The assessment of replication success based on relative effect size

Leonhard Held, Charlotte Micheloud and Samuel Pawel

Epidemiology, Biostatistics and Prevention Institute (EBPI)

and Center for Reproducible Science (CRS)

University of Zurich

Hirschengraben 84, 8001 Zurich, Switzerland

Email: leonhard.held@uzh.ch

21st September 2020

Abstract: Replication studies are increasingly conducted to confirm original findings. However, there is no established standard how to assess replication success and in practice many different approaches are used. The purpose of this paper is to refine and extend a recently proposed reverse-Bayes approach for the analysis of replication studies. We show how this method is directly related to the relative effect size, the ratio of the replication to the original effect estimate. This perspective leads to two important contributions: the golden level to recalibrate the assessment of replication success and a novel approach to calculate the replication sample size based on the specification of the minimum relative effect size. Compared to the standard approach to require statistical significance of both the original and replication study, replication success at the golden level offers uniform gains in project power and controls the Type-I error rate even if the replication sample size is slightly smaller than the original one. Sample size calculation based on replication success at the golden level tends to require smaller samples than the standard approach, if the original study is reasonably powered. An application to data from four large replication projects shows that the replication success approach leads to more appropriate inferences, as it penalizes shrinkage of the replication estimate compared to the original one, while ensuring that both effect estimates are sufficiently convincing on their own.

Key Words: Power; Relative Effect Size; Replication Studies; Sample Size; Sceptical p-value; Two-Trials Rule; Type-I error rate

1 Introduction

Replication studies are conducted in order to investigate whether an original finding can be confirmed in an independent study. Although replication has long been a central part of the scientific method in many fields, the so-called replication crisis (Ioannidis, 2005; Begley and Ioannidis, 2015) has led to increased interest in replication over the last decade. These developments eventually culminated in large-scale replication projects that were conducted in various fields (Errington et al., 2014; Open Science Collaboration, 2015; Camerer et al., 2016, 2018; Cova et al., 2018).

Declaring a replication as successful is, however, not a straightforward task, and currently used approaches include significance of both the original and replication studies, compatibility of their effect estimates, and meta-analysis of the effect estimates. In order to address this lack of an accepted definition of replicability, a new method has recently been proposed in Held (2020a). The approach combines the analysis of credibility (Matthews, 2001a,b) with a prior-data conflict assessment (Box, 1980) to define replication success. Conceptually, replication success is declared if the replication study is in conflict with a sceptical prior that would make the original study non-significant.

To introduce some notation, let $z_o = \hat{\theta}_o/\sigma_o$ and $z_r = \hat{\theta}_r/\sigma_r$ denote the z-statistic of the original and replication study, respectively. Here $\hat{\theta}_o$ and $\hat{\theta}_r$ are the corresponding effect estimates (assumed to be normally distributed) of the unknown effect θ with standard errors σ_o and σ_r, respectively. The corresponding one-sided p-values are denoted by $p_o = 1 - \Phi(z_o)$ and $p_r = 1 - \Phi(z_r)$, respectively, where $\Phi(\cdot)$ denotes the standard normal cumulative distribution function. Let $c = \sigma_o^2/\sigma_r^2$ denote the variance ratio of the squared standard errors of the original and replication effect estimates. Usually, the squared standard errors are 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. The variance ratio c can then be identified as the relative sample size $c = n_r/n_o$. The relative effect

size

$$d = \frac{\hat{\theta}_r}{\hat{\theta}_o} = \frac{1}{\sqrt{c}} \frac{z_r}{z_o}$$

quantifies the size of the replication effect estimate $\hat{\theta}_r$ relative to the original effect estimate $\hat{\theta}_o$. The corresponding shrinkage of the replication effect estimate will be denoted as $s = 1 - d$.

Suppose the original study achieved statistical significance at one-sided level α, so $p_o \leq \alpha$. The standard approach to assess replication success is based on significance of the replication effect estimate at the same level α, *i. e.* the replication is considered successful if also $p_r \leq \alpha$. This approach is known in drug development as the two-trials rule (Senn, 2007). Let $z_\alpha = \Phi^{-1}(1 - \alpha) > 0$ denote the z-value corresponding to the level α, then significance of the replication effect estimate is equivalent to the criterion

$$d \geq \frac{z_\alpha}{z_o \sqrt{c}}. \tag{1}$$

The right-hand side goes to zero for increasing c, so if the relative sample size c is large enough, significance of the replication study can be achieved with any arbitrarily small (but positive) relative effect size d.

In this paper we relate the approach by Held (2020a) to the relative effect size, propose the golden level, a recalibration for the assessment of replication success, and show that it leads to a more appropriate criterion to achieve replication success than the two-trials rule. Moreover, replication success at the golden level offers uniform gains in project power and controls the Type-I error rate even if the replication sample size is slightly smaller than the original one. We also propose a novel sample size calculation approach based on the specification of the minimum relative effect size. This approach tends to require smaller replication samples, if the original study is reasonably powered.

The paper is structured as follows. Section 2 reviews the Held (2020a) approach to

assess replication success and relates it to the relative effect size in Section 2.1. This leads to a novel recalibration of the procedure based on the golden level (Section 2.2). A comparison with the two-trials rule is made in Section 2.3. Section 3 introduces different approaches to design the replication study based on the result from the original study. Specifically, the new approach based on the relative effect size (Section 3.1) is compared to more traditional approaches based on the power to achieve significance of the replication effect estimate respectively replication success (Section 3.2). Type-I error rates and project power of the proposed procedure are compared with the two-trials rule in Section 4. Section 5 describes an application to data from four different replication projects and Section 6 closes with some discussion.

2 Replication success

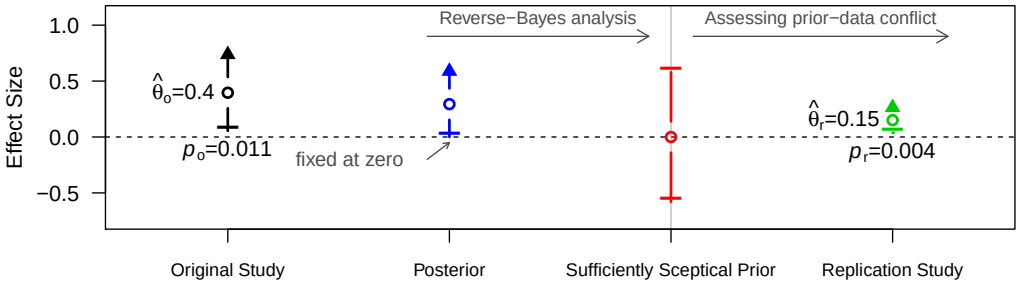

Figure 1: Example of the assessment of replication success. The original study from Pyc and Rawson (2010) has effect estimate $\hat{\theta}_o = 0.4$ on Fisher's z scale (95% CI from 0.05 to 0.74) and one-sided p-value $p_o = 0.011$. The left part of the figure illustrates the reverse-Bayes derivation of the sufficiently sceptical prior based on the original study result and the posterior with lower credible limit fixed at zero. The comparison of the sufficiently sceptical prior with the replication study result ($\hat{\theta}_r = 0.15$, 95% CI from 0.04 to 0.26, $p_r = 0.004$) in the right part of the figure is used to assess potential prior-data conflict.

Hereinafter we focus on the one-sided assessment of replication success to ensure that replication success can only occur if the original and replication effect estimates

go in the same direction. Figure 1 illustrates the Held (2020a) approach based on a replication study from the *Social Sciences Replication Project* (Camerer et al., 2018): the significant original finding by Pyc and Rawson (2010) at one-sided level $\alpha = 0.025$ is challenged with a sceptical prior, sufficiently concentrated around zero to make the original study result no longer convincing (Matthews, 2001a,b). Replication success is then defined as conflict between the sceptical prior and the result from the replication study in order to persuade the sceptic. Conflict is quantified by a prior-predictive tail probability p_{Box} (Box, 1980) where a small value $p_{\text{Box}} \leq \alpha$ defines replication success. In Figure 1 the original finding is only borderline significant, so the sufficiently sceptical prior is fairly wide. Furthermore, there is substantial shrinkage (62%) of the replication effect estimate and therefore hardly any conflict with the sufficiently sceptical prior (one-sided $p_{\text{Box}} = 0.31$). We are thus not able to declare replication success at level 2.5%.

Held (2020a) showed that if both $\text{sign}(z_o) = \text{sign}(z_r)$ and

$$\left(z_o^2/z_{\alpha_S}^2 - 1\right)\left(z_r^2/z_{\alpha_S}^2 - 1\right) \geq c \tag{2}$$

hold, replication success at the one-sided level α_S is achieved, where $z_{\alpha_S} = \Phi^{-1}(1 - \alpha_S)$. The requirement (2) can be assessed for different values of the level α_S and of particular interest is the smallest possible value of α_S where (2) holds, the so-called *sceptical p-value* p_S. We are thus interested in the value z_S^2 that fulfills

$$\left(z_o^2/z_S^2 - 1\right)\left(z_r^2/z_S^2 - 1\right) = c. \tag{3}$$

There is a unique solution of (3) which defines the one-sided sceptical p-value $p_S = 1 - \Phi(z_S)$ where $z_S := +\sqrt{z_S^2}$, provided $\text{sign}(z_o) = \text{sign}(z_r)$ holds. Replication success at level α_S is then achieved if $p_S \leq \alpha_S$. In the introductory example based on the original study by Pyc and Rawson (2010), the sceptical p-value turns out to be $p_S = 0.11$.

The sceptical p-value has a number of interesting properties, see Held (2020a, Section 3.1) for details. In particular,

$$p_S > \max\{p_o, p_r\} \tag{4}$$

always holds with $p_S \downarrow \max\{p_o, p_r\}$ for $c \downarrow 0$. Furthermore, if the p-values p_o and p_r are fixed, the sceptical p-value p_S increases with decreasing relative effect size d. The first property ensures that both the original and the replication study have to be sufficiently convincing on their own to ensure replication success. The second property guarantees that shrinkage of the replication effect estimate is penalized.

The level for replication success α_S has to be distinguished from the significance level α associated with the ordinary p-value. Held (2020a) has used the *nominal level* for replication success ($\alpha_S = \alpha$) for convenience, but in the following we will propose a recalibration of the procedure along with a new value for α_S, the *golden level*. The derivation is based on a natural requirement on the relative effect size (Section 2.2). In a nutshell, the golden level ensures that for original studies which were only borderline significant, replication success is only possible if the replication effect estimate is larger than the original one.

2.1 Relative effect size

Without loss of generality we assume that $\hat{\theta}_o > 0$ and that $p_o < \alpha_S$ has been observed in the original study, otherwise it would be impossible to achieve replication success at level α_S because p_S is always larger than p_o due to (4). The condition (2) for replication success can then be re-written as

$$z_r \geq z_{\alpha_S}\sqrt{1 + c/(K-1)} =: z_r^{\min}, \tag{5}$$

where $K = z_o^2/z_{\alpha_S}^2 > 1$. The right hand-side of (5) is the minimum replication z-value $z_r^{\min}$ required to achieve replication success. Note that the minimum replication z-value increases with increasing c, so increasing the sample size leads to a more stringent success requirement for the replication p-value. Equation (5) can be further transformed to a condition on the relative effect size

$$d \geq \frac{\sqrt{1 + c/(K-1)}}{\sqrt{K}\sqrt{c}} =: d_{\min}. \tag{6}$$

To achieve replication success, the relative effect size d must be at least as large as the right-hand side of (6), the *minimum relative effect size* $d_{\min}$, a function of K and the relative sample size c. Note that the minimum relative effect size simplifies to $d_{\min} = 1/\sqrt{K-1}$ for $c = 1$. If the relative sample size becomes very large, *i. e.* $c \to \infty$, we have $d_{\min} \downarrow d_\infty$ where

$$d_\infty = 1/\sqrt{K(K-1)}. \tag{7}$$

This shows that the minimum relative effect size in (6) does not go to zero for increasing c, so replication success cannot be achieved if the relative effect size d is smaller or equal to d_∞, no matter how large the replication study is. This has to be contrasted with the standard assessment of significance of the replication study, where the corresponding criterion (1) can be achieved for any positive relative effect size, regardless of how small, provided the replication sample size is sufficiently large.

2.2 The golden level

Significance of both the original and the replication study at level α is a necessary but not sufficient requirement for replication success at the nominal level ($\alpha_S = \alpha$). The nominal level may therefore be too stringent. It is more reasonable to calibrate the procedure in such a way that to establish replication success, original and replication study do not both necessarily need to be significant at level α, provided that the rep-

lication effect estimate does not shrink compared to the original one. We therefore choose a level α_S such that a borderline significant original study ($p_o = \alpha$) cannot lead to replication success if there is shrinkage $s > 0$ of the replication effect estimate. Mathematically, this translates to setting $d_\infty = 1$ and $K = z_\alpha^2 / z_{\alpha_S}^2$ in (7) and leads to the quadratic equation $K(K-1) = 1$ with solution $K = \varphi$ where $\varphi = (\sqrt{5}+1)/2 \approx 1.62$ is known as the golden ratio. Solving for z_{α_S} gives

$$z_{\alpha_S} \quad = \quad z_\alpha / \sqrt{\varphi} \tag{8}$$

and the corresponding *golden* level $\alpha_S = 1 - \Phi(z_{\alpha_S} = z_\alpha / \sqrt{\varphi})$ for replication success. This is our recommended default choice to assess replication success and we will study its properties in the following in more detail. For $z_\alpha = 1.96$ (one-sided $\alpha = 0.025$), the golden level is $\alpha_S = 0.062$. In the introductory example based on the replication of the Pyc and Rawson (2010) study, the sceptical p-value is $p_S = 0.11 > 0.062$, so the replication study was not successful at the golden level.

The condition $p_S \leq \alpha_S$ for replication success at the golden level is equivalent to $z_S \leq z_\alpha / \sqrt{\varphi}$, *i. e.* $z_S \sqrt{\varphi} \leq z_\alpha$. In practice it may be preferable to recalibrate the sceptical p-value $p_S = 1 - \Phi(z_S)$ to $\tilde{p}_S = 1 - \Phi(z_S \sqrt{\varphi})$, which then needs to be compared to α (rather than α_S) to assess replication success and can thus be interpreted on the same scale as an ordinary p-value. For example, the recalibrated sceptical p-value for the replication of Pyc and Rawson (2010) turns out to be $\tilde{p}_S = 0.061$.

2.3 Comparison with the two-trials rule

A useful benchmark for comparison is the two-trials rule in drug development (Kay, 2015, Section 9.4), which requires "at least two adequate and well-controlled studies, each convincing on its own, to establish effectiveness" (FDA, 1998, p. 3). This is usually achieved by independently replicating the result of a first study in a second study, both

significant at one-sided level $\alpha = 0.025$. It is worth noting that in practice the two trials are often run in parallel (Senn, 2007), so do not exactly resemble the replication setting.

The main difference between the replication success and the two-trials rule approach concerns how shrinkage of the replication effect estimate is handled. Figure 2 illustrates that shrinkage is penalized in the assessment of replication success, *i. e.* the original p-value needs to be quite small to achieve replication success for a relative effect size $d < 1$. In contrast, significance of the replication study can be achieved even if there is substantial shrinkage, provided the replication sample size is large enough.

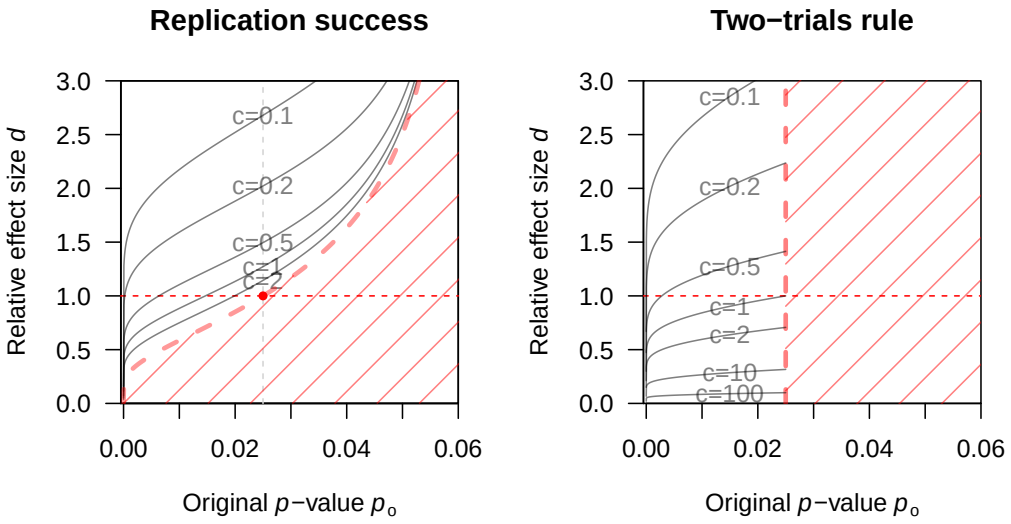

Figure 2: Comparison of replication success at the golden level ($p_S \leq \alpha_S = 0.062$) and the two-trials rule ($p_o \leq 0.025$ and $p_r \leq 0.025$). Red hatched areas indicate that success is impossible for original p-value p_o and relative effect size d. In the white areas success is possible and depends on the relative sample size c as indicated by the grey lines.

It is interesting to directly compare the two-trials rule and replication success at the golden level in terms of their stringency, *i. e.* the required relative effect size d to fulfill the criteria (1) and (6), respectively. If the original p-value is not significant at level α, only replication success can be achieved, but will require a replication effect estimate

larger than the original one. If the original p-value is smaller than α, then the situation depends on the relative sample size c. For example, when the replication sample size is chosen to be the same as in the original study ($c = 1$) and $\alpha = 0.025$, original studies with a p-value larger than 0.006 will require a smaller relative effect size d with the two-trials rule, while p-values smaller than 0.006 will require a smaller relative effect size d with the replication success method. This illustrates that the latter method is less stringent than the two-trials rule if the original study is already sufficiently convincing.

3 Design of the replication study

The methodology developed in the previous section is tailored for the analysis of replication studies, but can be adapted for the purpose of their design. In the following sample size calculation based on the minimum relative effect size is introduced and compared with the standard approach based on power.

3.1 Design based on relative effect size

Equation (6) can be inverted such that the replication sample size can be computed based on the specification of the minimum relative effect size. Specifically, the required relative sample size c to achieve replication success at level α_S with minimum relative effect size d_{min} is

$$c_{\text{RS}} = \begin{cases} \frac{K-1}{d_{\text{min}}^2 K(K-1)-1} & \text{if } d_{\text{min}} > d_\infty \\ \texttt{NA} & \text{else,} \end{cases} \tag{9}$$

where d_∞ is the bound (7) on the relative effect size. This novel way of calculating the sample size requires the specification of the minimum relative effect size d_{min} which can still be considered as acceptable. In practice, this will strongly depend on the field and the phenomenon studied by the researcher. We therefore do not recommend a default value, but rather that its specification should be based on domain knowledge.

The relative sample size (9) is available if $d_{\min} > d_\infty$, for $d_{\min} \downarrow d_\infty$ we have $c_{\text{RS}} \to \infty$. For an original study with $p_o = 0.025$ and the golden level $\alpha_S = 0.062$, the bound is $d_\infty = 1$. This ensures that for original studies that were only borderline significant, a necessary requirement for replication success is a replication effect estimate larger than the original one ($d_{\min} > 1$).

The two-trials rule can also be formulated in terms of the relative effect size. Re-arranging (1) gives the required relative sample size to fulfill the two-trials rule at level α with minimum relative effect size $d_{\min}$:

$$c_{\text{2TR}} = \begin{cases} \frac{z_\alpha^2}{d_{\min}^2 z_o^2} & \text{if } z_o \geq z_\alpha \\ \texttt{NA} & \text{else.} \end{cases} \tag{10}$$

If both (9) and (10) are available we have $K = z_o^2/z_{\alpha_S}^2 = \varphi\, z_o^2/z_\alpha^2$ at the golden level, so $c_{\text{2TR}} = \varphi/(d_{\min}^2 K)$ and hence the sample size ratio is

$$c_{\text{RS}}/c_{\text{2TR}} = \frac{\varphi - 1}{1 - (d_\infty/d_{\min})^2}. \tag{11}$$

Now $\varphi - 1 \approx 0.62$, so the relative sample size c_{RS} can be up to 38% smaller than c_{2TR} if $d_\infty/d_{\min}$ is small. If $d_\infty/d_{\min} > \varphi - 1$, then the replication success method at the golden level requires a larger sample size than the two-trials rule. This can be re-written with (7) to $K < 1/2 + \sqrt{\varphi^2/d_{\min}^2 + 1/4}$. For $\alpha = 0.025$, this translates to $p_o > 0.011$ for $d_{\min} = 1$ and $p_o > 0.007$ for $d_{\min} = 0.8$, showing that the replication success approach requires a larger sample size than the two-trials rule only if the p-value of the original study is relatively large.

It is interesting to compute the distribution of the sample size ratio (11) under the null respectively alternative hypothesis. The bound d_∞ in (11) depends on K which in turn depends on z_o whose distribution is either $\text{N}(0, 1)$ (under the null hypothesis) or

$N(\mu, 1)$ (under the alternative hypothesis), here

$$\mu = \Phi^{-1}(1-\alpha) + \Phi^{-1}(1-\beta), \tag{12}$$

where α is the assumed significance level and $1-\beta$ the power to detect the assumed effect in the original study (Matthews, 2006, Section 3.3). The distribution of the sample size ratio (11) can then be computed based on the implied truncated normal distribution of z_o. Table 1 gives the median, 10% and 90% quantile under the null and under the alternative hypothesis (with power of 80% and 90%, respectively) for minimum relative effect size $d_{\min} = 0.8$, 0.9 and 1. Under the null hypothesis, the median sample size ratio is between 1.07 and 1.22, so slightly larger with (9) compared to (10). Under the alternative, however, the median sample size ratio is between 0.65 and 0.71. This illustrates that the replication success approach tends to require a smaller sample size than the two-trials rule, if the alternative hypothesis is true and the original study is reasonably powered.

	Minimum relative effect size $d_{\min}$		
Hypothesis	0.8	0.9	1
Null	1.22 [0.75, 5.22]	1.14 [0.73, 4.66]	1.07 [0.71, 4.20]
Alternative (80% power)	0.71 [0.64, 1.35]	0.69 [0.63, 1.22]	0.68 [0.63, 1.12]
Alternative (90% power)	0.67 [0.63, 1.04]	0.66 [0.63, 0.96]	0.65 [0.63, 0.90]

Table 1: Quantiles of the distribution of the sample size ratio $c_{\text{RS}}/c_{\text{2TR}}$ (median with 10 and 90% quantile in brackets)

	Minimum relative effect size $d_{\min}$		
Hypothesis	0.8	0.9	1
Null	0.018	0.021	0.025
Alternative (80% power)	0.76	0.78	0.80
Alternative (90% power)	0.87	0.89	0.90

Table 2: Probability that the condition (13) is fulfilled

However, we need to assume that both c_{RS} and c_{2TR} exist, which is the case if both $d_{\min} > d_\infty$ and $z_o \geq z_\alpha$ holds. For $d_{\min} \leq 1$ this reduces to the condition

$$z_o > \frac{z_\alpha}{\sqrt{\varphi}} \sqrt{0.5 + \sqrt{1/4 + 1/d_{\min}^2}}. \tag{13}$$

Table 2 gives the probability that condition (13) holds. This has been calculated with Monte Carlo simulation. For $d_{\min} = 1$, the right-hand side of (13) reduces to z_α, so this probability is equal to the significance level respectively the assumed power. For $d_{\min} < 1$ the probability slightly decreases, for example under the null hypothesis and for $d_{\min} = 0.8$ the probability is 0.018 rather than 0.025.

3.2 Design based on power

The sample size of the replication study is usually calculated based on the estimate from the original study and a certain level of predictive or conditional power, depending on whether or not the uncertainty of the estimate is taken into account. Given that there is an effect, the power for *replication success* is the probability of replication success ($p_S \leq \alpha_S$) (Held, 2020a, Section 4) while the power for *significance* is the probability of a significant replication study ($p_r \leq \alpha$). The latter equals the power of the two-trials rule (2TR) provided the original study was significant ($p_o \leq \alpha$), otherwise the power is zero.

Figure 3 compares the power for replication success at the golden and at the nominal level with the power of the two-trials rule for relative sample size $c = 1$ (left) and $c = 5$ (right) as a function of the one-sided p-value from the original study p_o. Note that the predictive power is always closer to 50% due to incorporation of additional uncertainty (Held, 2020a; Micheloud and Held, 2020). Furthermore, the two-trials rule requires a significant original study and hence it is impossible to power a replication study when $p_o > 0.025$. For a replication study of the same size as the original study ($c = 1$),

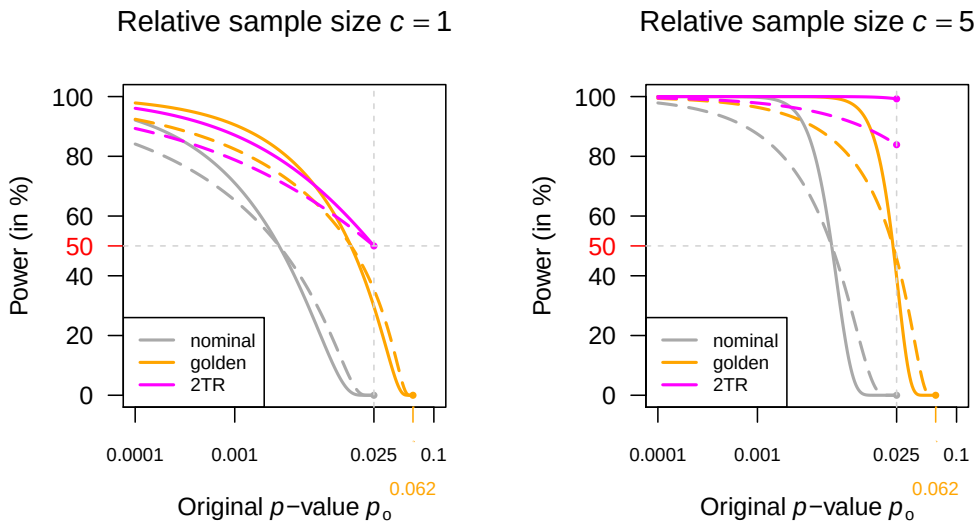

Figure 3: Power calculations for a replication study with sample size equal to the original study ($c = 1$, left) and increased by a factor of $c = 5$ (right). Shown is conditional (solid) and predictive (dashed) power for the two-trials rule (2TR) at level $\alpha = 0.025$ and for replication success (RS) at the corresponding nominal and golden level as a function of the one-sided p-value of the original study. Power values of exactly zero are omitted.

the probability of a significant replication result is only 50% if the original study is borderline significant (Goodman, 1992). For $c = 5$ the conditional power of the two-trials rule is 99% and the predictive power 84%. For very large c ($c \to \infty$) conditional power is 100%, independently of the original study result, whereas predictive power cannot be larger than $1 - p_o$ (Micheloud and Held, 2020). The power for replication success at the golden level is larger than the power of the two-trials rule if the original p-value is sufficiently small. Power curves for replication success at the nominal level are always smaller than for the golden level and approach zero for $p_o \uparrow 0.025$, whereas the power at the golden level is positive for any $p_o < 0.062$.

4 Error rates

Although Bayesian methods do not rely on the frequentist paradigm of repeated testing, it is still useful to investigate the frequentist operating characteristics of Bayesian methods (Dawid, 1982; Rubin, 1984; Grieve, 2016), and this also holds for the proposed reverse-Bayes assessment of replication success. In the following we assume that none of the two studies have been conducted and investigate the Type-I error rate and the project power (Maca et al., 2002).

4.1 Type-I error rate

The Type-I error rate of the two-trials rule is simply α^2 for any value of the relative effect size c. In contrast, the Type-I error rate of the proposed replication success assessment depends on c. For $c = 1$, Held (2020a, Section 3) showed that z_S^2 in (3) simplifies to half the harmonic mean of the squared test statistics z_o^2 and z_r^2. The connection $z_S^2 = z_H^2/4$ to the harmonic mean χ^2 test statistic z_H^2 (Held, 2020b), which has a $\chi^2(1)$-distribution under the null hypothesis, makes it straightforward to compute the Type-I error rate for $c = 1$ at level α_S:

$$\text{T1E} = \left\{1 - \Phi\left[2\,\Phi^{-1}\left(1 - \alpha_S\right)\right]\right\}/2. \tag{14}$$

For the golden level $\alpha_S = 0.062$ at $\alpha = 0.025$, the Type-I error rate (14) is 0.0515%, only slightly less than the Type-I error rate $\alpha^2 = 0.0625\%$ of the two-trials rule. For comparison, the Type-I error rate at the nominal level $\alpha_S = 0.025$ is 0.0022%, so considerably smaller than 0.0625%.

For $c \neq 1$, the Type-I error rate can be calculated through numerical integration:

$$\text{T1E} = \int_{z_{\alpha_S}}^{\infty} \Pr(z_r \geq z_r^{\min} \mid z_o, c, \alpha_S)\, \phi(z_o)\, dz_o, \tag{15}$$

where $\phi(\cdot)$ denotes the standard normal density function. The first term in the integral of (15) is the probability of replication success at level α_S conditional on a fixed original test statistic z_o and a relative sample size c. This can be computed with equation (5) as $\text{Pr}(z_r \geq z_r^{\min} \,|\, z_o, c, \alpha_S) = 1 - \Phi(z_r^{\min})$ because $z_r \sim \text{N}(0, 1)$ under the null hypothesis.

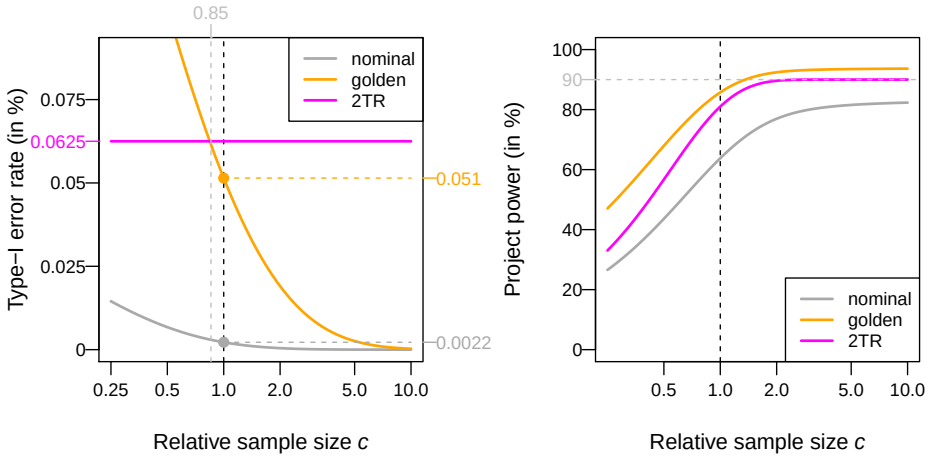

Figure 4: Type-I error rate (left) and project power (right) for fixed relative sample size c. Results are given for replication success (RS) at the nominal and golden level and compared with the two-trials rule (2TR) at $\alpha = 0.025$. The power of the original study is 90%

The left plot in Figure 4 displays the Type-I error rate for $\alpha = 0.025$ as a function of the relative sample size c. It can be seen that the Type-I error decreases with increasing relative sample size c. Due to (4), the Type-I error rate of the nominal level is always below the target 0.0625%. Although the Type-I error will eventually attain α^2 in the limit $c \downarrow 0$ (Held, 2020a, Section 3.4), the nominal level seems to be too stringent for realistic values of c. The Type-I error rate of the golden level is smaller than 0.0625% for $c > 0.85$. Appropriate Type-I error control is thus ensured even for replication studies where the sample size is slightly smaller than in the original study.

4.2 Project power

Under the alternative we have $z_o \sim \mathrm{N}(\mu, 1)$ where μ, as given in (12), depends on the assumed significance level α and the power $1-\beta$ to detect the assumed effect in the original study with a standard significance test (Matthews, 2006). In the following $\alpha = 0.025$ and $1-\beta = 90\%$ are used. The power of a significant replication study is $\Phi(\sqrt{c}\mu - z_\alpha)$, so depends on both μ and the relative sample size c. The project power of the two-trials rule is therefore $(1-\beta)\Phi(\sqrt{c}\mu - z_\alpha)$.

The project power for replication success is computed as

$$\mathrm{PP} = \int_{z_{\alpha_S}}^{\infty} \mathsf{Pr}(z_r \geq z_r^{\min} \,|\, z_o, c, \alpha_S)\, \phi(z_o - \mu)\, dz_o, \tag{16}$$

again with μ as given in (12), and shown in the right plot of Figure 12 as a function of c. For the golden level, the project power quickly increases to values above 90%, whereas the nominal level only reaches around 80% project power. The project power based on the two-trials rule is shown for comparison, which is always smaller than for the golden rule and converges to 90% for large c.

4.3 Adaptive designs

In practice the sample size of the replication study is usually chosen based on the result of the original study. We may distinguish four different methods. The first two calculate the relative sample size based on the minimum relative effect size $d_{\min}$ using formula (9) or (10), see Section 3.1. The other two calculate the relative sample size based on the power to achieve replication success or to fulfill the two-trials rule, see Section 3.2. As before we distinguish whether the analysis is based on replication success or the two-trials rule.

Calculation of Type-I error and project power can still be done with (15) respectively (16), noting that the relative sample size c in $\mathsf{Pr}(z_r \geq z_r^{\min} \,|\, z_o, c, \alpha_S)$ now depends on z_o.

The results are given in Supplementary Material and can be summarized as follows: Replication success in the analysis provides Type-I error control at the $\alpha^2 = 0.025^2$ level of the two-trials rule, if the replication study is powered to detect the effect from the original study with power 60% or larger. This holds both for design based on replication success and the two-trials rule. Likewise, an analysis based on replication success controls the Type-I error for any minimum relative effect size $d_{\min} < 1.04$ specified in the design based on replication success as described in Section 3.1. If design is based on the two-trials rule, then the relative effect size has to be slightly smaller ($d_{\min} < 0.97$).

Turning to project power, the uniformly better performance of replication success over the two-trials rule for fixed design shown in Figure 4 directly transfers to the adaptive designs with project power based on replication success in the analysis being always larger than project power based on the two-trials rule. Not surprisingly, design methods which require a smaller replication sample size lead to a smaller project power. This is the case if the design was based on the relative effect size for replication success rather than for the two-trials rule, where the median replication sample size of the former is considerably smaller than the latter, see Table 1.

5 Application

In this section, we illustrate the proposed methodology using data from four replication projects. All four projects reported effect estimates that were transformed to correlation coefficients (r). This scale allows for easy comparison of effect estimates from studies that investigate different phenomena, as it is bounded to the interval between minus one and one. Moreover, the Fisher z-transformation ($\hat{\theta} = \tanh^{-1}(r)$) can be applied to the correlation coefficients, resulting in the transformed estimates being asymptotically normal with variance which is only a function of the study sample

size n, *i. e.* $\text{Var}(\hat{\theta}) = 1/(n-3)$ (Fisher, 1921).

The first data set comprises the results from the *Reproducibility Project: Psychology* (Open Science Collaboration, 2015), whose aim was to replicate 100 studies, all of which were published in three major Psychology journals in 2008. For our purpose only the 73 study pairs from the "meta-analytic" subset are considered, since only for these studies the standard error of the Fisher z-transformed effect estimates can be computed (Johnson et al., 2016). The second data set comes from the *Experimental Economics Replication Project* (Camerer et al., 2016) which attempted to replicate 18 experimental economics studies published in two high impact economics journals between 2011 and 2015. The third data stem from the *Social Sciences Replication Project* (Camerer et al., 2018) where 21 replications of studies on the social sciences were carried out, all of which were originally published in the journals *Nature* and *Science* between 2010 and 2015. The last data set originates from the *Experimental Philosophy Replicability Project* (Cova et al., 2018) which involved 40 replications of studies from the emerging field of experimental philosophy. Since only for 31 studies effective sample size for original and replication study were available simultaneously, only these pairs were included. For more information on the data sets see also Pawel and Held (2020).

5.1 Analysis

Table 3 presents overall results for each of the replication projects. While the median relative effect size is below one for all of the four projects, there are still large differences. For example, the median relative effect size is only 0.29 in the psychology project, whereas it is 0.86 in the philosophy project. The degree of shrinkage is also reflected in the success rates (according to the two trials rule and the replication success approach), which are around 30% for the former and more than 70% for the latter. The percentage of successful replications is similar for the two-trials rule and the replication success approach, but there are discrepancies in all projects expect for

Experimental Economics.

Table 3: Results for each replication project: Relative effect size d (median with 25% and 75% quantiles on Fisher's z scale), percentage of successful replications with the two-trials rule (2TR) and the replication success (RS) approach (at the golden level), and number of studies where the methods disagree.

Project	relative effect size d	2TR (%)	RS (%)	discrepant
Psychology	0.29 [0.03, 0.77]	28.8	30.1	3/73
Social Sciences	0.52 [0.13, 0.65]	61.9	52.4	2/21
Experimental Philosophy	0.86 [0.47, 1.12]	74.2	71.0	1/31
Experimental Economics	0.67 [0.35, 0.92]	55.6	55.6	0/18

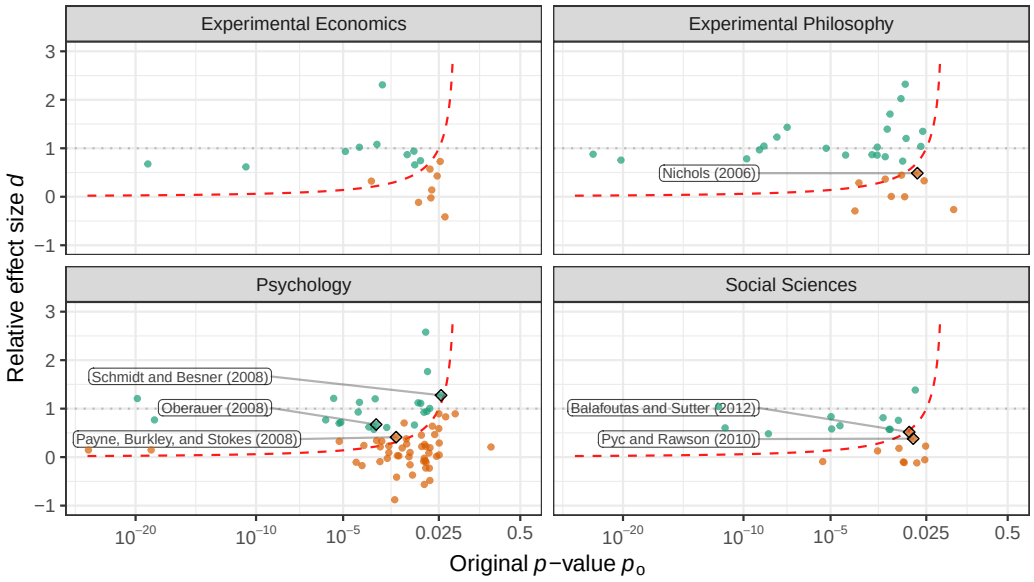

Figure 5: Green indicates that replication success was achieved at the golden level while orange indicates that it was not. The diamonds mark studies where the replication success approach and the two-trials rule disagree. The dashed red line indicates the bound below which replication success is impossible at the golden level with $\alpha = 0.025$.

Figure 5 displays the relative effect size d versus the original p-value p_o for each study pair and stratified by project. Note that one study pair from the philosophy project is not shown due to extremely small original p-value and another study pair

from the psychology project is not shown due to a very large relative effect size. We can see that for most of the study pairs, the replication success approach and the two-trials rule lead to the same conclusion, only six replications show conflicting results. They are highlighted with diamonds in Figure 5 and their characteristics are summarised in Table 4. Two studies from the psychology project show replication success but fail the

Table 4: Characteristics of studies for which the replication success approach (at the golden level) and the two-trials rule disagree (at one-sided $\alpha = 0.025$). Shown are relative sample size c, relative effect size d, original, replication and recalibrated sceptical p-value p_o, p_r and $\tilde{p}_S$.

Study	Project	c	d	p_o	p_r	$\tilde{p}_S$
Schmidt and Besner (2008)	Psychology	2.58	1.28	**0.028**	< 0.0001	0.024
Oberauer (2008)	Psychology	0.60	0.67	0.0003	**0.035**	0.017
Payne et al. (2008)	Psychology	2.65	0.41	0.001	0.023	**0.031**
Balafoutas and Sutter (2012)	Social Sciences	3.48	0.52	0.009	0.011	**0.04**
Pyc and Rawson (2010)	Social Sciences	9.18	0.38	0.011	0.004	**0.061**
Nichols (2006)	Experimental Philosophy	9.40	0.49	0.015	0.0006	**0.049**

two-trials rule. These studies show p-values that are slightly above the significance threshold in either original or replication study, but do not exhibit much shrinkage; In the replication of Oberauer (2008), the replication p-value was $p_r = 0.035$, a little too large to pass the two-trials rule. However, as the replication effect estimate shrunk only about 30% compared to the original one, replication success is still achieved. Conversely, the original p-value $p_o = 0.028$ in Schmidt and Besner (2008) was just above the significance level, yet the replication led to a highly significant result $p_r < 0.0001$ with the effect estimate being even 30% larger than the original counterpart, which therefore also resulted in replication success.

The remaining conflicting studies do not show replication success despite passing the two-trials rule. In all cases, there is substantial shrinkage of the replication effect estimate compared to the original one. For instance, in the replication study of Pyc and Rawson (2010), the estimate shrunk by 62% and the replication p-value was only significant because the sample size was increased by a factor of $c = 9.2$.

5.2 Design

Suppose that the four replication projects have not been conducted yet and we want to determine the sample size of the replication studies. The left plot in Figure 6 displays the percentage of original studies for which a finite replication sample size is available to achieve replication success (at the golden level with $\alpha = 0.025$) as a function of the required relative effect size d_{min} and stratified by replication project. If we require the replication effect estimate to be at least as large as the original one ($d_{\text{min}} = 1$), a finite replication sample size is available for all original studies which were significant at $\alpha = 0.025$. This is the case for all studies of the Social Science project, but not for the other three projects, where replication attempts have been made also for some non-significant original findings. For $d_{\text{min}} < 1$, the proportion decreases since the original studies have to be more convincing to achieve replication success. For $d_{\text{min}} = 0.8$ a finite replication sample size exists for around 85% of the studies. In the following, we will fix the minimum relative effect size d_{min} at this value.

The right plot in Figure 6 shows a comparison of the required relative sample size of the replication study based on the replication success approach with a minimum relative effect size of $d_{\text{min}} = 0.8$ to the two-trials rule with 80% power to detect the original effect estimate at the 2.5% significance level (one-sided). The violin densities and medians are based on studies for which both methods have a finite replication sample size. We can see that the relative sample size c is smaller for the replication success approach in the majority of the cases. Also the median relative sample size c is smaller for the replication success approach compared to the two-trials rule in all replication projects, illustrating that the replication success approach can lead to more efficient designs.

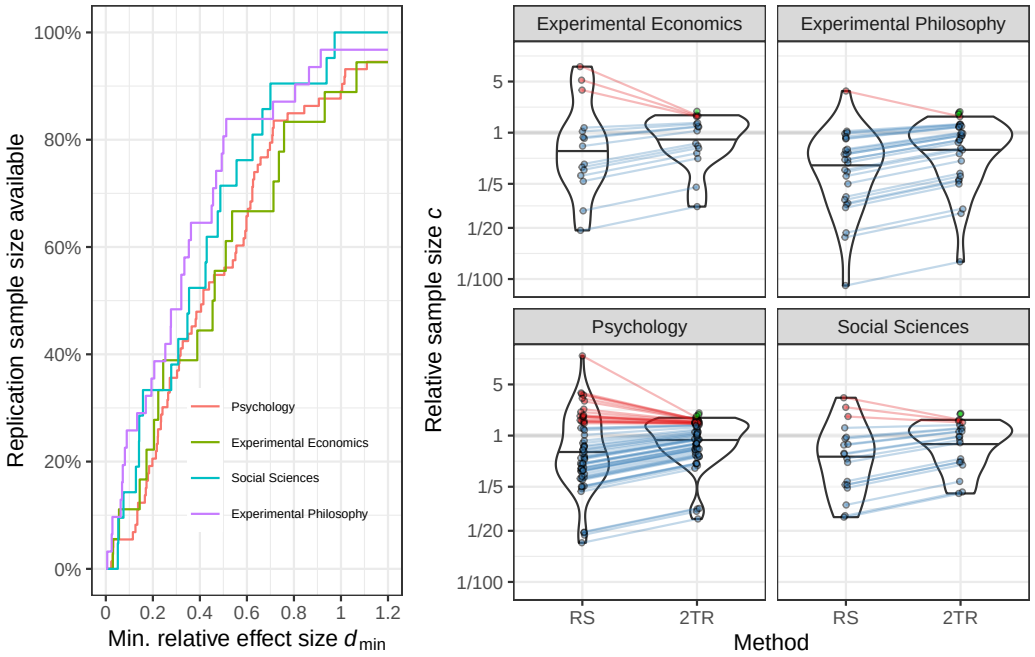

Figure 6: Percentage of original studies for which a finite replication sample size to achieve replication success (at the golden level with $\alpha = 0.025$) is available as a function of the relative effect size d (left). Comparison of the required relative sample size c determined with the replication success (RS) approach based on minimum relative effect size of $d_{\text{min}} = 0.8$ and the two trials rule (2TR) with 80% conditional power (right). Studies where the relative sample size c is smaller for the replication success approach are shown in blue, studies where it is larger in red. Green circles indicate original significant studies for which it is impossible to achieve replication success.

6 Discussion

In this paper, we have expanded on the replication success approach introduced in Held (2020a) and demonstrated its advantages over alternative methods such as the two-trials rule. In particular, the method provides an attractive compromise between hypothesis testing and estimation, as it penalizes shrinkage of the replication effect estimate compared to the original one, while ensuring that both are statistically significant to some extent. For instance, the method will indicate only a low degree of replication success when the replication study shows a much smaller but statistically significant significant effect estimate, whereas it can still indicate a large degree of success when either original or replication p-value are slightly above the significance level, provided their effect estimates are compatible.

We further refined the method by proposing the *golden level*, a new threshold for replication success. It guarantees that borderline significant original studies can only be replicated successfully if the replication effect estimate is larger than the original one. Compared to the two-trials rule, the golden level offers uniform gains in project power and controls the Type-I error rate if the replication sample size is not too small ($c > 0.85$). Empirical evaluation of data from four replication projects highlights that in most cases the methods are in agreement, however, for the study pairs where the approaches disagree, the replication success approach seems to lead to more sensible conclusions.

With this paper we further advanced the reverse-Bayes methodology for the analysis and design of replication studies, yet certain limitations and opportunities for future research remain: First, assuming normality of the effect estimates may be questionable, especially for small sample sizes, and more robust distributional assumptions could be considered. Second, in some types of analyses (*e. g.* regression or ANOVA) the effect estimate is a vector and the approach would need suitable adaptations. Finally, there is a recent trend to not only conduct one but several replications for one original study.

Also for this situation, the method would need to be adapted, *e.g.* the replication estimates could be first synthesized and an analysis of replication success could be performed subsequently.

Despite a lack of agreement as to which statistical method should be used to evaluate replication studies, conclusions based on different methods usually agree. Nevertheless, in some cases, classical methods such as the two-trials rule may produce anomalies. We argue that the replication success approach improves upon existing methods leading to more appropriate inferences and decisions that better reflect the available evidence.

Data and Software Availability Data analyzed in this article and software are available in the R-package `ReplicationSuccess`, which can be installed by running following command in an R console: `install.packages("ReplicationSuccess", repos= "http://R-Forge.R-project.org")`. Further information on data preprocessing can be found on the corresponding help page (with the command `?RProjects`).

Acknowledgments Support by the Swiss National Science Foundation (Project # 189295) is gratefully acknowledged.

References

Balafoutas, L. and Sutter, M. (2012). Affirmative action policies promote women and do not harm efficiency in the laboratory. *Science*, 335(6068):579–582.

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

Box, G. E. P. (1980). Sampling and Bayes' inference in scientific modelling and robust-

ness (with discussion). *Journal of the Royal Statistical Society, Series A*, 143:383–430. https://www.jstor.org/stable/2982063.

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.

Cova, F., Strickland, B., Abatista, A., Allard, A., Andow, J., Attie, M., Beebe, J., Berniūnas, R., Boudesseul, J., Colombo, M., Cushman, F., Diaz, R., N'Djaye Nikolai van Dongen, N., Dranseika, V., Earp, B. D., Torres, A. G., Hannikainen, I., Hernández-Conde, J. V., Hu, W., Jaquet, F., Khalifa, K., Kim, H., Kneer, M., Knobe, J., Kurthy, M., Lantian, A., Liao, S.-y., Machery, E., Moerenhout, T., Mott, C., Phelan, M., Phillips, J., Rambharose, N., Reuter, K., Romero, F., Sousa, P., Sprenger, J., Thalabard, E., Tobia, K., Viciana, H., Wilkenfeld, D., and Zhou, X. (2018). Estimating the reproducibility of experimental philosophy. *Review of Philosophy and Psychology*. https://doi.org/10.1007/s13164-018-0400-9.

Dawid, A. P. (1982). The well-calibrated Bayesian. *Journal of the American Statistical Association*, 77:605–610.

Errington, T. M., Iorns, E., Gunn, W., Tan, F. E., Lomax, J., and Nosek, B. A. (2014). An open investigation of the reproducibility of cancer biology research. *eLife*, 3.

FDA (1998). Providing clinical evidence of effectiveness for human drug and biological products. Technical report, US Food and Drug Administration. www.fda.gov/regulatory-information/search-fda-guidance-documents/providing-clinical-evidence-effectiveness-human-drug-and-biological-products.

Fisher, R. A. (1921). On the probable error of a coefficient of correlation deduced from a small sample. *Metron*, 1:3 – 32.

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.

Grieve, A. P. (2016). Idle thoughts of a 'well-calibrated' Bayesian in clinical drug development. *Pharmaceutical Statistics*, 15(2):96–108.

Held, L. (2020a). A new standard for the analysis and design of replication studies (with discussion). *Journal of the Royal Statistical Society, Series A*, 183:431–469.

Held, L. (2020b). The harmonic mean χ^2 test to substantiate scientific findings. *Journal of the Royal Statistical Society, Series C*. https://arxiv.org/abs/1911.10633.

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.

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.

Kay, R. (2015). *Statistical Thinking for Non-Statisticians in Drug Regulation*. John Wiley & Sons, Chichester, U.K., second edition.

Maca, J., Gallo, P., Branson, M., and Maurer, W. (2002). Reconsidering some aspects of the two-trials paradigm. *Journal of Biopharmaceutical Statistics*, 12(2):107–119.

Matthews, J. N. (2006). *Introduction to randomized controlled clinical trials*. Chapman and Hall/CRC. 10.1201/9781420011302.

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

Matthews, R. A. J. (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.

Micheloud, C. and Held, L. (2020). Power calculations for replication studies. Technical report. https://arxiv.org/abs/2004.10814.

Nichols, S. (2006). Folk intuitions on free will. *Journal of Cognition and Culture*, 6(1-2):57–86.

Oberauer, K. (2008). How to say no: Single- and dual-process theories of short-term recognition tested on negative probes. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 34(3):439–459.

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

Pawel, S. and Held, L. (2020). Probabilistic forecasting of replication studies. *PloS one*, 15(4):e0231416.

Payne, B. K., Burkley, M. A., and Stokes, M. B. (2008). Why do implicit and explicit attitude tests diverge? the role of structural fit. *Journal of Personality and Social Psychology*, 94(1):16–31.

Pyc, M. A. and Rawson, K. A. (2010). Why testing improves memory: Mediator effectiveness hypothesis. *Science*, 330(6002):335–335.

Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *The Annals of Statistics*, 12(4):1151–1172.

Schmidt, J. R. and Besner, D. (2008). The Stroop effect: Why proportion congruent has nothing to do with congruency and everything to do with contingency. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 34(3):514–523.

Senn, S. (2007). *Statistical Issues in Drug Development*. John Wiley & Sons, Chichester, U.K., second edition.

The assessment of replication success based on relative effect size

Supplement: Type-I error rate and project power of adaptive designs

Leonhard Held, Charlotte Micheloud and Samuel Pawel

21st September 2020

In practice the sample size of the replication study is usually chosen based on the result of the original study. We distinguish four different methods. The first two calculate the relative sample size based on the relative effect size d, either based on replication success or on the two-trials rule. The other two calculate the relative sample size based on power, again for replication success or the two-trials rule. We show results for both conditional and predictive power. Calculation of the Type-I error and the project power can be done with numerical integration in all cases.

First suppose that the sample size calculation is aiming to achieve replication success based on the specification of the relative effect size d. This is shown in Figure 1 (top left). It is re-assuring that any relative effect size smaller than 1.04 will maintain the Type-I error rate $\alpha^2 = 0.0625\%$ of the two-trials rule. We could also aim to fulfill the two-trials rule rather than replication success in the design of the replication study, see Figure 1 (top right). Type-I errors are slightly increased now with any relative effect size smaller than 0.97 maintaining the Type-I error rate of the two-trials rule.

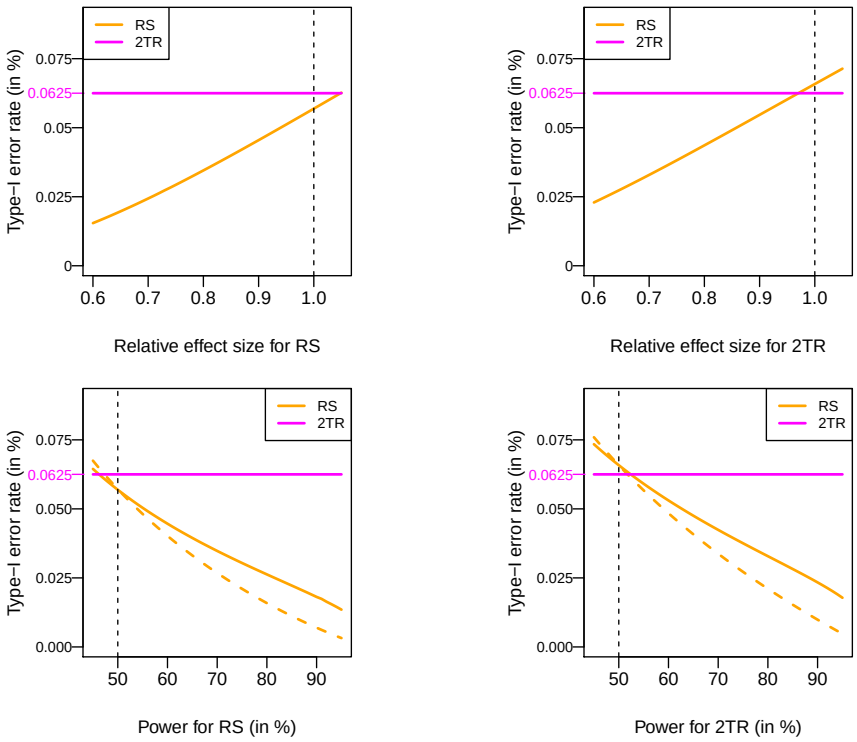

Figure 1: Type-I error rate for adaptive designs. In the top row the relative sample size was determined after the original study has been conducted based on specification of the relative effect size. In the bottow row the relative sample size was determined based on specification of the power (solid lines represent conditional, dashed lines predictive power). Results in the left column are based on design for replication success at the golden level at $\alpha = 0.025$. The right column is based on a design to achieve the two-trials-rule at $\alpha = 0.025$.

The bottom left plot of Figure 1 shows the Type-I error rate if the sample size was calculated based on a pre-specified power for replication success. The replication success approach controls the Type-I error for any conditional power larger than 47%. The Type-I error based on predictive power is slightly smaller for any power > 50%. Slightly larger values of the Type-I error rate are obtained if sample size calculation was based on the power to fulfill the two-trials rule (bottom right). The conditional version is the current standard in many replication projects and controls the Type-I

error for any power larger than 53%.

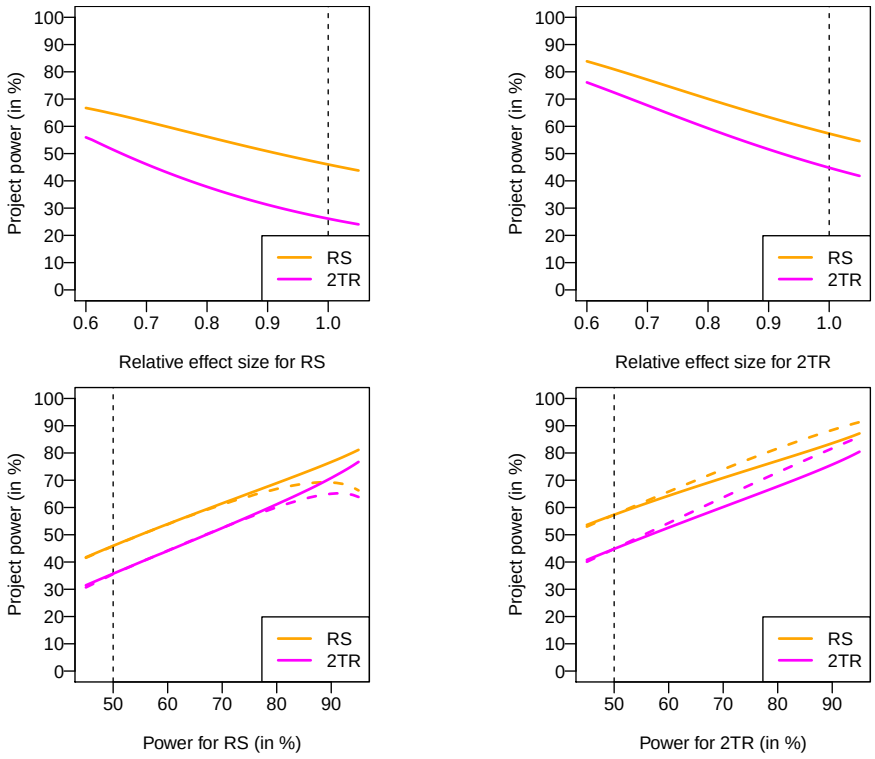

Figure 2: Project power for adaptive designs. In the top row the relative sample size was determined after the original study has been conducted based on specification of the relative effect size. In the bottow row the relative sample size was determined based on specification of the power (solid lines represent conditional, dashed lines predictive power). Results in the left column are based on design for replication success, in the right column based on design to achieve the two-trials-rule.

The four plots of Figure 2 show the project power for the different adaptive design. We now distinguish whether the replication success approach or the two-trials rule has been used in the design and in the analysis, respectively. The project power (top left plot) decreases with increasing relative effect size for replication success in the design (top left plot). Project power is 10-20% larger if the replication success approach rather than the two-trials rule is used in the analysis. A similar pattern can be seen if the

design was based on specification of relative effect size for the two-trials rule, where project power is in general increased due to the increased sample size.

The bottom left and right plots show the project power if the study was powered for replication success (left) respectively the two-trials rule (right). Project power based on replication success reaches values of around 70-80% if the power used in the sample size calculation was reasonably high (>80%). Values for predictive power are somewhat lower, in particular for large values of the power for replication success, where the probability that the replication sample size does not exist increases. A similar pattern can be seen if the design was based on the power to fulfill the two-trials rule, where the project power based on predictive power design is now slightly larger than for conditional power.