THE ASSESSMENT OF REPLICATION SUCCESS BASED ON RELATIVE EFFECT SIZE

BY LEONHARD HELD, CHARLOTTE MICHELOUD AND SAMUEL PAWEL

Epidemiology, Biostatistics and Prevention Institute, Center for Reproducible Science, University of Zurich,
leonhard.held@uzh.ch; charlotte.micheloud@uzh.ch; samuel.pawel@uzh.ch

Replication studies are increasingly conducted in order 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 a new proposal to recalibrate the assessment of replication success, the golden level. The recalibration ensures that for borderline significant original studies replication success can only be achieved if the replication effect estimate is larger than the original one. Conditional power for replication success can then take any desired value if the original study is significant and the replication sample size is large enough. 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 if the replication sample size is not smaller than the original one. An application to data from four large replication projects shows that the new 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.

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; Klein et al., 2014; Open Science Collaboration, 2015; Ebersole et al., 2016; Camerer et al., 2016, 2018; Cova et al., 2018; Klein et al., 2018).

Declaring a replication as successful is, however, not a straightforward task, and currently used approaches include statistical significance of both the original and replication study, compatibility of their effect estimates, and meta-analysis of the effect estimates. Many of the replication projects listed above also report the relative effect size, the ratio of the replication to the original effect estimate. For example, in Camerer et al. (2018) the replication effect estimates were only half as large as the original ones on average and even smaller in Open Science Collaboration (2015). This gives clear evidence of a systematic bias of the original studies and strongly suggests that the original and replication study should not be treated as exchangeable. However, all the approaches mentioned above will give the same results if the order of studies would be reversed.

In order to address this problem, 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). Replication success is declared if the replication study

---

is in conflict with a sceptical prior that would make the original study non-significant. This approach penalizes small relative effect sizes as we will see in more detail in the following.

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. The squared standard errors are 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. The variance ratio c can then be identified as the relative sample size $c = n_r/n_o$. The relative effect size

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

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), usually conducted at $\alpha = 0.025$. Let $z_\alpha = \Phi^{-1}(1 - \alpha) > 0$ denote the z-value corresponding to the level α, then significance of the replication study is achieved if $z_r \geq z_\alpha$, which is equivalent to the condition

$$(2) \qquad d \geq \frac{z_\alpha}{z_o \sqrt{c}}$$

on the relative effect size (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. However, declaring replication success when there is substantial shrinkage is contrary to common sense, as the replication effect estimate may not reflect an effect size of the same practical relevance as the original one, despite its statistical significance.

In this paper we first review the Held (2020a) approach for the assessment of replication success, followed by showing how it relates to the relative effect size (Section 2.1). This perspective is used in Section 2.2 and 2.3 to propose a recalibration of the method, the *golden level*, which leads to a more appropriate criterion for replication success compared to the two-trials rule (Section 2.4). In Section 3 we study power and Type-I error rates of the proposed method and compare it to the two-trials rule. The recalibrated method ensures that conditional power can take any desired value if the original study has been significant and the replication sample size is large enough (Section 3.1), controls the overall Type-I error if the replication sample size is not smaller than the original one (Section 3.2), and offers uniform gains in project power compared to the two-trials rule (Section 3.3). Section 4 describes an application to data from four replication projects and Section 5 closes with some discussion.
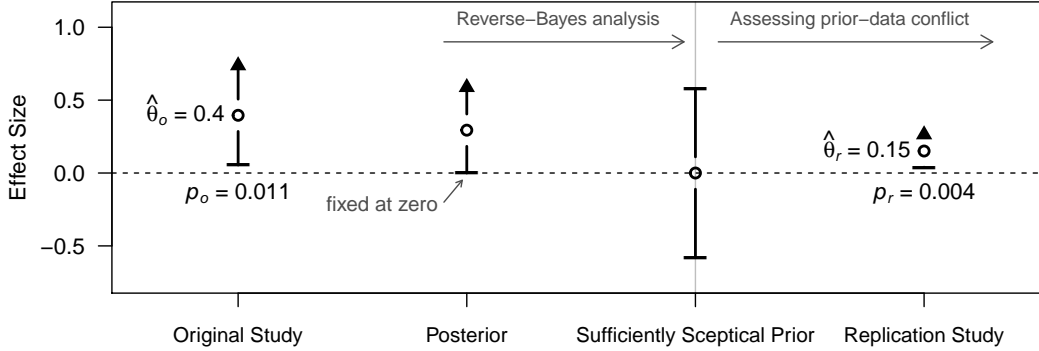
2. Replication success.

FIG 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 disprove 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 ($d = 0.15/0.4 = 0.38$) 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%.

The actual value of p_{Box} is difficult to interpret as it depends on the level α and does not even exist if the original p-value p_o exceeds α. However, Held (2020a) showed that if both $\text{sign}(z_o) = \text{sign}(z_r)$ and

$$(3) \qquad \left(z_o^2/z_{\alpha_S}^2 - 1\right)\left(z_r^2/z_{\alpha_S}^2 - 1\right) \geq c$$

hold, replication success at level α_S is achieved, where $z_{\alpha_S} = \Phi^{-1}(1 - \alpha_S)$. The requirement (3) can be assessed for any value of $\alpha_S > \max\{p_o, p_r\}$ and of particular interest is the smallest possible value of α_S where (3) holds, the so-called *sceptical p-value* p_S. We are thus interested in the value z_S^2 that fulfills

$$(4) \qquad \left(z_o^2/z_S^2 - 1\right)\left(z_r^2/z_S^2 - 1\right) = c.$$

There is a unique solution of (4) 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, $p_S > \max\{p_o, p_r\}$ 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 achieve 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* (Section 2.2). The derivation is based on a property of the required relative effect size for replication success, if the relative sample size is very large (Section 2.1). In a nutshell, the golden level ensures that for original studies which were only borderline significant ($p_o = \alpha$), replication success is only possible if the replication effect estimate is larger than the original one ($d > 1$).

2.1. *Relative effect size.* Without loss of generality we now 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. The condition (3) for replication success can then be re-written as

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

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 $z_r^{\min}$ increases with increasing c, so increasing the replication sample size leads to a more stringent success requirement for z_r and the corresponding replication p-value p_r.

Equation (5) can be further transformed to a condition on the relative effect size (1):

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

To achieve replication success, the relative effect size 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. If the relative sample size becomes very large, *i. e.* $c \to \infty$, we have $d_{\min} \downarrow d_\infty$ where

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

is the *limiting relative effect size*. This shows that the minimum relative effect size $d_{\min}$ 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. In contrast, the corresponding criterion (2) of the two-trials rule 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 replication 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 $z_{\alpha_S} = z_\alpha/\sqrt{\varphi}$ and the corresponding *golden level*

$$\alpha_S = 1 - \Phi(z_\alpha/\sqrt{\varphi}) \tag{8}$$

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 shown in Figure 1, the sceptical p-value is $p_S = 0.11 > 0.062$, so the replication of the Pyc and Rawson (2010) study was not successful.

The golden level (8) is derived from (7) with $d_\infty = 1$. However, we may also use a different value for the limiting relative effect size d_∞, say $d_\infty = 0.8$. Then replication success is only possible for a borderline significant result ($p_o = \alpha$) if there is less than $1 - d_\infty$ (20% for $d_\infty = 0.8$) shrinkage of the replication effect estimate. This approach is equivalent to a limiting relative effect size of 1 if the original p-value p_o is equal to a different level α', which can be derived as follows: First, solving (7) for $d_\infty > 0$ gives $K = z_\alpha^2/z_{\alpha_S}^2 = 1/2 + \sqrt{1/4 + 1/d_\infty^2}$. The new level α' fulfills $\varphi = z_{\alpha'}^2/z_{\alpha_S}^2$, so $z_\alpha^2/K = z_{\alpha'}^2/\varphi$ and therefore

$$\alpha' = 1 - \Phi\left(z_\alpha \sqrt{\varphi/K}\right). \tag{9}$$

For example, for $\alpha = 0.025$ and $d_\infty = 0.8$ we obtain $\alpha' = 0.033$.

2.3. *Recalibration of the sceptical p-value.* The condition $p_S \leq \alpha_S$ for replication success at the golden level is equivalent to $z_S \geq z_\alpha/\sqrt{\varphi}$, i. e. $z_S \sqrt{\varphi} \geq z_\alpha$. 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 $\tilde{p}_S = 0.061$ and does not lead to replication success at any level $\alpha < 0.061$, including the standard 0.025 level.

2.4. *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.

It is interesting to directly compare the two-trials rule and replication success at the golden level in terms of the required relative effect size d to fulfill the criteria (2) and (6), respectively, see Figure 2. 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. For example, four studies with one-sided $p_o \in (0.025, 0.03)$ have been included in the *Reproducibility Project: Psychology* (Open Science Collaboration, 2015) and one of them
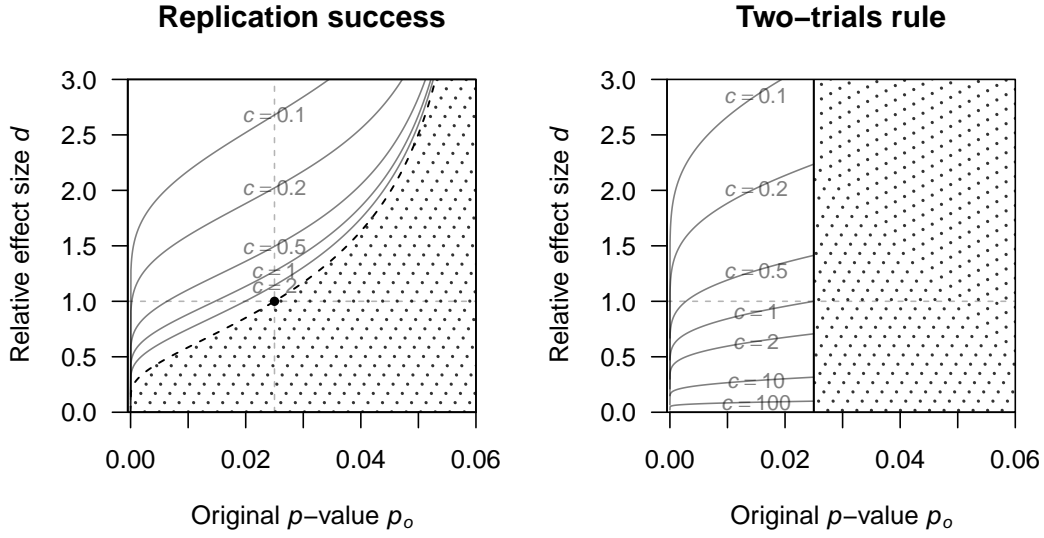
FIG 2. *Comparison of replication success at the golden level ($p_S \leq \alpha_S = 0.062$) and the two-trials rule ($p_o \leq 0.025$ and $p_r \leq 0.025$). The dotted areas indicate that success is impossible for original p-value p_o and relative effect size d. In the white areas success is possible and depends on the relative sample size c as indicated by the grey lines. The dashed black line in the left plot indicates the limiting relative effect size d_∞.*

achieves replication success (see Section 4 for details). By definition, such non-significant original findings can never fulfill the two-trials rule.

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. Power and Type-I Error Rate. Although Bayesian methods do not rely on the frequentist paradigm of repeated testing, it is still useful to investigate their frequentist operating characteristics (Dawid, 1982; Rubin, 1984; Grieve, 2016) and this also holds for the proposed reverse-Bayes assessment of replication success. We first condition on the results from the original study and compare the power to achieve replication success with the two-trials rule in Section 3.1. We then assume that none of the two studies have been conducted and investigate the overall Type-I error rate (Section 3.2) and the project power (Section 3.3) (Maca et al., 2002) over both studies in combination for fixed relative sample size c.

3.1. *Conditional power.* Figure 3 compares the power for replication success (see Held, 2020a, Section 4 for details) 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 p_o from the original study. Shown is the conditional power assuming the unknown parameter θ is equal to the original effect estimate $\hat{\theta}_o$. Then $\hat{\theta}_r | \hat{\theta}_o \sim N(\hat{\theta}_o, \kappa^2/n_r)$ and it follows that $d | \hat{\theta}_o \sim N(1, 1/(cz_o^2))$. The conditional power for replication success can therefore be calculated as

$$(10) \qquad \Pr(d \geq d_{\min} | \hat{\theta}_o) = \Phi\left[\sqrt{c}z_o(1 - d_{\min}) \right],$$
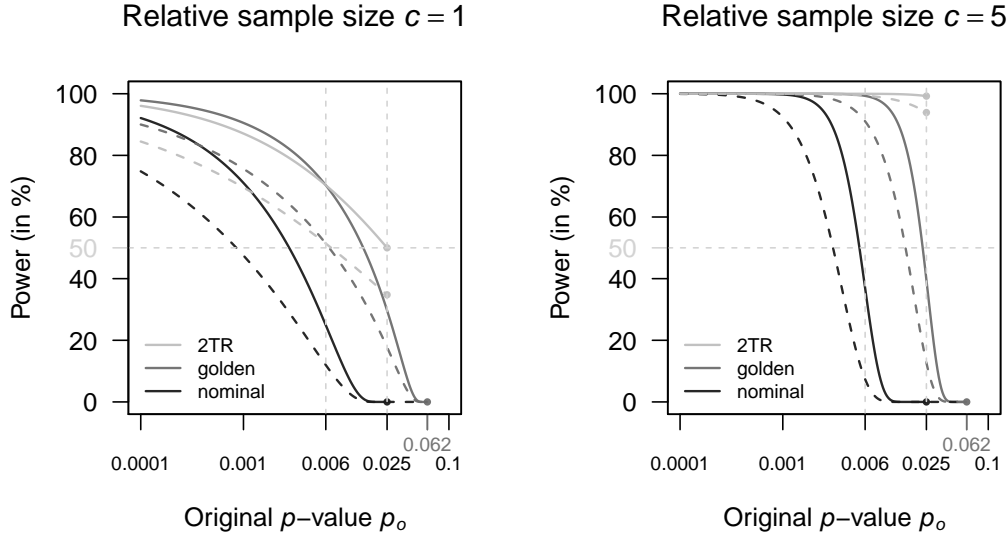
FIG 3. *Conditional power as a function of the one-sided p-value of the original study with relative sample size $c = 1$ (left) and $c = 5$ (right). Shown is conditional power assuming the unknown parameter is equal to the original effect estimate (solid) and conditional power based on 20% shrinkage of the original effect estimate (dashed) for the two-trials rule (2TR) at level $\alpha = 0.025$ and for replication success at the corresponding golden and nominal level. Power values of exactly zero are omitted.*

where $d_{\min}$ is given in (6). Predictive power, which is conditional power averaged over a $N(\hat{\theta}_o, \sigma_o^2)$ distribution for the effect size θ, could also be calculated, then $d \mid \hat{\theta}_o \sim N(1, (1 + 1/c)/z_o^2)$. Conditional and predictive power of the two-trials rule also depend on z_o, c and α and are given in Micheloud and Held (2021).

The two-trials rule requires a significant original study and hence it is impossible to power a replication study when $p_o > 0.025$. The same applies for replication success at the nominal level, where the power is zero for any $p_o > 0.025$, regardless of the replication sample size. This is different for the golden level, where the conditional power of an original study with $0.025 < p_o < 0.062$ is low, but not zero. However, if the original p-value p_o is slightly smaller than 0.025, the two-trials rule has a larger power, both for $c = 1$ and $c = 5$. But if the original p-value is sufficiently small ($p_o < 0.006$ for $c = 1$), the power for replication success at the golden level is larger than the power of the two-trials rule.

Compared to $c = 1$, the conditional power for $c = 5$ of both the two-trials rule and the replication success approach at the golden level increases if $p_o \leq \alpha$. A remarkable feature of the replication success approach at the golden level is that conditional power can be pushed towards 100% for large enough c if $p_o < \alpha$, but not otherwise. This can be seen from (10) because $d_{\min} < 1$ for $p_o < \alpha$ and large enough relative sample size c. On the other hand, for $p_o > \alpha$ conditional power for replication success will tend to 0% for increasing c because $d_{\min} > 1$ for all c. Finally, for $p_o = \alpha$ the limit is 50%. The same property can be observed at the nominal level, however at the smaller threshold $1 - \Phi(z_\alpha \sqrt{\varphi})$ which is 0.006 for $\alpha = 0.025$. Only if $p_o < 0.006$ will the conditional power for replication success attain 100% for $c \to \infty$. This further highlights the stringency of the nominal level.

The approach described so far takes the original study at face-value since it assumes that $\hat{\theta}_o$ is equal to the unknown effect size θ. In practice, however, there are often good reasons to believe that original effect estimates have a tendency to be inflated (*e. g.* due to publication

bias). One way to address this issue is to base power calculations on a shrunken version of the original effect estimate, where the amount of shrinkage is guided by domain knowledge and a risk of bias assessment of the original study. For illustration, Figure 3 also shows conditional power based on 20% shrinkage of the original effect estimate which reduces the conditional power for all methods, especially for a relative sample size $c = 1$. Conditional power for replication success at the golden level can now be pushed towards 100% only for $p_o < 0.018$, which can be derived by solving (9) for α with $\alpha' = 0.025$ and $d_\infty = 0.8$. To be able to push conditional power based on 20% shrinkage towards 100% for all $p_o < 0.025$, equation (9) would have to be used directly to relax the level from $\alpha = 0.025$ to $\alpha' = 0.033$.

3.2. *Overall Type-I error rate.* The two studies are assumed to be independent with Type-I error rate fixed at α for each of them, so the Type-I error rate of the two-trials rule over the entire project 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 the relative sample size c.

For $c = 1$, Held (2020a, Section 3) showed that z_S^2 in (4) 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 at level α_S for $c = 1$ as

$$\text{(11)} \qquad \text{T1E} = \left\{ 1 - \Phi\left[2\,\Phi^{-1}(1 - \alpha_S) \right] \right\}/2.$$

For the golden level $\alpha_S = 0.062$ at $\alpha = 0.025$, the Type-I error rate (11) is 0.0515%, 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%, much smaller than 0.0625%.

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

$$\text{(12)} \qquad \text{T1E} = \int_{z_{\alpha_S}}^{\infty} \Pr(z_r \geq z_r^{\min} \mid z_o, c, \alpha_S)\, \phi(z_o)\, dz_o,$$

where $\phi(\cdot)$ denotes the standard normal density function. The first term in the integral of (12) 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. Now $z_r \sim \text{N}(0, 1)$ under the null hypothesis, so this term simplifies to $\Pr(z_r \geq z_r^{\min} \mid z_o, c, \alpha_S) = 1 - \Phi(z_r^{\min})$ where $z_r^{\min}$ in (5) depends on z_o, c, and α_S.

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 of the replication success approach decreases with increasing relative sample size c. This also follows from (12) where $\Pr(z_r \geq z_r^{\min} \mid z_o, c, \alpha_S) = 1 - \Phi(z_r^{\min})$ decreases with increasing c, because $z_r^{\min}$ increases with increasing c, see equation (5).

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.

Figure 5 compares for $c = 1$ the Type-I error rate (11) of replication success at the golden and at the nominal level with the two-trials rule for different values of α. The Type-I error rate of the two-trials rule is α^2 and the replication success approach at the nominal level always has a much smaller Type-I error rate than α^2. At the golden level the Type-I error rate of the replication success approach is much closer to α^2, still slightly smaller if $\alpha < 0.058$. For
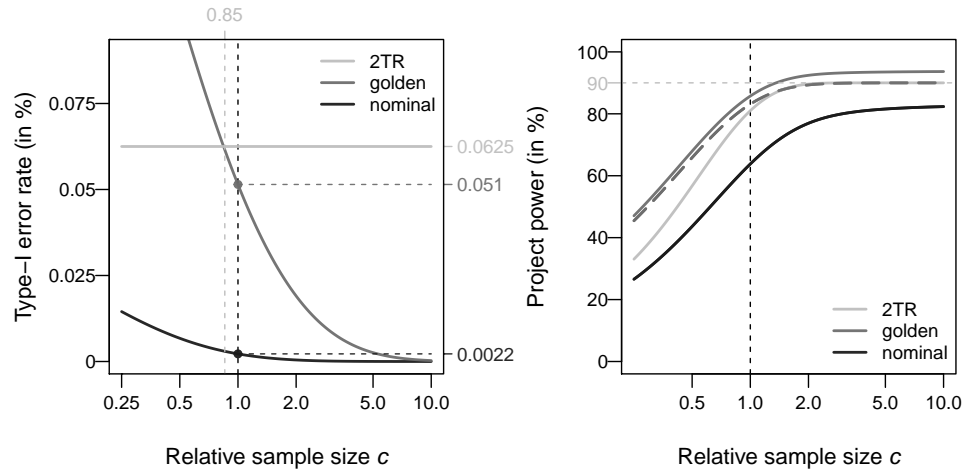
FIG 4. *Overall Type-I error rate (left) and project power (right) for fixed relative sample size c. Results are given for replication success at the nominal and golden level and compared with the two-trials rule (2TR) at $\alpha = 0.025$. The dashed darkgrey line is the project power at the golden level based on significant original studies ($p_o \leq 0.025$). The power of the original study is 90%.*
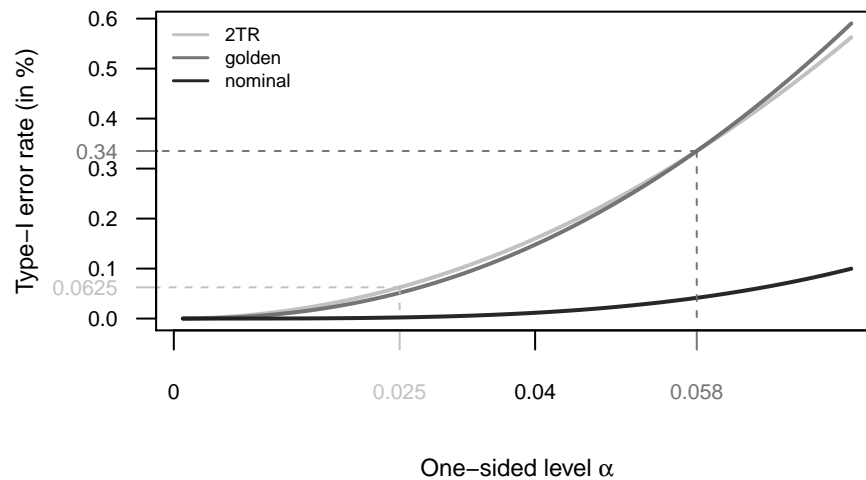

FIG 5. *Overall Type-I error rate if the replication sample size equal to the original study (c = 1). The two-trials rule (2TR) is compared to replication success at the golden and nominal level for different values of α.*

$\alpha = 0.058$ the Type-I error rate is equal to the Type-I error rate $0.058^2 = 0.34\%$ of the two-trials rule and for $\alpha > 0.058$ the Type-I error rate is slightly larger than α^2. The Type-I error rate for replication success decreases with increasing c, so as long as the replication sample

size is not smaller than the original sample size, Type-I error control at α^2 is guaranteed at the golden level for any one-sided level $\alpha < 0.058$.

3.3. *Project power.* Under the alternative we have $z_o \sim N(\mu, 1)$ with $\mu = z_\alpha + z_\beta$ where α is the assumed significance level and $1 - \beta = \Phi(\mu - z_\alpha)$ is the power to detect the assumed effect $\theta = \mu \sigma_o$ in the original study (Matthews, 2006, Section 3.3). In the following $\alpha = 0.025$ and $\beta = 0.1$ are used. The power of a significant replication study with sample size $n_r = cn_o$ is

$$\Phi(\theta/\sigma_r - z_\alpha) = \Phi(\sqrt{c}\mu - z_\alpha),$$

so depends on μ and the relative sample size c. The project power of the two-trials rule is therefore $(1 - \beta)\,\Phi(\sqrt{c}\mu - z_\alpha)$ and increases with increasing c.

The project power for replication success is computed as

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

and shown in the right plot of Figure 4 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 level and converges to 90% for large c.

The advantage in power stems partly from replication success still being possible when the original p-value is larger than 0.025, but smaller than 0.062. If we assume that a replication study is only conducted if the original study is significant (with $p_o \leq 0.025$), then the project power based on the golden level (the dashed line in Figure 4) is slightly smaller and for $c > 1$ barely different than for the two-trials rule. More substantial gains are still visible for $c < 1$. However, the restriction to original studies with $p_o \leq 0.025$ may not reflect current practice in large-scale replication projects. For example, 5 out of 143 replication studies considered in Section 4 do have original p-values between 0.025 and 0.062.

4. 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 and 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).

Table 1 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 at the golden level), which are around 30% for the former and more than 70% for the latter. The proportion of successful replications is similar for the two-trials rule and the replication success approach. In the Experimental Economics project the methods perfectly agree, while in the other three projects the methods disagree for a few studies.

TABLE 1

Results for each replication project: Relative effect size d (median with 25% and 75% quantiles on Fisher's z scale), proportion 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.

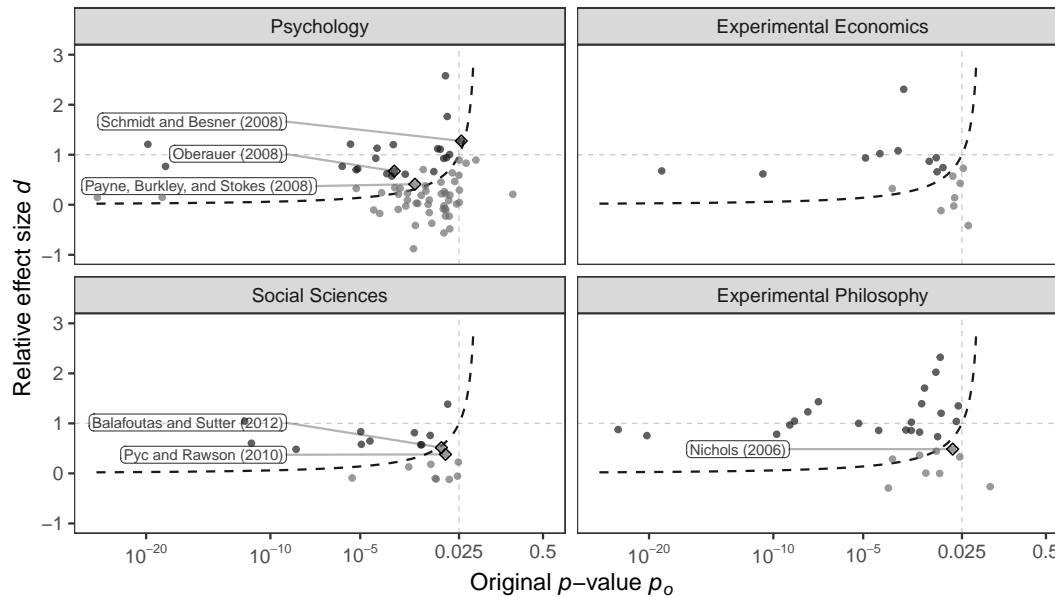| Project	relative effect size d	2TR (%)	RS (%)	discrepant
Psychology	0.29 [0.03, 0.77]	28.8	30.1	3/73
Experimental Economics	0.67 [0.35, 0.92]	55.6	55.6	0/18
Social Sciences	0.52 [0.13, 0.65]	61.9	52.4	2/21
Experimental Philosophy	0.86 [0.47, 1.12]	74.2	71.0	1/31

FIG 6. *Relative effect size d versus original p-value p_o. Black indicates that replication success was achieved at the golden level while grey indicates that it was not. The diamonds mark studies where the replication success approach (at the golden level) and the two-trials rule disagree. The dashed black line indicates the limiting relative effect size at the golden level with $\alpha = 0.025$.*

Figure 6 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 6 and their characteristics are summarised in Table 2. Two studies from the Psychology project

TABLE 2
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, Burkley and Stokes (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**

show replication success but fail the 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 $c = 9.2$.

This analysis was based on the default choice $d_\infty = 1$ at $\alpha = 0.025$ for the golden level as described in Section 2.2. We may also choose a different value for the limiting relative effect size d_∞ at $\alpha = 0.025$ which then corresponds to $d_\infty = 1$ at a different level α' as given in (9). Figure 7 compares the proportion of successful replications with the replication success approach for $d_\infty \in (0.5, 1.1)$ with the two-trials rule at the corresponding levels $\alpha' \in (0.06, 0.022)$ for all four replication projects. We can see that the two proportions agree fairly well for all values of α' considered. The number of discrepant studies in each project varies between 0 and 3. Only in the Psychology project there are some studies which are successful with the replication success approach but not the two-trials rule and some studies successful with the two-trials rule but not the replication success approach. The proportion of studies where both methods are successful (also shown in Figure 7) is then smaller than the proportion of successful replications with either one of the two methods. The three discrepant studies from the Psychology project listed in the top three rows of Table 2 are an example of this particular feature.

5. 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
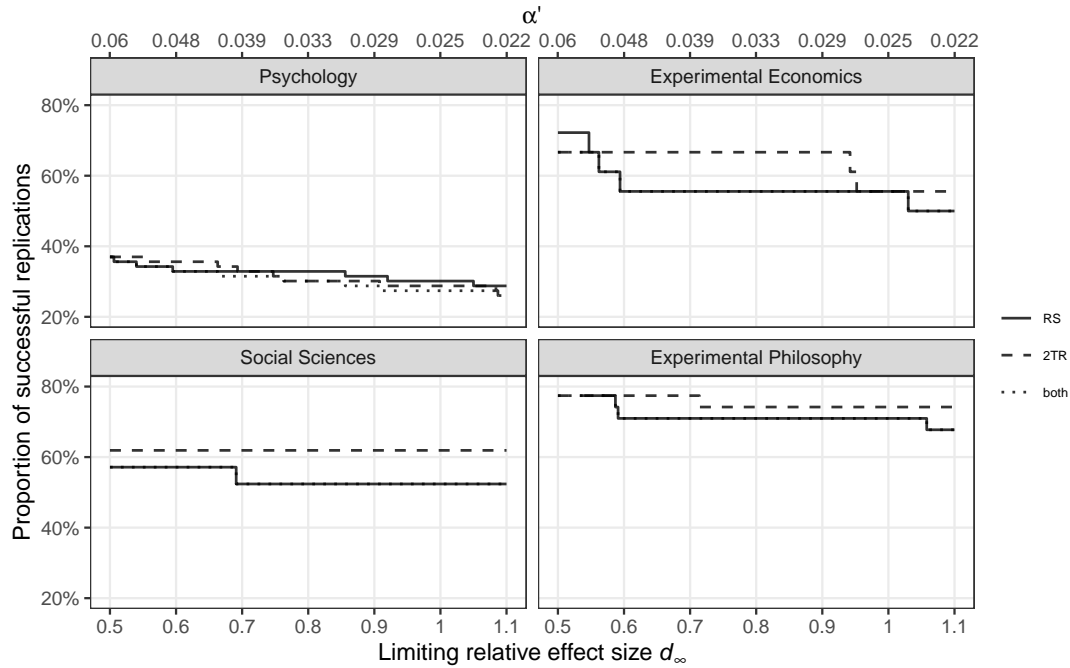
FIG 7. *Proportion of successful replications as a function of the limiting relative effect size d_∞ at $\alpha = 0.025$. The upper axis gives the equivalent level α' where the corresponding limiting relative effect size is 1. The replication success (RS) approach is compared with the two-trials rule (2TR).*

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.

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 at any one-sided level $\alpha < 0.058$ if the replication sample size is not smaller than the original one. 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. The good performance has been recently confirmed by a comparison of different replication success metrics through a simulation study in the presence of publication bias (Muradchanian et al., 2021).

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. However, in extreme cases the performance of the sceptical p-value may be considered as strange or even counterintuitive. Specifically, if the original study was only borderline significant, a highly significant replication study can only lead to success if the replication effect estimate is larger than the original one. To understand this behaviour it is important to realize that the proposed approach does not synthesize the evidence from the two studies (like a standard meta-analysis).

The sceptical p-value is designed to confirm claims of new discoveries through replication, but will remain "stubborn" (Ly and Wagenmakers, 2020) if the original study was not particularly convincing, even if the replication study provides overwhelming evidence for an effect. It will lead to a different result if the order of studies was reversed, as long as original and replication study do not have the same sample size ($c \neq 1$). The related harmonic mean χ^2-test (Held, 2020b) for evidence synthesis of two or more studies also requires each study to be convincing on its own to a certain degree, but treats them as exchangeable.

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. Third, there is a recent trend to not only conduct one but several replications for one original study (*e. g.* Klein et al., 2014; Ebersole et al., 2016; Klein et al., 2018). 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.

Throughout the paper we have assumed that the relative sample size is fixed in advance. In practice the sample size of the replication study is often chosen based on the result of the original study (Anderson and Maxwell, 2017). Power calculations as shown in Figure 3 can then be inverted to determine the appropriate sample size of the replication study. We can also invert equation (6) to obtain the required replication sample size based on the specification of the minimum relative effect size $d_{\min}$ to achieve replication success. This novel way of calculating the sample size requires the specification of the minimum relative effect size which can still be considered as acceptable. Sample size calculations based on the two-trials rule can also be formulated in terms of the minimum relative effect size by inverting equation (2). We will report on a detailed comparison of the different approaches in future work.

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** 305–324. https://doi.org/10.1080/00273171.2017.1289361.

Balafoutas, L. and Sutter, M. (2012). Affirmative Action Policies Promote Women and Do Not Harm Efficiency in the Laboratory. *Science* **335** 579–582. https://doi.org/10.1126/science.1211180.

Begley, C. G. and Ioannidis, J. P. A. (2015). Reproducibility in Science. *Circulation Research* **116** 116–126. https://doi.org/10.1161/CIRCRESAHA.114.303819.

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://doi.org/10.2307/2982063.

Camerer, C. F., Dreber, A., Forsell, E., Ho, T. H., Huber, J., Johannesson, M., Kirchler, M., Almenberg, J., Altmejd, A. et al. (2016). Evaluating replicability of laboratory experiments in economics. *Science* **351** 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. et al. (2018). Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour* **2** 637–644. https://doi.org/10.1038/s41562-018-0399-z.

OPEN SCIENCE COLLABORATION (2015). Estimating the reproducibility of psychological science. *Science* **349** aac4716. https://doi.org/10.1126/science.aac4716.

COVA, F., STRICKLAND, B., ABATISTA, A., ALLARD, A., ANDOW, J., ATTIE, M., BEEBE, J., BERNIŪNAS, R., BOUDESSEUL, J. et al. (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. https://doi.org/10.1080/01621459.1982.10477856.

EBERSOLE, C. R., ATHERTON, O. E., BELANGER, A. L., SKULBORSTAD, H. M., ALLEN, J. M., BANKS, J. B., BARANSKI, E., BERNSTEIN, M. J., BONFIGLIO, D. B. V. et al. (2016). Many labs 3: Evaluating participant pool quality across the academic semester via replication. *Journal of Experimental Social Psychology* **67** 68–82. https://doi.org/10.1016/j.jesp.2015.10.012.

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**. https://doi.org/10.7554/elife.04333.

FDA (1998). Providing clinical evidence of effectiveness for human drug and biological products. 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. https://doi.org/10.2307/2331802.

GRIEVE, A. P. (2016). Idle thoughts of a 'well-calibrated' Bayesian in clinical drug development. *Pharmaceutical Statistics* **15** 96–108. https://doi.org/10.1002/pst.1736.

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. https://doi.org/10.1111/rssa.12493.

HELD, L. (2020b). The harmonic mean χ^2 test to substantiate scientific findings. *Journal of the Royal Statistical Society, Series C* **69** 697–708. https://doi.org/10.1111/rssc.12410.

IOANNIDIS, J. P. A. (2005). Why Most Published Research Findings Are False. *PLoS Medicine* **2** 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** 1–10. https://doi.org/10.1080/01621459.2016.1240079.

KAY, R. (2015). *Statistical Thinking for Non-Statisticians in Drug Regulation*, Second ed. John Wiley & Sons, Chichester, U.K. https://doi.org/10.1002/9781118451885.

KLEIN, R. A., RATLIFF, K. A., VIANELLO, M., ADAMS, R. B., BAHNÍK, V., BERNSTEIN, M. J., BOCIAN, K., BRANDT, M. J., BROOKS, B. et al. (2014). Investigating variation in replicability: A "many labs" replication project. *Social Psychology* **45** 142–152. https://doi.org/10.1027/1864-9335/a000178.

KLEIN, R. A., VIANELLO, M., HASSELMAN, F., ADAMS, B. G., REGINALD B. ADAMS, J., ALPER, S., AVEYARD, M., AXT, J. R., BABALOLA, M. T., ŠTĚPÁN BAHNÍK et al. (2018). Many labs 2: Investigating variation in replicability across samples and settings. *Advances in Methods and Practices in Psychological Science* **1** 443–490. https://doi.org/10.1177/2515245918810225.

LY, A. and WAGENMAKERS, E. J. (2020). Discussion of "A new standard for the analysis and design of replication studies" by Leonhard Held. *Journal of the Royal Statistical Society, Series A* **183** 460-461. https://doi.org/10.1111/rssa.12544.

MACA, J., GALLO, P., BRANSON, M. and MAURER, W. (2002). Reconsidering some aspects of the two-trials paradigm. *Journal of Biopharmaceutical Statistics* **12** 107–119. https://doi.org/10.1081/bip-120006450.

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.

MATTHEWS, J. N. (2006). *Introduction to randomized controlled clinical trials*. Chapman and Hall/CRC https://doi.org/10.1201/9781420011302.

MICHELOUD, C. and HELD, L. (2021). Power Calculations for Replication Studies. *Statistical Science*. to appear.

MURADCHANIAN, J., HOEKSTRA, R., KIERS, H. and VAN RAVENZWAAIJ, D. (2021). How Best to Quantify Replication Success? A Simulation Study on the Comparison of Replication Success Metrics. *Royal Society Open Science* **8** 201697. https://doi.org/10.1098/rsos.201697.

NICHOLS, S. (2006). Folk Intuitions on Free Will. *Journal of Cognition and Culture* **6** 57–86. https://doi.org/10.1163/156853706776931385.

16

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** 439–459. https://doi.org/10.1037/0278-7393.34.3.439.

PAWEL, S. and HELD, L. (2020). Probabilistic forecasting of replication studies. *PLOS ONE* **15** e0231416. https://doi.org/10.1371/journal.pone.0231416.

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** 16–31. https://doi.org/10.1037/0022-3514.94.1.16.

PYC, M. A. and RAWSON, K. A. (2010). Why Testing Improves Memory: Mediator Effectiveness Hypothesis. *Science* **330** 335–335. https://doi.org/10.1126/science.1191465.

RUBIN, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *The Annals of Statistics* **12** 1151-1172. https://doi.org/10.1214/aos/1176346785.

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** 514–523. https://doi.org/10.1037/0278-7393.34.3.514.

SENN, S. (2007). *Statistical Issues in Drug Development*, Second ed. John Wiley & Sons, Chichester, U.K.

Computational details.

```
sessionInfo()
## R version 4.0.5 (2021-03-31)
## Platform: x86_64-pc-linux-gnu (64-bit)
## Running under: Ubuntu 20.04.4 LTS
##
## Matrix products: default
## BLAS:   /usr/lib/x86_64-linux-gnu/openblas-pthread/libblas.so.3
## LAPACK: /usr/lib/x86_64-linux-gnu/openblas-pthread/liblapack.so.3
##
## locale:
##  [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
##  [3] LC_TIME=en_US.UTF-8        LC_COLLATE=en_US.UTF-8
##  [5] LC_MONETARY=en_US.UTF-8    LC_MESSAGES=en_US.UTF-8
##  [7] LC_PAPER=en_US.UTF-8       LC_NAME=C
##  [9] LC_ADDRESS=C               LC_TELEPHONE=C
## [11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
##  [1] gplots_3.1.1            ggrepel_0.9.1           fanplot_3.4.2
##  [4] stringr_1.4.0           scales_1.1.1            biostatUZH_1.8.0
##  [7] survival_3.2-10         ggplot2_3.3.3           dplyr_1.0.6
## [10] ReplicationSuccess_0.1-4 xtable_1.8-4           knitr_1.33
##
## loaded via a namespace (and not attached):
##  [1] Rcpp_1.0.6        pillar_1.6.1      compiler_4.0.5   bitops_1.0-7
##  [5] tools_4.0.5       digest_0.6.27     boot_1.3-27      evaluate_0.14
##  [9] lifecycle_1.0.0   tibble_3.1.2      rootSolve_1.8.2.1 gtable_0.3.0
## [13] lattice_0.20-41   pkgconfig_2.0.3   rlang_0.4.11     Matrix_1.3-2
## [17] DBI_1.1.1         xfun_0.23         withr_2.4.2      caTools_1.18.2
## [21] gtools_3.8.2      generics_0.1.0    vctrs_0.3.8      grid_4.0.5
## [25] tidyselect_1.1.1  glue_1.4.2        R6_2.5.0         fansi_0.4.2
## [29] farver_2.1.0      purrr_0.3.4       magrittr_2.0.1   codetools_0.2-18
## [33] ellipsis_0.3.2    splines_4.0.5     assertthat_0.2.1 colorspace_2.0-1
## [37] labeling_0.4.2    KernSmooth_2.23-18 utf8_1.2.1      stringi_1.6.1
## [41] munsell_0.5.0     crayon_1.4.1
```