptical prior) in the right part of Figure 1 with the replication study result helps to assess potential conflict, but in general a more principled statistical approach is needed.

One option is to consider the original study as credible, if the absolute value of the effect estimate $\hat{\theta}_r$ from the replication study is larger than the scepticism limit S (Matthews, 2001a,b). In the above example the effect estimate in the replication study ($\hat{\theta}_r = 0.33$) is larger than the scepticism limit ($S = 0.22$), so the original study would be considered credible at the 5% level. However, a disadvantage of this approach is that it does not take the variance σ_r^2 of the replication estimate $\hat{\theta}_r$ (in the following also assumed to be normally distributed) into account. To address this issue, I propose to quantify prior-data conflict based on the prior-predictive distribution of $\hat{\theta}_r$, a normal distribution with mean zero and variance $\tau^2 + \sigma_r^2$ (Box, 1980; Spiegelhalter et al., 2004, Section 5.8). This leads to the test statistic

$$t_{\text{Box}} = \frac{\hat{\theta}_r}{\sqrt{\tau^2 + \sigma_r^2}} \tag{3}$$

and the tail probability $p_{\text{Box}} = \Pr(\chi^2(1) \geq t_{\text{Box}}^2)$ as the corresponding upper tail of a χ^2-distribution with one degree of freedom. Small values of p_{Box} indicate a conflict between the sufficiently sceptical prior and the estimate from the replication study and I define *replication success* at level α if $p_{\text{Box}} \leq \alpha$.

In the introductory example shown in Figure 1, the prior-predictive assessment of conflict between the sceptical prior and the replication effect estimate gives $t_{\text{Box}} =$

1.65 with Box's tail probability $p_{\text{Box}} = 0.098 > 0.05$, so the replication study is not successful at the 5% level, although both the original and the replication study are significant at that level. This illustrates that replication success is a more stringent criterion than significance alone. For $\alpha = 10\%$, Box's tail probability is somewhat smaller ($p_{\text{Box}} = 0.078$), and we can declare replication success at the 10% level.

The example illustrates how Box's tail probability can be used to assess replication success at level α. However, it is difficult to interpret the actual value of p_{Box} as it depends on the choice of α. Furthermore, assessment of replication success is only possible if the original study result is significant at level α as otherwise the sufficiently sceptical prior would not exist and p_{Box} could not be computed. These issues motivate the work described in the next section where I introduce the sceptical p-value, a measure for replication success that is independent of the level α.

3. The Sceptical p-Value

Instead of dichotomizing replication studies into successful yes/no at some arbitrary threshold α, I now propose the *sceptical p-value*[1] p_S to assess replication success quantitatively. The idea is to determine the largest confidence level $1 - p_S$ for the original confidence interval, where we are able to declare replication success at level p_S. Replication success at a pre-specified level α is then equivalent to $p_S \leq \alpha$. This parallels the duality of ordinary p-values and confidence intervals, where the largest confidence level $1 - p$ where we are able to declare significance can be used to compute the ordinary p-value p.

To determine p_S, let $c = \sigma_o^2/\sigma_r^2$ denote the ratio of the variances of the original and replication effect estimates and let $t_r = \hat{\theta}_r/\sigma_r$ denote the test statistic of the replication

[1] In previous unpublished work (Held, 2017), not related to replication studies, this was called the p-value for credibility. Uri Simonsohn pointed out to me that in psychological science the term credibility is already linked to a different concept.

study. With (2) we can derive the prior-predictive variance of $\hat{\theta}_r$:

$$\tau^2 + \sigma_r^2 = \sigma_r^2 \left(\frac{c}{t_o^2/z_{\alpha/2}^2 - 1} + 1 \right). \tag{4}$$

Using (3) and (4), the requirement $t_{\text{Box}}^2 = \hat{\theta}_r^2/(\tau^2 + \sigma_r^2) \geq z_{\alpha/2}^2$ for replication success turns out to be equivalent to

$$\left(t_o^2/z_{\alpha/2}^2 - 1\right)\left(t_r^2/z_{\alpha/2}^2 - 1\right) \geq c, \tag{5}$$

see Appendix B for a derivation. Significance of the original study implies that $z_{\alpha/2}^2 < t_o^2$ holds, therefore $z_{\alpha/2}^2 < t_r^2$ must also hold to ensure that the left hand side of equation (5) is positive.

The required squared quantile $z_S^2 = z_{p_S/2}^2$ to obtain equality in (5) defines the sceptical p-value $p_S = 2\left[1 - \Phi(|z_S|)\right]$ via

$$\left(t_o^2/z_S^2 - 1\right)\left(t_r^2/z_S^2 - 1\right) = c \tag{6}$$

and the requirement $p_S \leq \alpha$ for replication success at level α translates to $z_S^2 \geq z_{\alpha/2}^2$. Equation (6) can be re-written as

$$(c-1)z_S^4 + 2z_S^2\, t_A^2 = t_A^2\, t_H^2, \tag{7}$$

where $t_A^2 = (t_o^2 + t_r^2)/2$ is the arithmetic and $t_H^2 = 2/(1/t_o^2 + 1/t_r^2)$ the harmonic mean of the squared test statistics t_o^2 and t_r^2. The only solution of (7) that fulfills the requirement $0 \leq z_S^2 < \min\{t_o^2, t_r^2\}$ is

$$z_S^2 \;=\; \begin{cases} t_H^2/2 & \text{for } c = 1 \text{ and} \\ \frac{1}{c-1}\left\{\sqrt{t_A^2\left[t_A^2 + (c-1)t_H^2\right]} - t_A^2\right\} & \text{for } c \neq 1. \end{cases} \tag{8}$$

In the introductory example the original and the replication confidence intervals have the same width, so $c = 1$ and z_S^2 is simply half the harmonic mean of $t_o^2 = 12.19$ and $t_r^2 = 3.99$, *i. e.* $z_S^2 = 3.00$, $|z_S| = 1.73$ and $p_S = 2\,[1 - \Phi(1.73)] = 0.083$. We can thus declare replication success at any level $\alpha \geq 0.083$.

3.1. Properties

The requirement $z_S^2 < \min\{t_o^2, t_r^2\}$ implies that the sceptical p-value p_S is always larger than both the original and the replication p-values p_o and p_r. Closer inspection of equation (6) shows that z_S^2 is increasing with increasing t_o^2 (for fixed t_r^2 and c) and also with increasing t_r^2 (for fixed t_o^2 and c). Therefore, the smaller p_o (or p_r), the smaller p_S (for fixed c). Furthermore, for fixed test statistics t_o and t_r (so fixed p-values p_o and p_r), the solution z_S^2 of (6) will decrease with increasing variance ratio

$$c = \frac{\sigma_o^2}{\sigma_r^2} = \frac{\hat{\theta}_o^2}{\hat{\theta}_r^2} \frac{t_r^2}{t_o^2}.$$

Since t_r^2/t_o^2 is fixed, c increases with increasing squared effect size ratio $\hat{\theta}_o^2/\hat{\theta}_r^2$. In other words, for the same ordinary p-values p_o and p_r, the sceptical p-value p_S increases with smaller absolute replication effect estimate relative to the original effect estimate. This is a desired property, as replication studies with smaller effect estimates than the original estimates are considered less credible (Simonsohn, 2015).

To illustrate the dependence of p_S on the variance ratio c, consider a scenario where $p_o = p_r = 0.01$, so $t_o^2 = t_r^2$ and therefore $c = \hat{\theta}_o^2/\hat{\theta}_r^2$. First assume equal effect sizes $\hat{\theta}_o = \hat{\theta}_r$, so $c = 1$. The sceptical p-value turns out to be $p_S = 0.069$. For $\hat{\theta}_o = 2\,\hat{\theta}_r$ ($c = 4$) we obtain a larger value ($p_S = 0.14$) because the effect estimate $\hat{\theta}_r$ of the replication study is just half as large as the original estimate $\hat{\theta}_o$. On the other hand, for $\hat{\theta}_r = 2\,\hat{\theta}_o$ ($c = 1/4$) the sceptical p-value gets smaller ($p_S = 0.035$). This asymmetry in the incorporation of the original and replication study data is natural, placing less weight

on replication studies with relatively small effect estimates. This is the case in the introductory example, where substantial shrinkage of the replication effect estimate leads to a relatively large sceptical p-value.

It is also interesting to study limiting values of the sceptical p-value. If we let $\sigma_o^2 \downarrow 0$ for fixed $\hat{\theta}_o \neq 0$, (7) reduces to the requirement $z_S^2 = t_r^2$, as shown in Appendix C. Thus, the ordinary p-value of the replication study can be seen as a special case of the sceptical p-value if the uncertainty of the original effect estimate is ignored. On the other hand, ignoring the uncertainty of $\hat{\theta}_r \neq 0$ via $\sigma_r^2 \downarrow 0$ leads to $z_S^2 \downarrow z_M^2$ where

$$z_M^2 = \frac{\sqrt{d\,(d+4)} - d}{2}\, t_o^2, \tag{9}$$

with $d = \hat{\theta}_r^2 / \hat{\theta}_o^2$, see Appendix C for a proof. Using the criterion $z_M^2 \geq z_{\alpha/2}^2$ rather than $z_S^2 \geq z_{\alpha/2}^2$ to assess replication success corresponds to the Matthews (2001a) approach. For any value of d, z_M^2 is smaller than t_o^2 but can be larger than t_r^2. Ignoring the uncertainty of the replication effect estimate may thus lead to the declaration of replication success, even if the replication study is not conventionally significant on its own.

We may also consider the case $c \downarrow 0$ for fixed t_o^2 and t_r^2, where (8) increases with limit

$$z_S^2 \uparrow \min\{t_o^2, t_r^2\}, \tag{10}$$

as shown in Appendix D. Therefore $p_S \downarrow \max\{p_o, p_r\}$ for $c \downarrow 0$, which we will use in Section 3.4.

3.2. Relationship to Intrinsic Credibility

The concept of intrinsic credibility has been proposed in Matthews (2018) to check the credibility of "out of the blue" findings without any prior support. In the present context this corresponds to an original study in the absence of a replication study. The idea is to evaluate the credibility of the original study if we would be able to observe

exactly the same result in the replication study.

The approach by Matthews (2018) corresponds to the case where we ignore the uncertainty of the (hypothetical) replication study and thus leads to (9) with $d = 1$: $z_M^2 = (\sqrt{5}-1)/2\, t_o^2 \approx 0.618\, t_o^2$. However, if we incorporate the uncertainty using the prior-predictive approach by Box (1980), then we obtain $z_S^2 = 0.5\, t_o^2$ as a special case of (8) for $c = 1$ and $t_o^2 = t_r^2$. Now p_S reduces to the p-value for intrinsic credibility,

$$p_{IC} = 2\left[1 - \Phi\left(t_o/\sqrt{2}\right)\right], \tag{11}$$

as proposed in Held (2018) for the assessment of claims of new discoveries. Intrinsic credibility at level α is achieved if $p_{IC} \leq \alpha$, *i.e.* $t_o^2 \geq 2\, z_{\alpha/2}^2$, which is equivalent to $p_o \leq \alpha_{IC}$, where

$$\alpha_{IC} \;=\; 2\left\{1 - \Phi\left(\sqrt{2}\, z_{\alpha/2}\right)\right\} \tag{12}$$

is the p-value threshold for intrinsic credibility. For $\alpha = 0.05$ we obtain $\alpha_{IC} = 0.0056$, for $\alpha = 0.10$ we have $\alpha_{IC} = 0.02$. These thresholds will become important in Section 4.

3.3. One-Sided Sceptical p-Values

The procedure described above is designed for standard two-sided confidence intervals and assesses replication success in a two-sided fashion, as the sign of $\hat{\theta}_r$ does not matter in the computation of the sceptical p-value. In extreme cases, it may therefore happen that a replication study is classified as successful although the sign of $\hat{\theta}_o$ and $\hat{\theta}_r$ differ, reflecting the form of the alternative hypothesis H_1: $\theta \neq 0$ in standard two-sided hypothesis tests. This "replication paradox" may also occur in a Bayes factor approach, see Ly et al. (2017) for details.

It is therefore of interest to adapt the sceptical p-value to the one-sided setting. Without loss of generality consider the one-sided alternative H_1: $\theta > 0$ to H_0: $\theta = 0$

and assume that we observe $\hat{\theta}_o > 0$. We now need to use a one-sided confidence interval for θ whose lower limit at level $\tilde{\alpha}$ equals the lower limit L of the corresponding two-sided confidence interval at level $2\tilde{\alpha}$. The variance of the sceptical prior therefore equals (2) with $z_{\alpha/2}$ replaced by $z_{\tilde{\alpha}}$. The corresponding one-sided requirement for replication success is now $t_{\text{Box}} \geq z_{\tilde{\alpha}}$, ensuring that the replication paradox cannot occur. To derive the one-sided sceptical p-value $\tilde{p}_S$, we can therefore still use (6) with $z_S^2 = z_{\tilde{p}_S}^2$, so the one-sided sceptical p-value is just half of the two-sided sceptical p-value, $\tilde{p}_S = p_S/2$, if $z_S \geq 0$. Of course, the same property holds for ordinary p-values, which implies that the one-sided sceptical p-value will always be larger than the ordinary one-sided p-values from the original and replication study, respectively.

One-sided sceptical p-values are appropriate if the study protocol of the original study is already formulated in a one-sided fashion, as is often the case in confirmatory pharmaceutical trials (often with $\tilde{\alpha} = 2.5\%$). In exploratory research, however, the direction of the assumed effect is rarely pre-specified and then the two-sided sceptical p-value should be used. A post-hoc (after the original study results are known) formulation of a one-sided alternative would require halving the original two-sided significance level α to $\alpha/2$, say. Use of the one-sided sceptical p-value would then be equivalent to the two-sided procedure at level α, if the sign of the original and replication effect estimates agree. This modified procedure ensures that the replication paradox cannot occur.

3.4. The Distribution Under The Null

It is interesting to compare the distributions of p_o (or p_r), p_{IC} and p_S under the assumption of no effect, under which the ordinary p-value is uniformly distributed. We can easily derive the density of p_{IC} with a change-of-variables using (11): $f(p_{IC}) = 2\sqrt{\pi}\,\varphi\{t(p_{IC})\}$, here $\varphi(.)$ is the standard normal density function and $t(p_{IC}) = \Phi^{-1}(1 - p_{IC}/2)$.

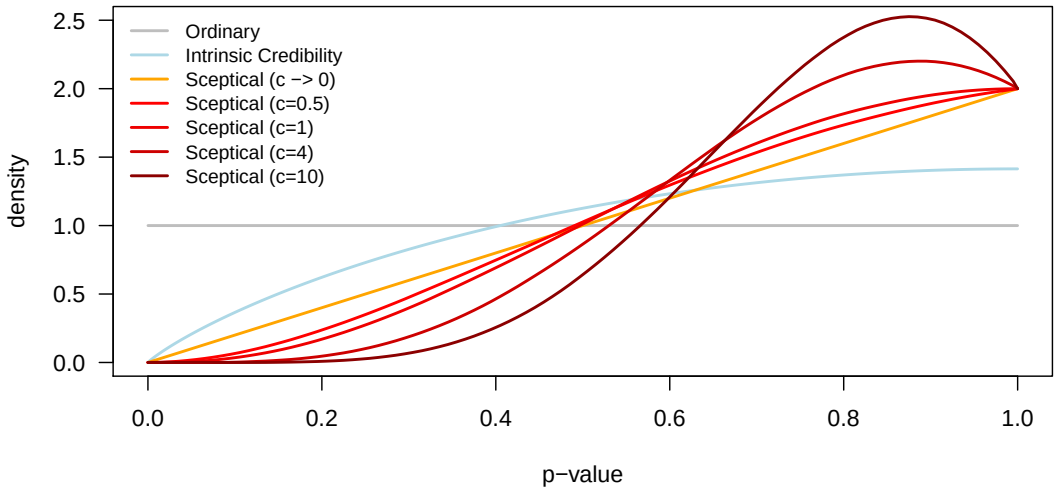

Figure 2: The density function of the sceptical p-value for different values of the variance ratio c under the assumption of no effect. The density for the limiting case $c \to 0$ as well as the density of the p-value for intrinsic credibility and the ordinary p-value are also shown.

The distribution of p_S can be studied via stochastic simulation. Density estimates are displayed in Figure 2 for different values of the variance ratio c based on 5×10^6 samples each. We can see that the risk of small "false positive" sceptical p-values is drastically reduced, compared to ordinary p-values based on one study only. Note that the variance is usually inversely proportional to the sample size of each study, *i. e.* $\sigma_o^2 = \kappa^2/n_o$ and $\sigma_r^2 = \kappa^2/n_r$ for some unit variance κ^2, say. Then $c = n_r/n_o$, so the variance factor c is increasing with increasing sample size n_r of the replication study. The distribution of p_S in Figure 2 is shifted to the right with increasing c, so an increasing sample size of the replication study reduces the risk of a false claim of replication success.

From (10) we know that for $c \downarrow 0$ we have $p_S \downarrow \max\{p_o, p_r\}$, which follows a triangular $\text{Be}(2,1)$ distribution if p_r and p_o are independently uniform. The correspond-

ing density function is shown in Figure 2, as well as the density function of p_o and p_{IC}. The triangular distribution gives the upper bound α^2 for the tail probability $\Pr(p_S \leq \alpha \,|\, H_0)$ for sufficiently small α and any value of the variance ratio c. For example, for $\alpha = 0.05$ we obtain $\Pr(p_S \leq 0.05 \,|\, H_0) \leq 0.0025$ for any c. This is to be compared with $\Pr(p_o \leq 0.05 \,|\, H_0) = 0.05$ and $\Pr(p_{IC} \leq 0.05 \,|\, H_0) = 0.0056$. However, α^2 is not a particularly sharp bound. If, for example, the replication sample size equals the original sample size ($c = 1$), then $\Pr(p_S \leq 0.05 \,|\, H_0) \approx 0.0001$ is much smaller than 0.0025.

4. Power and Sample Size Calculations

Replication success is not only a function of the two p-values from the original and replication study, but also of sample size, which enters in the variance ratio c. The computation of the power or the required sample size to achieve replication success is hence more challenging than in standard sample size calculations. A larger sample size will be required since replication success (defined as $p_S \leq \alpha$) implies significance of the replication study ($p_r < p_S$). Furthermore, the required sample size will depend on the p-value p_o from the original study. If there is strong evidence from the original study, p_o is very small and the required sample size for replication success will be smaller than if p_o is relatively large.

The results from the original study will enter in two different ways, as prior distribution for the effect size and in the assessment of replication success. For the former I will distinguish two cases, a normal prior and a point prior. The normal prior incorporates the uncertainty of $\hat{\theta}_o$ while the point prior does not. Suppose n_o is the size of the original study sample and n_r the sample size of the replication study, so $\sigma_o^2 = \kappa^2/n_o$ and $\sigma_r^2 = \kappa^2/n_r$, where κ^2 is the unit variance from one observation. Then $c = n_r/n_o$, which would also hold in a balanced two-sample design with re-

spective sample size n_o and n_r per group. Under an initial uniform prior for θ, the sampling distribution $\hat{\theta}_o \sim \mathrm{N}(\theta, \sigma_o^2)$ of the original study now serves as prior distribution $\theta \,|\, \hat{\theta}_o \sim \mathrm{N}\left(\hat{\theta}_o, \sigma_o^2 = \kappa^2/n_o\right)$ with prior-predictive distribution

$$\hat{\theta}_r \,|\, \hat{\theta}_o \sim \mathrm{N}\left(\hat{\theta}_o, \sigma_o^2 + \sigma_r^2 = \kappa^2 \left\{\frac{1}{n_o} + \frac{1}{n_r}\right\}\right) \tag{13}$$

for the observed effect $\hat{\theta}_r$ in the replication study. Then t_r^2 follows a scaled non-central χ^2-distribution with one degree of freedom, scaling factor $(n_r + n_o)/n_o$ and non-centrality parameter $\lambda = (n_r n_o)/(n_r + n_o) \cdot \hat{\theta}_o^2/\kappa^2$, as shown in Appendix E. For the alternative point prior at $\theta = \hat{\theta}_o$, t_r^2 follows a non-central χ^2-distribution with one degree of freedom and non-centrality parameter $\lambda = n_r \cdot \hat{\theta}_o^2/\kappa^2$.

To compute the power for success of a replication study with fixed sample n_r, the entry $c = \sigma_o^2/\sigma_r^2 = n_r/n_o$ in (5) is fixed. Then z_S^2 and the sceptical p-value p_S are monotone functions of t_r^2 and we can compute the power for replication success at any level α. We can also calculate the required replication sample size n_r at some pre-defined power for replication success. Both goals require application of numerical root-finding algorithms. Computational details are omitted here.

4.1. Power calculations

Figure 3 compares the power for significance with the power for replication success for a replication study with sample size equal to the original study ($n_r = n_o$) at level $\alpha = 5\%$ as a function of the p-value p_o of the original study. Power calculations for significance aim to detect the effect estimate $\hat{\theta}_o$ from the original study with a standard two-sided significance test. Not accounting for the associated uncertainty corresponds to a point prior at $\hat{\theta}_o$, whereas a normal distribution with mean $\hat{\theta}_o$ and variance σ_o^2 (Spiegelhalter et al., 2004, Equation 6.4) is used to incorporate the uncertainty. The results are in accordance with those reported in Goodman (1992). In particular, for

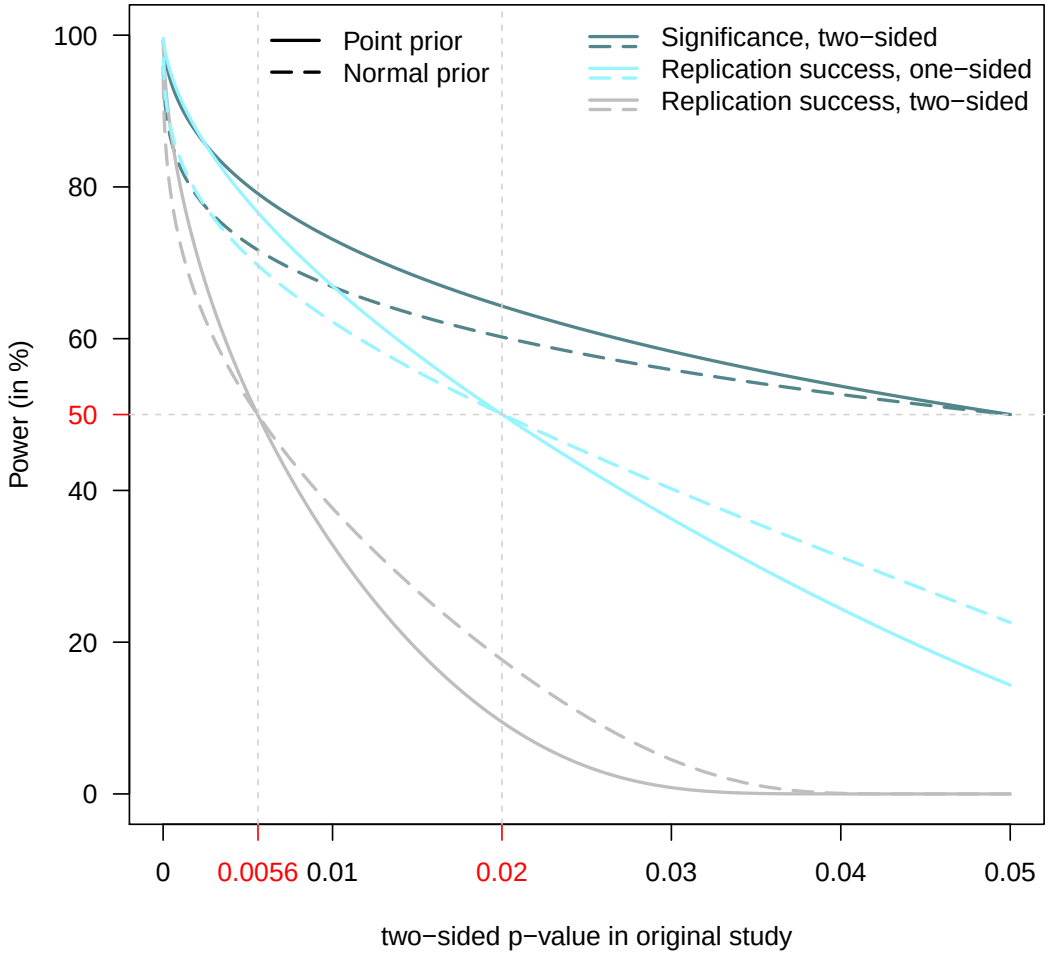

Figure 3: Power calculations for a replication study with sample size equal to the original study. Shown is the power for significance and for replication success, both at level $\alpha = 5\%$, as a function of the p-value of the original study.

$p_o = 0.05$ the power is 50% both under the point prior and the normal prior. The power increases for smaller p-values p_o, with higher values under the point prior.

The power for replication success (both two-sided and one-sided) is also shown in Figure 3, again based on the point or the normal prior and $\alpha = 5\%$. As expected,

the power for replication success is lower than for significance, and drops quickly to zero for values of p_o close to 0.05. Remarkably, in the two-sided case under both the point and the normal prior, a power of 50% is attained at $p_o = 0.0056$. This is the threshold (12) for intrinsic credibility at level $\alpha = 5\%$, as described in Section 3.2. In the one-sided case a power of 50% is obtained at $p_o = 0.02$, the threshold for intrinsic credibility at two-sided level $\alpha = 10\%$. Therefore, only intrinsically credible results (based on the Held (2018) threshold (12)) ensure that the power for success of an identically designed replication study exceeds 50%. This intriguing feature highlights the difficulty to achieve replication success if the evidence from the original study is only suggestive and provides a new argument for the recently proposed 0.005 p-value threshold for claims of new discoveries (Johnson, 2013; Benjamin et al., 2018; Ioannidis, 2018).

This surprising result can be explained as follows: If the non-centrality parameter λ of a non-central χ^2-distribution is reasonably large, say $\lambda > 4$, then the median is approximately equal to λ. Under the point prior and for $n_r = n_o$, the non-centrality parameter of the distribution of t_r^2 is $\lambda = t_o^2$, so $\text{Med}(t_r^2) \approx t_o^2$. The sceptical p-value is defined for $n_o = n_r$ through $z_S^2 = t_H^2/2 = (1/t_o^2 + 1/t_r^2)^{-1}$, so the median of z_S^2 is approximately $t_o^2/2$. Replication success is thus achieved with 50% probability for $z_{\alpha/2}^2 = z_S^2 \approx t_o^2/2$, *i. e.* $t_o^2 \approx 2\, z_{\alpha/2}^2$. This corresponds to the intrinsic credibility threshold (12) for the ordinary p-value p_o from the original study.

We obtain essentially the same result under the normal prior, where now $\lambda = t_o^2/2$, which combined with a scaling factor of 2 also leads to $\text{Med}(t_r^2) \approx t_o^2$ for sufficiently large t_o^2. The rest of the argument is as above, but note that this approximation is slightly less precise, because the non-centrality parameter is half as large than under the point prior. The approximation is, however, still very good: For $\alpha = 5\%$, the exact power for replication success at $p_o = 2(1 - \Phi(\sqrt{2}\, z_{0.025} = 2.77)$ is 50.00002% for the point prior and 50.00461% for the normal prior.

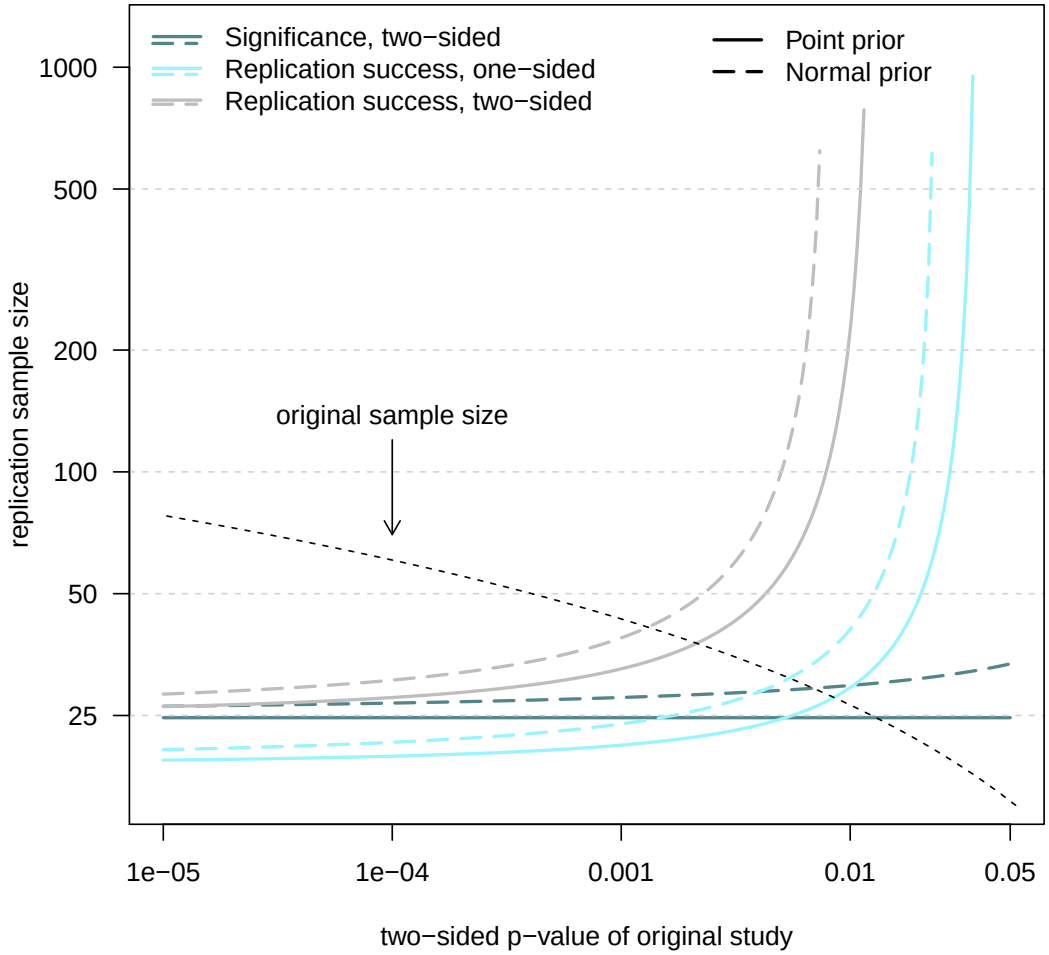

Figure 4: Replication sample size n_r to achieve replication success with 80% power for $\hat{\theta}_o = 0.5$, $\kappa = 1$ and $\alpha = 0.05$. The top axis gives the sample size of the original study as a function of the original two-sided p-value p_o (bottom axis).

4.2. Sample size calculations

Figure 4 compares different sample size calculations for $\alpha = 0.05$, $\hat{\theta}_o = 0.5$, $\kappa = 1$ and a pre-specified power of 80% to achieve replication success. The sample size n_o of

the original study varies between 15 and 78, which corresponds to ordinary p-values p_o between 10^{-5} and 0.05. Standard (two-sided) sample size calculations to detect $\hat{\theta}_o$ give $n_r = 24.7$, independent of the significance of the original study. Incorporating the uncertainty from the original study with a normal prior gives sample sizes between 26.3 and 33.5.

As expected, the required (two-sided) sample size n_r for replication success is larger than the one for significance alone. Furthermore, somewhat larger sample sizes are required under the normal prior. The required sample size n_r now depends dramatically on the p-value p_o of the original study. If p_o is small, say smaller than 0.001, then n_r is even smaller than the original sample size n_o. However, the required sample size explodes for larger p-values with an asymptote around $p_o = 0.011$ under the point prior and $p_o = 0.007$ under the normal prior. For original p-values between 0.01 and 0.05, it is thus nearly impossible to obtain 80% power for replication success. The curves shift only a little to the right when we assess replication success in a one-sided fashion, pushing the asymptotes towards $p_0 = 0.034$ and $p_0 = 0.023$, respectively.

5. Replication Success in Psychological Science

I now re-analyse data from the Open Science Collaboration (2015) project on the replicability of psychological science. It is possible to transform effect sizes to the correlation scale for 73 studies, the so-called Meta-Analytic subset (Johnson et al., 2016). Application of Fisher's z-transformation $z(\rho) = \tanh^{-1}(\rho)$ to the estimated correlations $\hat{\rho}$ justifies a normal assumption with the standard error being a function only of the study sample size n: $\text{se}(z(\hat{\rho})) = 1/\sqrt{n-3}$.

Figure 5 displays the replication versus the original correlation estimates. Eight of the 73 original studies are not significant at the standard $\alpha = 5\%$ level, three of them with p-values between 0.05 and 0.06. There have been 21 significant replication

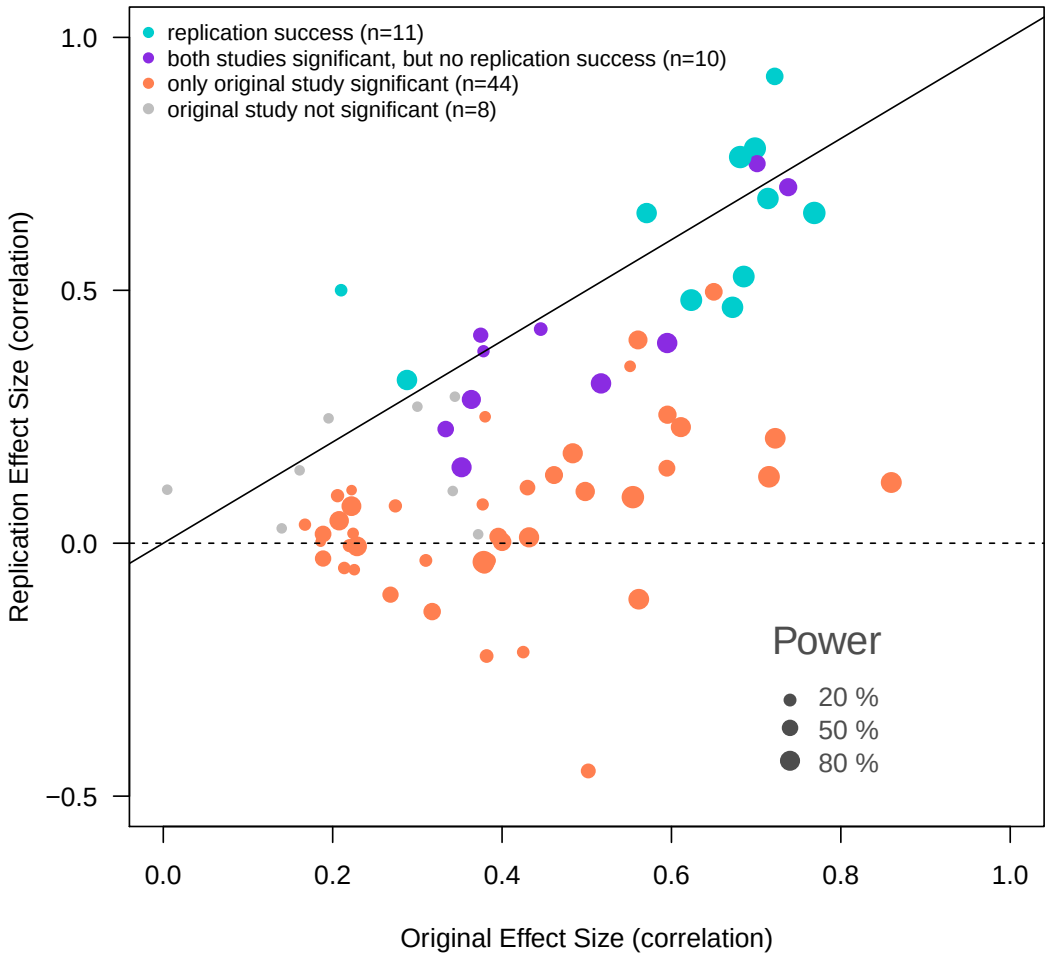

Figure 5: Application to Open Science Collaboration (2015) data: The circles represent the effect estimates (correlations) of original and replication studies. The circle size represents the power for replication success at the two-sided 5% level based on the normal prior. Replication success and significance is assessed at the two-sided 5% level and indicated by the color of the circles.

studies following from the 65 significant original studies. The sceptical p-value allows us to rank the studies by the degree of replication success. Table 1 lists the 24 most successful replication studies with $p_S \leq 0.15$ of which the top 11 have been successful at the two-sided 5% level ($p_S \leq 0.05$). The remaining 13 studies in Table 1 (with $p_S > 0.05$) show some interesting features. For example, study 18 has a non-significant replication result but still leads to replication success at the 10% level. Conversely, there are several studies with both $p_o \leq 0.05$ and $p_r \leq 0.05$ but $p_S > 0.10$. This illustrates once again, that the sceptical p-value does not only take significance of the original and replication study into account, but also effect and sample sizes, both entering in the variance ratio c.

6. Discussion

There is pressing need for new methodology to design and analyse replication studies. The proposed approach follows the spirit of "evolution rather than revolution" (Matthews, 2018) and provides a framework for extracting more insight from replication studies based on standard metrics (effect estimates, confidence intervals and p-values). Instead of synthesising original and replication study results through a meta-analysis, the original study result is challenged with the sufficiently sceptical prior. Replication success is then defined as conflict between the sufficiently sceptical prior and the replication effect estimate. The resulting sceptical p-value is a natural extension of the ordinary p-value of the replication study. While the latter quantifies the conflict between the point null hypothesis and the replication data, the former quantifies the conflict between the sufficiently sceptical prior and the replication data. The sceptical p-value p_S thus takes into account the original study result, whereas the ordinary p-value does not.

Significance of both the original and the replication study is a necessary but not

	Original study			Replication study			Replication Success	
	n_o	$\hat{\rho}_o$	p_o	n_r	$\hat{\rho}_r$	p_r	Power	p_S
1	126	0.68	< 0.0001	177	0.76	< 0.0001	> 99.9	< 0.0001
2	78	0.77	< 0.0001	38	0.65	< 0.0001	> 99.9	< 0.0001
3	30	0.70	< 0.0001	31	0.78	< 0.0001	96.9	0.0005
4	174	0.29	0.0001	141	0.32	< 0.0001	83.4	0.005
5	32	0.57	0.0005	32	0.65	< 0.0001	83.0	0.007
6	22	0.71	< 0.0001	22	0.68	0.0003	91.9	0.008
7	38	0.62	< 0.0001	39	0.48	0.002	95.3	0.011
8	30	0.69	< 0.0001	27	0.53	0.004	94.6	0.015
9	117	0.21	0.023	236	0.50	< 0.0001	17.4	0.033
10	23	0.67	0.0003	31	0.47	0.007	92.2	0.038
11	9	0.72	0.026	18	0.92	< 0.0001	59.4	0.048
12	154	0.36	< 0.0001	50	0.28	0.045	72.0	0.052
13	40	0.37	0.017	95	0.41	< 0.0001	39.7	0.06
14	11	0.70	0.014	11	0.75	0.006	55.9	0.067
15	25	0.59	0.001	33	0.40	0.022	82.5	0.072
16	41	0.52	0.0004	41	0.32	0.044	82.9	0.08
17	9	0.74	0.021	16	0.70	0.002	64.2	0.091
18	33	0.56	0.0005	21	0.40	0.071	70.9	0.096
19	33	0.38	0.029	72	0.38	0.0009	15.2	0.10
20	25	0.45	0.025	39	0.42	0.007	26.6	0.10
21	57	0.33	0.011	118	0.23	0.014	50.6	0.11
22	96	0.20	0.057	243	0.25	< 0.0001	0.0	0.12
23	16	0.65	0.005	13	0.50	0.085	60.6	0.13
24	69	0.35	0.003	178	0.15	0.045	80.5	0.14

Table 1: Results for the 24 most successful replication studies with $p_S \leq 0.15$, as listed in the last column. The penultimate column gives the power for replication success (in %) based on the normal prior at level $\alpha = 5\%$.

sufficient requirement for replication success. The proposed framework thus forms a more stringent statistical basis for the traditional FDA (and EMA) approach for drug approval following the "two pivotal study paradigm" requiring two significant findings from two independent confirmatory trials (Lee, 2018). In particular, the concept of replication success also takes into account effect and sample sizes. However, the difficulty to achieve replication success if the evidence from the original study is only suggestive underlines the need for more stringent p-value thresholds for claims of new discoveries. The threshold for intrinsic credibility (12) is a natural choice for this task.

It would be interesting to combine original and replication effect estimates into an overall summary measure within the proposed reverse-Bayes framework. Some down-weighting of the original study result would in general be required depending on the degree of replication success. The overall estimate could then be used as a new "original" effect estimate in order to assess the success of a second replication study. This would open up new ways to iteratively challenge existing knowledge through a series of replication studies and would provide an interesting alternative to traditional evidence synthesis methods.

The proposed reverse-Bayes approach assumes a simple mathematical framework, where likelihood, prior and posterior are all normally distributed. It will be of interest to extend this framework to other settings, for example to the t-distribution.

Data and Software Availability Data analyzed in this article are originally from Open Science Collaboration (2015) and have been downloaded from `https://osf.io/fgjvw/`. Software to compute the sceptical p-value and the power or sample size of replication studies will be made available in the R-package `pCalibrate` available on the Comprehensive R Archive Network (`https://CRAN.R-project.org/package=pCalibrate`).

Acknowledgments I am grateful to Robert Matthews, Uri Simonsohn and the members of the UZH Department of Biostatistics for helpful discussions and suggestions. I also acknowledge helpful comments by a referee on a related grant proposal of mine.

References

Anderson, S. F. and Maxwell, S. E. (2017). Addressing the "replication crisis": Using original studies to design replication studies with appropriate statistical power. *Multivariate Behavioral Research*, 52(3):305–324. `https://doi.org/10.1080/00273171.2017.1289361`.

Bayarri, M. J. and Mayoral, M. (2002). Bayesian design of "successful" replications. *The American Statistician*, 56:207–214. `https://www.doi.org/10.1198/000313002155`.

Begley, C. G. and Ioannidis, J. P. (2015). Reproducibility in science. *Circulation Research*, 116(1):116–126. http://circres.ahajournals.org/content/116/1/116.

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., et al. (2018). Redefine statistical significance. *Nature Human Behaviour*, 2:6–10. http://dx.doi.org/10.1038/s41562-017-0189-z.

Box, G. E. P. (1980). Sampling and Bayes' inference in scientific modelling and robustness (with discussion). *Journal of the Royal Statistical Society, Series A*, 143:383–430. https://www.jstor.org/stable/2982063.

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., and Munafò, M. R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, 14. http://dx.doi.org/10.1038/nrn3475.

Camerer, C. F., Dreber, A., Forsell, E., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Almenberg, J., Altmejd, A., Chan, T., Heikensten, E., Holzmeister, F., Imai, T., Isaksson, S., Nave, G., Pfeiffer, T., Razen, M., and Wu, H. (2016). Evaluating replicability of laboratory experiments in economics. *Science*, 351(6280):1433–1436. https://doi.org/10.1126/science.aaf0918.

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Nave, G., Nosek, B. A., Pfeiffer, T., Altmejd, A., Buttrick, N., Chan, T., Chen, Y., Forsell, E., Gampa, A., Heikensten, E., Hummer, L., Imai, T., Isaksson, S., Manfredi, D., Rose, J., Wagenmakers, E.-J., and Wu, H. (2018). Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour*, 2(9):637–644. https://doi.org/10.1038/s41562-018-0399-z.

Colquhoun, D. (2017). The reproducibility of research and the misinterpretation of p-values. *Royal Society Open Science*, 4(12). http://dx.doi.org/10.1098/rsos.171085.

Fletcher, A., Spiegelhalter, D., Staessen, J., Thijs, L., and Bulpitt, C. (1993). Implications for trials in progress of publication of positive results. *The Lancet*, 342(8872):653–657. https://doi.org/10.1016/0140-6736(93)91762-b.

Good, I. J. (1950). *Probability and the Weighing of Evidence*. Griffin, London, UK.

Good, I. J. (1983). *Good Thinking: The Foundations of Probability and Its Applications*. University of Minnesota Press, Minneapolis.

Goodman, S. N. (1992). A comment on replication, p-values and evidence. *Statistics in Medicine*, 11(7):875–879. https://doi.org/10.1002/sim.4780110705.

Goodman, S. N., Fanelli, D., and Ioannidis, J. P. A. (2016). What does research reproducibility mean? *Science Translational Medicine*, 8(341):341ps12–341ps12. https://doi.org/10.1126/scitranslmed.aaf5027.

Greenland, S. (2006). Bayesian perspectives for epidemiological research: I. foundations and basic methods. *International Journal of Epidemiology*, 35:765–775. https://doi.org/10.1093/ije/dyi312.

Greenland, S. (2011). Null misinterpretation in statistical testing and its impact on health risk assessment. *Preventive Medicine*, 53:225–228. https://doi.org/10.1016/j.ypmed.2011.08.010.

Held, L. (2013). Reverse-Bayes analysis of two common misinterpretations of significance tests. *Clinical Trials*, 10:236–242. https://doi.org/10.1177/1740774512468807.

Held, L. (2017). *P*-values for credibility. Technical report. https://arxiv.org/abs/1712.03032.

Held, L. (2018). The assessment of intrinsic credibility and a new argument for $p <$ 0.005. Technical report. https://arxiv.org/abs/1803.10052.

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

Ioannidis, J. P. A. (2018). The proposal to lower p value thresholds to .005. *JAMA: The Journal of the American Medical Association*. https://doi.org/10.1001/jama.2018.1536.

Johnson, V. E. (2013). Revised standards for statistical evidence. *Proceedings of the National Academy of Sciences of the United States of America*, 110(48):19313–19317. https://doi.org/10.1073/pnas.1313476110.

Johnson, V. E., Payne, R. D., Wang, T., Asher, A., and Mandal, S. (2016). On the reproducibility of psychological science. *Journal of the American Statistical Association*, 112(517):1–10. https://doi.org/10.1080/01621459.2016.1240079.

Lee, K.-S. (2018). When the Alpha is the Omega: *P*-values, "Sustantial Evidence," and the 0.05 Standard at FDA. *Food and Drug Law Journal*, 72(4):595–635. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6169785/.

Ly, A., Etz, A., Marsman, M., and Wagenmakers, E.-J. (2017). Replication Bayes factors from evidence updating. Technical report. http://psyarxiv.com/u8m2s.

Matthews, R. (2001a). Methods for assessing the credibility of clinical trial outcomes. *Drug Information Journal*, 35:1469–1478. https://doi.org/10.1177/009286150103500442.

Matthews, R. (2001b). Why *should* clinicians care about Bayesian methods? (with discussion). *Journal of Statistical Planning and Inference*, 94:43–71. https://doi.org/10.1016/S0378-3758(00)00232-9.

Matthews, R. (2018). Beyond "significance": principles and practice of the Analysis of Credibility. *Royal Society Open Science*, 5(1). http://dx.doi.org/10.1098/rsos.171047.

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

Patil, P., Peng, R. D., and Leek, J. T. (2016). What should researchers expect when they replicate studies? a statistical view of replicability in psychological science. *Perspectives on Psychological Science*, 11(4):539–544. https://doi.org/10.1177/1745691616646366.

Schönbrodt, F. D. and Wagenmakers, E.-J. (2018). Bayes factor design analysis: Planning for compelling evidence. *Psychonomic Bulletin & Review*, 25(1):128–142. https://doi.org/10.3758/s13423-017-1230-y.

Simonsohn, U. (2015). Small telescopes: Detectability and the evaluation of replication results. *Psychological Science*, 26(5):559–569. https://doi.org/10.1177/0956797614567341.

Spiegelhalter, D. J., Abrams, K. R., and Myles, J. P. (2004). *Bayesian Approaches to Clinical Trials and Health-Care Evaluation*. Wiley, New York.

Spiegelhalter, D. J., Freedman, L. S., and Parmar, M. K. B. (1994). Bayesian approaches to randomized trials. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 157(3):357. https://doi.org/10.2307/2983527.

van Aert, R. C. M. and van Assen, M. A. L. M. (2017). Bayesian evaluation of effect size after replicating an original study. *PLOS ONE*, 12(4):1–23. https://doi.org/10.1371/journal.pone.0175302.

Verhagen, J. and Wagenmakers, E.-J. (2014). Bayesian tests to quantify the result of a replication attempt. *Journal of Experimental Psychology*, 143:1457–1475. http://dx.doi.org/10.1037/a0036731.

Appendix

A. Proof of equation (2)

With $U, L = \hat{\theta} \pm z_{\alpha/2}\,\sigma$ we have $U\,L = \hat{\theta}^2 - z_{\alpha/2}^2\,\sigma^2$ and $U - L = \text{sign}\,(U - L)\,2\,z_{\alpha/2}\,\sigma$. We therefore obtain with (1) and $\tau = S/z_{\alpha/2}$:

$$\tau^2 = \frac{S^2}{z_{\alpha/2}^2} = \frac{(U-L)^4}{16 z_{\alpha/2}^2 UL} = \frac{(2 z_{\alpha/2}\sigma)^4}{16 z_{\alpha/2}^2 UL} = \frac{z_{\alpha/2}^2 \sigma^4}{\hat{\theta}^2 - z_{\alpha/2}^2\sigma^2} = \frac{z_{\alpha/2}^2\sigma^2}{t^2 - z_{\alpha/2}^2} = \frac{\sigma^2}{t^2/z_{\alpha/2}^2 - 1}.$$

B. Proof of equation (5)

We have

$$t^2_{\text{Box}} = \frac{\hat{\theta}_r^2}{\tau^2+\sigma_r^2} = \frac{\hat{\theta}_r^2}{\sigma_r^2}\left(\frac{c}{t_o^2/z^2_{\alpha/2}-1}+1\right)^{-1} = t_r^2\left(\frac{t_o^2/z^2_{\alpha/2}-1}{c+t_o^2/z^2_{\alpha/2}-1}\right),$$

so the requirement for replication success $t^2_{\text{Box}} \geq z^2_{\alpha/2}$ is equivalent to

$$\frac{t_r^2}{z^2_{\alpha/2}}\left(\frac{t_o^2}{z^2_{\alpha/2}}-1\right) \geq c + t_o^2/z^2_{\alpha/2} - 1.$$

Subtracting $t_o^2/z^2_{\alpha/2}-1$ on both sides leads to (5).

C. The limiting cases $\sigma_o^2 \downarrow 0$ and $\sigma_r^2 \downarrow 0$

Equation (7) can be re-written as

$$\frac{c-1}{t_A^2} z_S^4 + 2\, z_S^2 = t_H^2, \tag{14}$$

where

$$\frac{c-1}{t_A^2} = \frac{\sigma_o^2-\sigma_r^2}{\sigma_r^2}\frac{2}{t_o^2+t_r^2} = \frac{2\,\sigma_o^2(\sigma_o^2-\sigma_r^2)}{\hat{\theta}_o^2\sigma_r^2+\hat{\theta}_r^2\sigma_o^2}.$$

For $\sigma_o^2 \downarrow 0$ we thus have $(c-1)/t_A^2 \to 0$ and $t_H^2 \to 2\,t_r^2$ so equation (14) reduces to $2\,z_S^2 = 2\,t_r^2$ and hence $z_S^2 = t_r^2$. For $\sigma_r^2 \downarrow 0$ we have $(c-1)/t_A^2 \to (2\,\sigma_o^2)/\hat{\theta}_r^2$ and $t_H^2 \to 2\,t_o^2$ so equation (14) reduces to $(\sigma_o^2/\hat{\theta}_r^2)\, z_S^4 + z_S^2 = t_o^2$. The solution of this equation is with $d = \hat{\theta}_r^2/\hat{\theta}_o^2$:

$$\begin{aligned}
z_S^2 &= \frac{\hat{\theta}_r^2}{2\,\sigma_o^2}\left\{\sqrt{1+4\sigma_o^2 t_o^2/\hat{\theta}_r^2}-1\right\} \\
&= \frac{\hat{\theta}_r^2}{2\,\sigma_o^2}\left\{\sqrt{1+4/d}-1\right\} \\
&= \frac{\hat{\theta}_o^2}{2\,\sigma_o^2}\left\{\sqrt{d^2+4\,d}-d\right\} \\
&= \frac{t_o^2}{2}\left\{\sqrt{d(d+4)}-d\right\},
\end{aligned}$$

which is equation (9).

D. Proof of result (10)

Equation (8) reduces for $c \downarrow 0$ to

$$z_S^2 = t_A^2 - \sqrt{t_A^2\left[t_A^2 - t_H^2\right]} = t_A^2 - \sqrt{\frac{t_A^2}{2}\frac{(t_o^2 - t_r^2)^2}{t_o^2 + t_r^2}} = t_A^2 - \frac{\left|t_o^2 - t_r^2\right|}{2} = \min\{t_o^2, t_r^2\}.$$

The derivative of (8) with respect to c is (for $c \neq 1$)

$$\begin{aligned}
\frac{d\, z_S^2}{d\, c} &= -\frac{1}{c-1}\left\{z_S^2 - \frac{1}{2}\frac{t_A^2 t_H^2}{(c-1)z_S^2 + t_A^2}\right\} \\
&= -\frac{z_S^2}{c-1}\left\{1 - \frac{1}{2}\frac{(c-1)z_S^2 + 2\,t_A^2}{(c-1)z_S^2 + t_A^2}\right\} \\
&= -\frac{1}{2}\,\frac{z_S^4}{(c-1)z_S^2 + t_A^2} \qquad (15)
\end{aligned}$$

where the middle line follows from (7) and the last line also holds for $c = 1$. It is easy to see from (8) that $(c-1)z_S^2 + t_A^2 > 0$ for all c, and therefore (15) is negative for all c.

E. Proof of results in Section 4

For notational simplicity I omit the conditioning on $\hat{\theta}_o$ in the following. Equation (13) implies a distribution on $t_r = \hat{\theta}_r/\sigma_r = \sqrt{n_r}\,\hat{\theta}_r/\kappa$,

$$t_r \sim \mathrm{N}\left(\sqrt{n_r}\,\frac{\hat{\theta}_o}{\kappa}, \frac{n_r + n_o}{n_o}\right),$$

so $t_r = \sqrt{(n_r + n_o)/n_o}\,\tilde{t}_r$ where

$$\tilde{t}_r \sim \mathrm{N}\left(\sqrt{\frac{n_r n_o}{n_r + n_o}}\,\frac{\hat{\theta}_o}{\kappa}, 1\right).$$

Therefore $t_r^2 = (n_r + n_o)/n_o \cdot \tilde{t}_r^2$ follows a scaled non-central χ^2-distribution with 1 degree of freedom, scaling factor $(n_r + n_o)/n_o$ and non-centrality parameter $\lambda = (n_r n_o)/(n_r + n_o) \cdot \hat{\theta}_o^2/\kappa^2$. Things simplify somewhat for a point prior $\theta = \hat{\theta}_o$ at the estimate from the original study. Then $\hat{\theta}_r \,|\, \hat{\theta}_o \sim \mathrm{N}\left(\hat{\theta}_o, \kappa^2/n_r\right)$ so $t_r \sim \mathrm{N}\left(\sqrt{n_r}\,\hat{\theta}_o/\kappa, 1\right)$. Therefore t_r^2 now follows a non-central χ^2-distribution with 1 degree of freedom and non-centrality parameter $\lambda = n_r\,\hat{\theta}_o^2/\kappa^2$.