University of Zurich[UZH]

Zurich Open Repository and Archive
University of Zurich
University Library
Strickhofstrasse 39
CH-8057 Zurich
www.zora.uzh.ch

Year: 2023

Assessing replicability with the sceptical p-value: Type-I error control and sample size planning

Micheloud, Charlotte ; Balabdaoui, Fadoua ; Held, Leonhard

DOI: https://doi.org/10.1111/stan.12312

Posted at the Zurich Open Repository and Archive, University of Zurich
ZORA URL: https://doi.org/10.5167/uzh-253936
Journal Article
Published Version

The following work is licensed under a Creative Commons: Attribution 4.0 International (CC BY 4.0) License.

Originally published at:
Micheloud, Charlotte; Balabdaoui, Fadoua; Held, Leonhard (2023). Assessing replicability with the sceptical p-value: Type-I error control and sample size planning. Statistica Neerlandica, 77(4):573-591.
DOI: https://doi.org/10.1111/stan.12312

Check for updates

Received: 17 May 2023 | Accepted: 8 June 2023

DOI: 10.1111/stan.12312

ORIGINAL ARTICLE

WILEY

Assessing replicability with the sceptical *p*-value: Type-I error control and sample size planning

Charlotte Micheloud[1,2] | **Fadoua Balabdaoui**[3] | **Leonhard Held**[1,2]

[1]Epidemiology, Biostatistics and Prevention Institute (EBPI), University of Zurich, Zurich, Switzerland

[2]Center for Reproducible Science (CRS), University of Zurich, Zurich, Switzerland

[3]Seminar für Statistik, ETH Zurich, Zurich, Switzerland

Correspondence
Charlotte Micheloud, Epidemiology, Biostatistics and Prevention Institute (EBPI), University of Zurich, Hirschengraben 84, 8001 Zurich, Switzerland.
Email: charlotte.micheloud@uzh.ch

Funding information
Schweizerischer Nationalfonds zur Förderung der Wissenschaftlichen Forschung

We study a statistical framework for replicability based on a recently proposed quantitative measure of replication success, the sceptical *p*-value. A recalibration is proposed to obtain exact overall Type-I error control if the effect is null in both studies and additional bounds on the partial and conditional Type-I error rate, which represent the case where only one study has a null effect. The approach avoids the double dichotomization for significance of the two-trials rule and has larger project power to detect existing effects over both studies in combination. It can also be used for power calculations and requires a smaller replication sample size than the two-trials rule for already convincing original studies. We illustrate the performance of the proposed methodology in an application to data from the Experimental Economics Replication Project.

KEYWORDS

design of replication studies, power calculations, replicability, sceptical *p*-value, two-trials rule, Type-I error control

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.
© 2023 The Authors. *Statistica Neerlandica* published by John Wiley & Sons Ltd on behalf of Netherlands Society for Statistics and Operations Research.

1 | INTRODUCTION

Replication plays a key role to build confidence in the scientific merit of published results. The so-called replication crisis has led to increased interest in replication studies over the last decade (National Academies of Sciences, Engineering, and Medicine, 2019; Royal Netherlands Academy of Arts and Science, 2018) with large-scale replication projects being conducted in various fields (Camerer et al., 2016, 2018; Errington et al., 2021; Open Science Collaboration, 2015). Deciding whether a replication is successful is, however, not a straightforward task, and different statistical methods are currently being used. For example, the Reproducibility Project: Cancer Biology (Errington et al., 2021), an 8-year effort to replicate experiments from high-impact cancer biology papers, has used no less than seven different methods to assess replicability, including significance of both the original and replication study, compatibility of the original and replication effect estimates, and computation of a meta-analytic combined effect estimate with confidence interval.

Declaring a replication as successful if both the original and replication study are significant at level α (usually one-sided .025) is known as the two-trials rule in drug development and serves as a useful benchmark. Specifically, it has an overall Type-I error (T1E) rate of α^2 (Senn, 2007) if the effect is null in both studies and the partial T1E rate, the risk of a false claim of replication success if a true effect is present in at most one of the two studies, is bounded by α (Heller, Bogomolov, & Benjamini, 2014). However, the "double dichotomization" of the two-trials rule has serious limitations. First, it is common practice to replicate interesting findings even if the original study does not pass the much-critized "bright-line" threshold of $\alpha = .025$. For example, in the Open Science Collaboration (2015) Psychology replication project, four studies have been included despite "falling a bit short of the [two-sided] .05 criterion—$p = .0508, .0514, .0516$, and .0567—but all of these were interpreted as positive effects." Similarly, the Experimental Economics Replication Project (Camerer et al., 2016) has chosen to replicate 18 studies, two of which have not been significant at the conventional two-sided .05 standard. Strict application of the two-trials rule, however, would make it pointless to try to replicate such nonsignificant original studies. Second, the two-trials rule has been shown to have relatively low project power, that is, power to detect existing effects over both studies in combination (Held, 2020b; Maca, Gallo, Branson, & Maurer, 2002). These issues suggest investigating alternative methods to assess replication success.

A recent proposal by Held (2020a) combines a reverse-Bayes approach (see Held, Matthews, Ott, & Pawel, 2022, for a recent review) with a prior-predictive check for conflict (Box, 1980) and gives rise to a quantitative measure of replication success, the sceptical p-value. The sceptical p-value depends on the two study-specific p-values, but also on the ratio of the variances of the original and replication effect estimates. The method treats the original and replication study not as exchangeable and specifically penalizes shrinkage of the replication effect estimate, compared to the original one. The effect size perspective has been further explored to propose a modification based on the golden ratio (Held, Micheloud, & Pawel, 2022), in the following called the golden sceptical p-value. While significance of both studies is a necessary but not sufficient success criterion in the original formulation, the golden sceptical p-value also allows original studies with a "trend to significance" to be successful at replication, but only if the effect estimate at replication is larger than at original.

The golden sceptical p-value addresses some of the problems of the two-trials rule. It can flag replication success if the original or replication p-value does not meet the significance threshold α and provides larger project power than the two-trials rule. However, the probability for replication success if the observed original effect estimate is the true effect while being nonsignificant

14679574, 2023, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/stan.12312 by Eth Zürich, Wiley Online Library on [17/10/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

is always smaller than 50%. This is less extreme than the two-trials rule where nonsignificant original studies can never lead to replication success, but precludes sample size planning for replication studies of nonsignificant original findings at commonly used power values such as 80% or 90%. Furthermore, neither the original (nominal) nor the golden sceptical p-value has an exact overall T1E rate of α^2 if the effect is null in both studies. The T1E rate of the nominal one is always below α^2, whereas the T1E rate of the golden one can exceed α^2 if the variance ratio (original to replication) is smaller than one. An alternative reverse-Bayes approach based on Bayes factors has also been proposed, but the resulting sceptical Bayes factor can also not be used for sample size planning if the original result was not convincing on its own (Pawel & Held, 2022, Section 3.3).

In this paper we study the sceptical p-value from a frequentist perspective and examine its T1E rate in greater detail. We aim to control the overall T1E rate rather than the partial T1E rate to allow for a fair comparison with the two-trials rule (Rosenkranz, 2023). Any other method with partial T1E rate $< \alpha$ (such as the nominal sceptical p-value) will have a smaller overall T1E rate than the two-trials rule, and its success region will be a subset of the success region of the two-trials rule. The ultimate goal is hence to recalibrate the sceptical p-value to achieve exact overall T1E control at level α^2 and to enable sample size calculations also for nonsignificant original studies. However, any method with the same overall T1E rate as the two-trials rule will have an increased partial T1E rate. This is also the case for the sceptical p-value, but we will show that the increase in conditional T1E rate, the risk of a false claim of success if the replication study is properly powered based on the original result, is always below $2\alpha = 5\%$, which is considered a "sensible option" by Rosenkranz (2002).

In Section 2, we describe the underlying statistical framework for replicability and consider T1E rates under two different null hypotheses, the intersection and the union null (Heller et al., 2014) in Section 2.1. The two-trials rule (Section 2.2) and the harmonic mean χ^2-test (Section 2.3) are identified as special cases of this framework with exact overall T1E control of α^2 under the intersection null. The relevant null distribution is then derived in all other cases and the sceptical p-value is recalibrated in Section 2.4 to achieve exact overall T1E control for every possible value of the variance ratio. Limiting cases and further properties are described in Sections 2.5 and 2.6. In Section 3, the sceptical p-value is used as a dichotomous criterion for replication success with focus on the partial T1E rate under the union null hypothesis in Section 3.1. The sceptical p-value and the two-trials rule are then compared in terms of success regions (Section 3.2), project power (Section 3.3) and for the design of replication studies, with particular focus on the conditional T1E rate (Section 3.4). An application to data from the Experimental Economics Replication Project is given in Section 4. We close with some discussion in Section 5.

2 | A STATISTICAL FRAMEWORK FOR REPLICABILITY

Let $\hat{\theta}_i$ denote the estimate of the unknown effect size θ_i and σ_i the corresponding standard error from the original and replication study, $i \in \{o, r\}$. As in standard meta-analysis we assume that the $\hat{\theta}_i$'s are independent and follow a normal distribution with mean θ_i and known variance σ_i^2. Let $z_i = \hat{\theta}_i/\sigma_i$ denote the test statistic for the null hypothesis H_0^i: $\theta_i = 0$, $i \in \{o, r\}$, and $p_i = 1 - \Phi(z_i)$ the corresponding one-sided p-value for the alternative H_1^i: $\theta_i > 0$, here $\Phi(.)$ denotes the standard normal cumulative distribution function. Replication success at level γ is achieved if

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

14679574, 2023, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/stan.12312 by Eth Zürich, Wiley Online Library on [17/10/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

holds, here $x_+ = \max\{0, x\}$, $c = \sigma_o^2/\sigma_r^2 > 0$ is the variance ratio and $z_\gamma = \Phi^{-1}(1-\gamma) > 0$ is the threshold at replication success level γ.

The two-sided formulation only requires (1), irrespectively of the signs of the estimates $\hat{\theta}_o$ and $\hat{\theta}_r$, but suffers from the "replication paradox" (Ly, Etz, Marsman, & Wagenmakers, 2019) because replication success can occur even if the effect estimates $\hat{\theta}_o$ and $\hat{\theta}_r$ are in opposite directions. The one-sided formulation avoids this problem with the additional requirement that the two estimates are both in the same prespecified (w.l.o.g. positive) direction,

$$\hat{\theta}_o > 0 \text{ and } \hat{\theta}_r > 0, \tag{2}$$

and so we usually require both (1) and (2) to achieve replication success (if not stated otherwise).

The success conditions (1) and (2) can be motivated from a recent proposal to define replication success with a two-step procedure (Held, 2020a): First, a significant original study at one-sided level γ is challenged by a normal prior with mean zero modeling the belief of a hypothetical sceptic who regards the absence of an effect to be the most likely reality (Matthews, 2018). The prior variance is chosen such that the posterior probability that the effect is negative is γ. Second, the conflict between the replication study result and the sceptical prior is quantified with a prior-predictive tail probability p_{Box} (Box, 1980). Replication success at level γ is then achieved if $p_{\text{Box}} \leq \gamma$, that is, if there is more conflict between the sceptical prior and the replication study than there was evidence against the null hypothesis based on the original data.

We are often interested in the smallest possible value of z_γ^2 which solves (1) and denote this value as $z_S^2 \in (0, \min\{z_o^2, z_r^2\})$, defined as the smallest positive root of

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

This is a quadratic equation in z_S^2 and can be solved analytically. Any $z_S = +\sqrt{z_S^2} \geq z_\gamma$ will hence lead to replication success at level γ, so the threshold z_γ in (1) serves as a critical value for the test statistic z_S. If the effect estimates fulfill (2), the transformation $p_S = 1 - \Phi(z_S)$ defines the (one-sided) sceptical p-value in its original formulation and the criterion $z_S \geq z_\gamma$ for replication success translates to $p_S \leq \gamma$. If (2) does not hold we set $p_S = \Phi(z_S)$ (Held, 2020a, Section 3.3).

2.1 | Null hypotheses and Type-I error rates

The T1E rate is the probability of a false claim of replication success under a given null hypothesis. In the replication setting with two studies, this probability can be considered under two different null hypotheses (Heller et al., 2014). The *intersection null hypothesis* is a point null hypothesis, defined as the intersection of the study-specific null hypotheses H_0^i: $\theta_i = 0$, $i \in \{o, r\}$:

$$H_0^o \cap H_0^r. \tag{4}$$

The probability of a false claim of replication success with respect to the intersection null (4) is the *overall* T1E rate.

The *no-replicability* or *union null hypothesis* is defined as the complement of the alternative that the effect is nonnull in both studies. This is a composite null hypothesis, which also includes

14679574, 2023, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/stan.12312 by Eth Zürich, Wiley Online Library on [17/10/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

the possibility that only one study has a null effect:

$$H_0^o \cup H_0^r. \tag{5}$$

The probability of a false claim of replication success with respect to the union null (5), the *partial* T1E rate, depends on the values of θ_o and θ_r. One of them is zero but the other one may not be zero. The partial T1E rate has been recently investigated by Zhan, Kunz, and Stallard (2023) for the two-trials rule. In Section 3.4, we also study the T1E rate under H_0^r only, conditional on the result of the original study. This *conditional* T1E rate is of primary interest because in practice the design of the replication study depends on the result from the original study (Anderson & Kelley, 2022).

A necessary but not sufficient condition for the replication success criterion (1) to hold is $\min\{|z_o|, |z_r|\} > z_\gamma$, as otherwise the left-hand side of (1) is zero. Combined with (2) this translates to the necessary but not sufficient requirement $p_{\max} = \max\{p_o, p_r\} < \gamma$. Under the union null hypothesis, either p_o or p_r is uniformly distributed, so γ is a bound on the partial T1E rate of the sceptical p-value for any value of the variance ratio c. Likewise, the overall T1E rate is smaller than γ^2 due to independence of the two studies.

This raises the question what value for γ to use in (1). The *nominal* success level is the standard significance level $\gamma = \alpha$, so controls the overall and partial T1E rate at α^2 respectively α for every value of c. However, T1E control is not exact and the overall T1E rate can be considerably smaller than α^2 (Held, Micheloud, & Pawel, 2022, Section 3.2). Replication success at the nominal level is hence only possible if both p-values are smaller than α and is much more stringent than the two-trials rule. The *golden* success level is $\gamma(\alpha) = 1 - \Phi(z_\alpha/\sqrt{\varphi}) > \alpha$, where $z_\alpha = \Phi^{-1}(1-\alpha)$ and $\varphi = (\sqrt{5}+1)/2$ is the golden ratio. For example, $\gamma(\alpha = .025) = .062$. It is therefore less restrictive than the nominal level in the assessment of replication success. By construction, it controls the partial T1E rate at level $\gamma(\alpha)$ and the overall T1E rate at $\gamma(\alpha)^2$ and even at α^2 if $c \geq 1$ and $\alpha \leq .058$ (Held, Micheloud, & Pawel, 2022, Section 3.2), but the actual overall T1E rate can be much smaller than the corresponding bound. Comparing p_S to the golden level $\gamma(\alpha)$ is equivalent to comparing the golden sceptical p-value $\tilde{p}_S = 1 - \Phi(\sqrt{\varphi}\, z_S)$ to α.

In what follows we describe how to obtain exact overall T1E control of the sceptical p-value for any particular value of c. This is motivated through the identification of the two-trials rule as a special case of the general formulation (1) respectively (3) and leads to a *controlled* success level $\gamma_c(\alpha)$ that depends on both α and c.

2.2 | The two-trials rule

The two-trials rule requires significance of both studies at the one-sided significance level α, so corresponds to $z_o \geq z_\alpha$ and $z_r \geq z_\alpha$, and translates to the single criterion $p_{\max} \leq \alpha$. Under the intersection null (4) both p_o and p_r are uniformly distributed and so $p_{\max}$ follows a triangular Beta(2, 1) distribution with cumulative distribution function (cdf) $F(p) = p^2$, so that

$$\Pr(p_{\max} \leq \alpha) = \alpha^2 \ \text{ holds for all } \ \alpha \in (0, 1). \tag{6}$$

In the sequel, a p-value with this property will be said to have *exact squared T1E control*. The square of a triangular Beta(2, 1) distribution follows a uniform distribution, so that

14679574, 2023, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/stan.12312 by Eth Zürich, Wiley Online Library on [17/10/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

$p = p^2_{\max} = \max\{p^2_o, p^2_r\}$ has cdf

$$\Pr(p \leq \alpha) = \alpha \text{ for all } \alpha \in (0, 1). \tag{7}$$

In what follows, a p-value fulfilling (7) will be said to have *exact linear T1E control*. Note that linear T1E control is the traditional requirement for p-values (Casella & Berger, 2002, p. 397) whereas p-values with squared T1E control (such as $p_{\max}$) are useful if the p-value is based on two independent studies.

Another way to obtain $p_{\max}$ as the p-value from the two-trials rule with squared T1E control is by considering (3), but replacing the variance ratio c by 0, so $z^2_S = z^2_{\min} = \min\{z^2_o, z^2_r\}$. Now the distribution of $Y = \min\{z^2_o, z^2_r\}$ under the intersection null is of interest. It can be shown that the random variable Y has cdf

$$F_0(y) = 1 - 4[1 - \Phi(\sqrt{y})]^2 \text{ for } y \geq 0, \tag{8}$$

see Supporting Material A for a derivation. Now Y does not take into account the direction of the effect estimates and hence the corresponding p-value

$$4\,p = 1 - F_0(y = z^2_{\min}) = 4[1 - \Phi(\min\{|z_o|, |z_r|\})]^2 \tag{9}$$

is two-sided with exact linear T1E control. We use the notation $4\,p$, as there are two studies (original and replication) with four possible sign combinations of $\hat{\theta}_o$ and $\hat{\theta}_r$. If the combination (2) is fulfilled, the one-sided p-value is therefore obtained from (9) through division by 4:

$$p = \left[1 - F_0(z^2_{\min})\right]/4 = [1 - \Phi(\min\{z_o, z_r\})]^2 = (\max\{p_o, p_r\})^2 = \max\{p^2_o, p^2_r\}$$

and so $p = p^2_{\max}$ respectively $p_{\max} = \sqrt{p}$.

This insight suggests a strategy to obtain a sceptical p-value with exact squared overall T1E control: If we can derive the distribution function $F_c(.)$ of z^2_S in (3) under the intersection null for any value of $c > 0$, then we can use the transformation $p = [1 - F_c(z^2_S)]/4$, provided (2) holds. The *controlled* sceptical p-value then is $p^*_S = \sqrt{p}$ and has exact squared overall T1E control. If (2) does not hold, we set $p^*_S = 1 - \sqrt{p}$. For simplicity, we call p^*_S the sceptical p-value in the following, if no misunderstandings can arise. Figure 1 summarizes the different steps from the null distribution function $F_c(.)$ to the assessment of replication success at overall T1E rate α^2 with the two-trials rule and the sceptical p-value.

2.3 | The null distribution for equal variances

We first consider the case $c = 1$, where the null distribution of z^2_S is available from the harmonic mean χ^2-test (Held, 2020b). The solution of (3) then is

$$z^2_S = z^2_H/2 \tag{10}$$

where $z^2_H = 2/(1/z^2_o + 1/z^2_r)$ is the harmonic mean of the squared test statistics z^2_o and z^2_r. It can be shown that z^2_S has a gamma Ga(1/2, 2) distribution under the intersection null (4), where z^2_o and z^2_r are independent $\chi^2(1)$-distributed (Pillai & Meng, 2016, eq. (2.3)). The cdf $F_{c=1}(y)$ of $Y = z^2_S$ is

14679574, 2023, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/stan.12312 by Eth Zürich, Wiley Online Library on [17/10/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

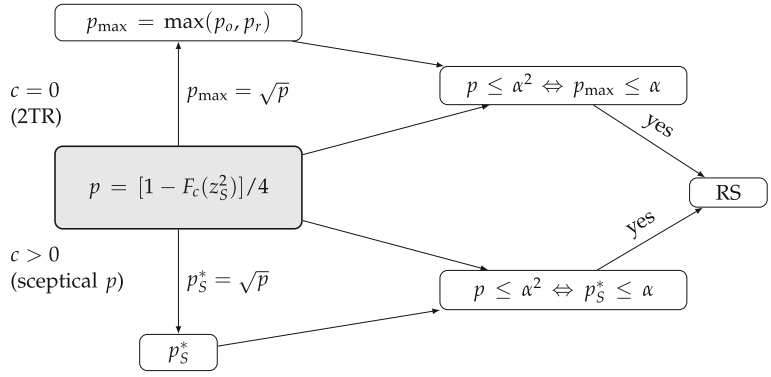

FIGURE 1 Replication success (RS) with the two-trials rule (2TR) and the controlled sceptical p-value p_S^*. The cdf $F_c(.)$ is based on the distribution of z_S^2 in (3) under the intersection null and the p-values are one-sided, calculated under the assumption that (2) holds.

thus readily available and a two-sided p-value with exact linear T1E control can be calculated:

$$4\,p = 1 - F_{c=1}(y = z_S^2). \tag{11}$$

Division by 4 gives the corresponding one-sided p-value p if (2) is fulfilled and the square root $p_S^* = \sqrt{p}$ defines the controlled sceptical p-value with exact squared T1E control for $c = 1$.

2.4 | The null distribution for unequal variances

For $c > 0$ and $c \neq 1$ there is a unique solution of (3) that fulfills the requirement $0 \leq z_S^2 \leq \min\{z_o^2, z_r^2\}$:

$$z_S^2 = \frac{z_A^2}{c-1}\left\{\sqrt{1 + (c-1)z_H^2/z_A^2} - 1\right\}, \tag{12}$$

where z_A^2 is the arithmetic and z_H^2 the harmonic mean of z_o^2 and z_r^2. To obtain $F_c(.)$, consider the probabilistic version of equation (12), where the random variable $Y = z_S^2$ depends on the two random variables z_o^2 and z_r^2 through z_A^2 and z_H^2. Under the intersection null hypothesis (4), z_o^2 and z_r^2 are independent $\chi^2(1)$-distributed. Then z_A^2 and z_H^2/z_A^2 in (12) are also independent (Grimmett & Stirzaker, 2001, Section 4.7), which facilitates the computation of the cdf $F_c(y) = \Pr(Y \leq y \mid c)$ of Y. In Supporting Material B we show that

$$F_c(y) = 1 - \frac{1}{\pi}\int_0^1 \frac{\exp\{-g(y,t,c)\}}{\sqrt{t(1-t)}}\,dt \tag{13}$$

where

$$g(y,t,c) = \frac{(c-1)y}{\sqrt{1 + (c-1)t} - 1}.$$

14679574, 2023, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/stan.12312 by Eth Zürich, Wiley Online Library on [17/10/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

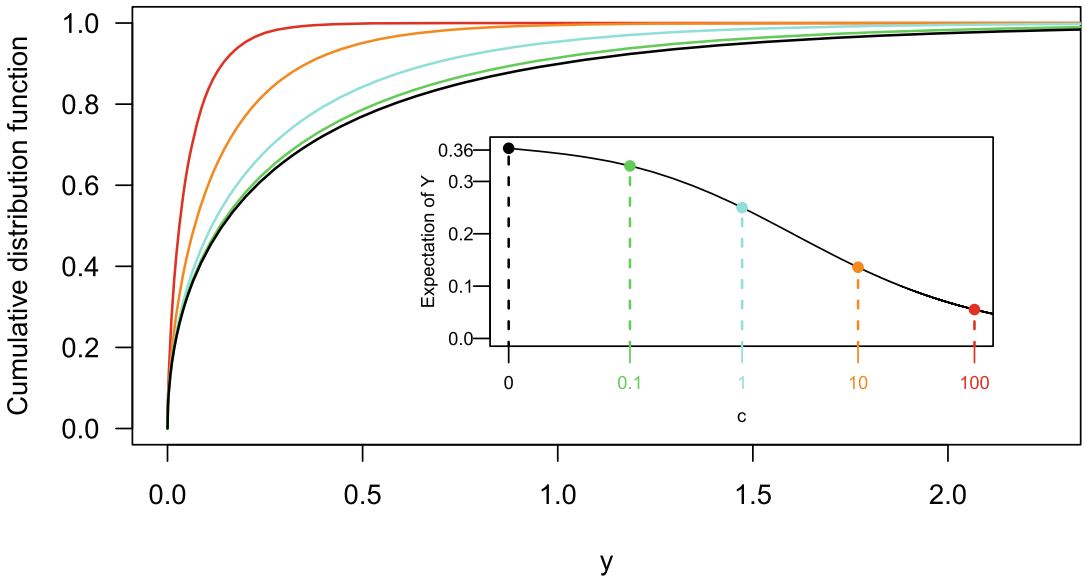

FIGURE 2 Cumulative distribution function $F_c(y)$ of $Y = z_S^2$ (main plot) and expectation of Y (inset plot) under the intersection null hypothesis (4) for different values of c.

Figure 2 compares $F_c(y)$ and the expectation of Y for different values of c, including the special cases $c = 0$ and $c = 1$. Evaluation of (13) is possible with numerical integration techniques and so a two-sided p-value with exact linear T1E control can be calculated:

$$4\,p = 1 - F_c(z_S^2) \tag{14}$$

with z_S^2 as defined in (12). If (2) is fulfilled, the corresponding one-sided p-value has exact linear T1E control and $p_S^* = \sqrt{p}$ defines the one-sided controlled sceptical p-value with exact squared T1E control.

2.5 | Limiting cases

For $c \downarrow 0$ the two-sided p-value $4\,p$ in (14) converges to $4\,p_{\max}^2$. This follows from the fact that $z_S^2 \uparrow z_{\min}^2$ for $c \downarrow 0$ (Held, 2020a, eq. (11)) with cdf (8) under the intersection null. The two-trials rule described in Section 2.2 is therefore a limiting case of our framework if we are willing to ignore the interpretation of c as the variance ratio.

For $c \to \infty$ the two-sided p-value $4\,p$ in (14) converges to

$$4\,p_\infty = \lim_{c\to\infty} 4\,p = \frac{1}{\pi}\int_0^1 \frac{\exp\left(-z_G^2/\sqrt{t}\right)}{\sqrt{t(1-t)}}\,dt, \tag{15}$$

where $z_G^2 = \sqrt{z_o^2 z_r^2} = |z_o z_r|$ is the geometric mean of the squared test statistics z_o^2 and z_r^2 (proof to be found in Supporting Material C). Note that $4\,p_\infty = 1$ if either $z_o = 0$ or $z_r = 0$, as $f(x) = 1/\{\pi\sqrt{x(1-x)}\}$ is the density of a $X \sim \text{Beta}(1/2, 1/2)$ random variable and integrates to 1.

14679574, 2023, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/stan.12312 by Eth Zürich, Wiley Online Library on [17/10/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

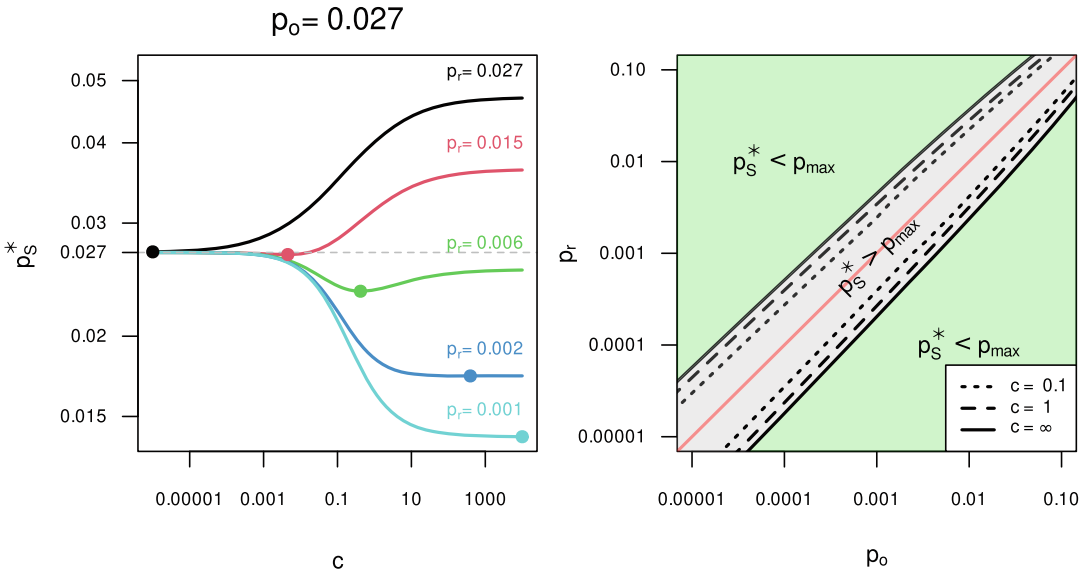

FIGURE 3 Left panel: Sceptical p-value p_S^* as a function of the variance ratio c for $p_o = .027$ and selected values of p_r. The dots represent the infimum of p_S^* for selected values of p_r and the dashed line indicates $p_{\max}$. Right panel: Combinations of p_o and p_r for which p_S^* is always smaller than $p_{\max}$ (green area), for which p_S^* is always larger than $p_{\max}$ (red line) and for which it depends on the variance ratio c (gray region). In the gray region above (respectively below) the red line, p_S^* is smaller than $p_{\max}$ for combinations of p_o and p_r laying above (respectively below) the black line for the corresponding c, and vice-versa.

Furthermore, (15) is a two-sided p-value with exact linear T1E control, that is, $4\,p_\infty$ is uniformly distributed on the unit interval if z_o^2 and z_r^2 are i.i.d. $\chi^2(1)$, see Supporting Material D for a proof.

Other well-known methods that combine two (or more) p-values are Fisher's (Fisher, 1958), Stouffer's (Stouffer et al., 1949) and Pearson's (Pearson, 1933) methods. Fisher's method is based on the product of the p-values, Pearson's method on the product of one minus the p-values, Stouffer's method on the sum of the z-values, whereas (15) is based on the product of the z-values. We will compare the different methods to combine p-values in Sections 3.1 and 3.2 in more detail.

2.6 | Properties of the sceptical p-value

For fixed p_o and p_r, the nominal and golden versions of the sceptical p-value monotonically increase with increasing variance ratio c (Held, 2020a, Section 3.1) and eventually reach 0.5 for $c \to \infty$. The controlled sceptical p-value p_S^* behaves differently, as illustrated in Figure 3 (left), which shows p_S^* as a function of c for selected values of p_o and p_r.

As described in Section 2.5, p_S^* converges to $p_{\max}$ for $c \downarrow 0$. The functional behavior of p_S^* as a function of c can be studied through inspection of the derivative of p_S^* with respect to c, see Supporting Material E and F. If $p_o = p_r$ then p_S^* increases monotonically with increasing c. If the difference between p_o and p_r is relatively large, p_S^* decreases monotonically. If $|p_o - p_r|$ is relatively small but not zero, the sceptical p-value first decreases and then increases with increasing c. The infimum of p_S^* is hence $p_{\max}$ in the first case, p_∞ from (15) in the second, and in between those two values in the third case.

14679574, 2023, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/stan.12312 by Eth Zürich, Wiley Online Library on [17/10/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

These properties allow us to compare p_S^* and $p_{\max}$ for the same values of p_o and p_r, see Figure 3 (right). In the first case ($p_o = p_r$), p_S^* is always larger than $p_{\max}$, for any value of the variance ratio c. If p_o and p_r differ considerably, p_S^* is always smaller than $p_{\max}$ (the green region). The gray area depicts combinations of p_o and p_r for which the sceptical p-value p_S^* is not monotone as a function of c, but first decreases, then increases and eventually gets larger than $p_{\max}$ for large c. Whether p_S^* is smaller or larger than $p_{\max}$ now depends on the value of c, as indicated with dotted and dashed lines in Figure 3 (right) for $c = 0.1$ and $c = 1$, respectively.

3 | REPLICATION SUCCESS RATES AND REGIONS

In this section, we consider properties of the sceptical p-value based on the dichotomous criterion $p_S^* \leq \alpha$ for replication success. We start in Section 3.1 with deriving the corresponding value of the replication success level γ in (1) so that the overall T1E rate is exactly α^2 for a particular value of $c > 0$. As discussed in Section 2.1, the replication success level γ is also a bound on the partial T1E rate of the sceptical p-value. We will then compare success regions for different values of c in Section 3.2 and investigate project power in Section 3.3. Finally, Section 3.4 outlines how the sceptical p-value can be used for sample size calculations. This is facilitated through the interpretation of the variance ratio c as relative sample size, $c = n_r/n_o$, since the variances of the effect estimates are usually inversely proportional to the corresponding sample sizes n_o and n_r, that is, $\sigma_o^2 = \kappa^2/n_o$ and $\sigma_r^2 = \kappa^2/n_r$ for some unit variance κ^2 (Held, 2020a).

3.1 | Partial Type-I error control

If we want to know the bound on the partial T1E rate of the controlled sceptical p-value, we need to derive the corresponding value of γ in (1), now denoted as $\gamma_c = \gamma_c(\alpha)$ as it depends not only on the target overall T1E rate α^2, but also the relative sample size c. Comparing p_S to $\gamma_c(\alpha)$ is then equivalent to comparing p_S^* to α.

For $c = 1$, we have $\gamma_1 = 1 - \Phi(\Phi^{-1}(1 - 2\alpha^2)/2)$, see Held (2020b, Section 2.1). For example, $\gamma_1(\alpha = .025) = .065$, so the partial T1E rate is bounded by .065 for $c = 1$. The null distribution function (13) of z_S^2 can be used to compute the bound γ_c for $c \neq 1$, but now numerical methods are needed. Briefly, the overall T1E rate for any two values of c and γ_c can be computed with numerical integration (Held, Micheloud, & Pawel, 2022, Section 3.2). Root-finding methods are then used to find the value of γ_c which gives the target overall T1E rate of α^2. The inset plot in Figure 4 shows the bound γ_c as a function of c for exact overall T1E control at $\alpha^2 = .000625$. The bound on the partial T1E rate can get quite large for large relative sample sizes c. For example, it is $\gamma_{10}(\alpha = .025) = .14$ for $c = 10$. However, this large partial T1E rate is balanced by the fact that such large relative sample sizes are only needed for unconvincing original studies, where success is unlikely. The conditional T1E rate for such unconvincing original studies combined with large relative sample sizes is in fact very small, as shown in Section 3.4.

A summary of the overall and partial T1E rates of different methods is given in Table 1. Fisher's, Stouffer's, and Pearson's methods control the overall T1E rate exactly at significance level α^2. However, as the first two methods do not impose a threshold on the individual p-values p_o and p_r (Rosenkranz, 2002), the partial T1E rate is bounded by one and replication success can occur if one of the two p-values is very large. The partial T1E rate of Pearson's method is bounded by $c_P(\alpha) = 1 - \exp(-0.5\chi_4^2(\alpha^2))$, for example $c_P(\alpha = .025) = .035$.

14679574, 2023, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/stan.12312 by Eth Zürich, Wiley Online Library on [17/10/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

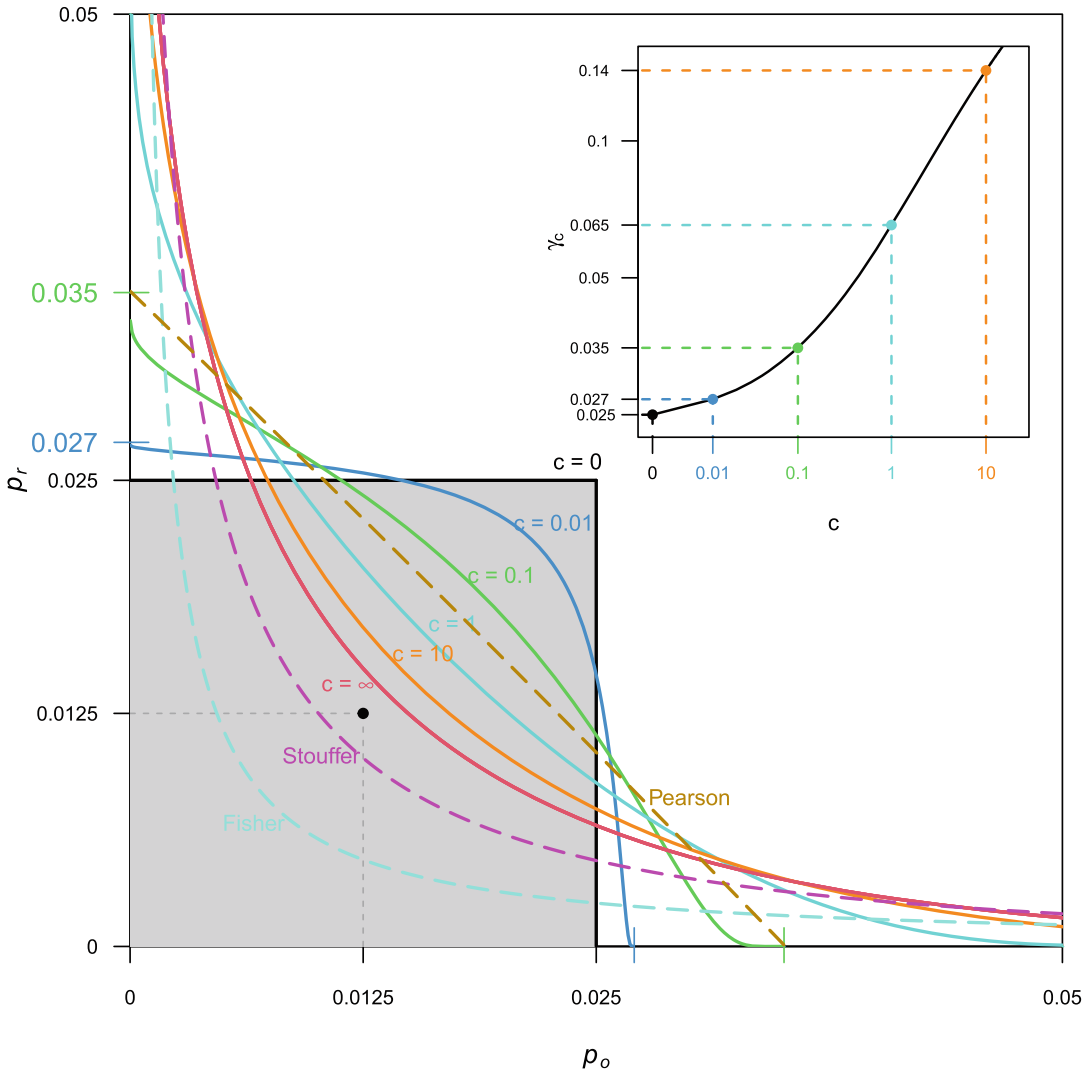

FIGURE 4 Success region of the sceptical p-value as a function of p_o and p_r for different values of the relative sample size c. The colored labels on the y-axis are the different values of the bound γ_c on the partial T1E rate ($\gamma_1 = .065$ and $\gamma_{10} = .14$ are outside the axis range, but can be read off the inset plot). The two-trials rule success region is the squared gray area below the black line where $c = 0$ and $\gamma_0 = \alpha$. Fisher's, Stouffer's, and Pearson's methods have been added for comparison purposes. All methods control the overall T1E rate at $\alpha^2 = .025^2 = .000625$, and so the area under each curve is equal to this value.

3.2 | Success regions

The (one-sided) framework (1) can now be used to determine the region where two p-values p_o and p_r lead to replication success. The main plot in Figure 4 compares the success regions for $\alpha = .025$ and selected values of the variance ratio (respectively relative sample size) c, so the area under each curve is equal to $\alpha^2 = .000625$. Each success region is bounded on both axes by the corresponding success level γ_c. The two-trials rule success region corresponds to $\gamma_0 = .025$ and is the squared gray area below the black line. The success region of the sceptical p-value is close to

14679574, 2023, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/stan.12312 by Eth Zürich, Wiley Online Library on [17/10/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 1 T1E rate of different methods to assess replication success.

		T1E control	
Method	**Parameters**	**Overall**	**Partial**
Two-trials rule	p_o, p_r	$= \alpha^2$	$< \alpha$
Fisher	p_o, p_r	$= \alpha^2$	< 1
Stouffer	p_o, p_r	$= \alpha^2$	< 1
Pearson	p_o, p_r	$= \alpha^2$	$< c_P(\alpha)$
Nominal p_S	p_o, p_r, c	$< \alpha^2$	$< \alpha$
Golden $\tilde{p}_S$	p_o, p_r, c	$< \gamma(\alpha)^2$	$< \gamma(\alpha)$
Controlled p_S^*	p_o, p_r, c	$= \alpha^2$	$< \gamma_c(\alpha)$

Note: The overall T1E rate is calculated under the intersection null (4). The partial T1E rate is calculated under the union null (5).

the two-trials rule's success region for small c, but becomes more and more in favor of p-values of different size as c increases. The case $c = \infty$ is based on the one-sided p-value p_∞ available from (15). Also shown are the success regions based on Fisher's, Stouffer's and Pearson's method. The first two are even less in favor of p-values of the same magnitude. For example, if both p-values are equal to $\alpha/2 = .0125$ (the solid black point in Figure 4), replication success will be flagged with the sceptical p-value for any value of c, but not with Fisher's nor Stouffer's method. Pearson's method has a success region similar to that of the sceptical p-value with $c = 0.1$, and the same bound on the partial T1E rate.

3.3 | Project power

Suppose none of the two studies has been conducted yet and so the probability to declare replication success is calculated over both studies in combination for a fixed relative sample size c. Using numerical integrations adapted from Held, Micheloud, & Pawel (2022, Section 3.3), the project power of the sceptical p-value is considered in this section.

The project power is the probability to declare replication success when both effects are equal and nonnull. The distribution of z_o is then $\mathrm{N}(\mu = z_\alpha + z_\beta, 1)$, where $1 - \beta$ is the power to detect the true original effect $\theta_o = \mu\ \sigma_o$ (Matthews, 2006, Section 3.3), and the distribution of z_r is $z_r \sim \mathrm{N}(\sqrt{c}\mu, 1)$. Figure 5 shows the project power of the sceptical p-value and the two-trials rule with $\alpha = .025$ and original power $1 - \beta = 80\%$ (left), respectively 90% (right). The project power based on the two-trials rule converges to 80%, respectively 90%, for large relative sample size c. The project power based on the sceptical p-value is always larger than with the two-trials rule, and increases to values close to 100% for large c. For example, for 80% original power and $c = 2$ the project power of the two trials rule is 78%, while the project power of the sceptical p-value is already 87%.

3.4 | Design of replication studies

In this section we assume that the original study has already been conducted and a replication study is planned. It has been recently argued that the method used for sample size planning

14679574, 2023, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/stan.12312 by Eth Zürich, Wiley Online Library on [17/10/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

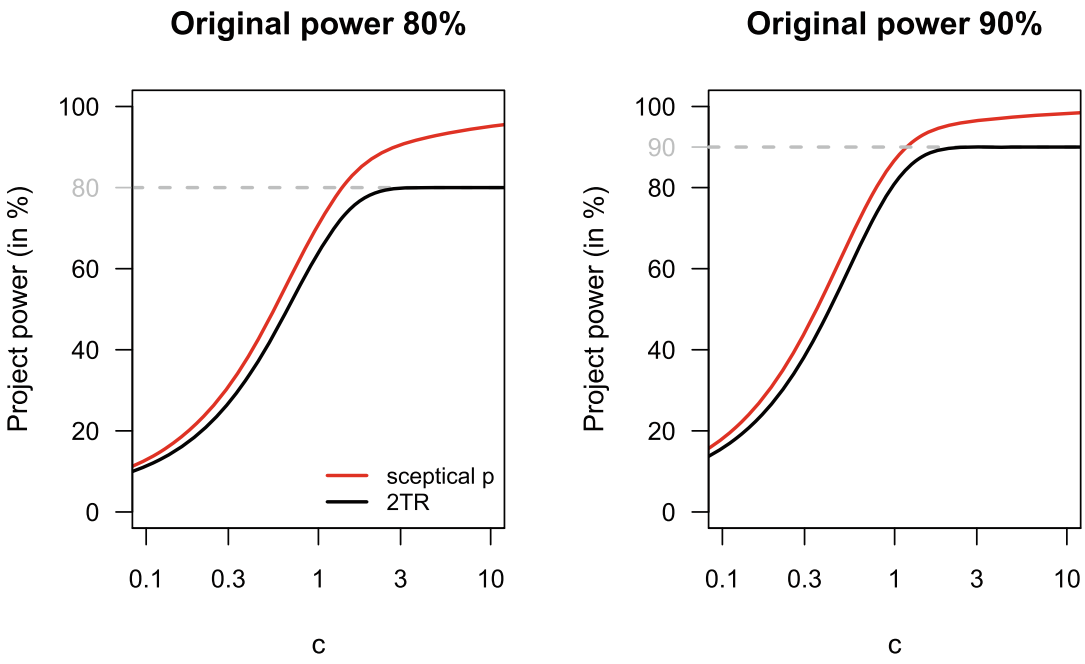

FIGURE 5 Project power as a function of the relative sample size c for $\alpha = .025$. The original power is 80% in the left plot, and 90% in the right one. Results are given for the sceptical p-value and compared with the two-trials rule (2TR).

should always match the one used for the analysis (Anderson & Kelley, 2022). Hence, we develop methods to design the replication study based on the sceptical p-value and compare it to the design based on the two-trials rule, that is, significance of the replication study.

The probability to declare replication success with a particular sample size n_r is known as the power of the replication study and is often calculated conditional on the effect estimate from the original study. Predictive power (Spiegelhalter & Freedman, 1986) can also be used and takes the uncertainty of the original effect estimate into account. Formulas for the power of the two-trials rule and the sceptical p-value at fixed success level can be found in Micheloud and Held (2022, Section 2.1) and Held, Micheloud, & Pawel (2022, Section 3.1), respectively.

Figure 6 (left) shows the ratio of conditional power calculated with the sceptical p-value versus the two-trials rule with $\alpha = .025$ as a function of the relative sample size c and the original p-value p_o. The sceptical p-value has larger power (ratio > 1) if the original study is already convincing. For example, for $c = 1$, this is the case if $p_o < .01$, otherwise the two-trials rule has larger power. However, if $p_o > \alpha$, the power of the two-trials rule is 0, but not the power of the sceptical p-value as long as $p_o < \gamma_c$.

Instead of calculating the power for a fixed replication sample size n_r, we can also fix the power to a desired value and calculate the required sample size n_r. Sample size calculation with the controlled sceptical p-value does not have a closed-form expression, because the success level γ_c in (1) depends on the relative sample size c. Root-finding algorithms are therefore required to find the value of c which leads to the desired power. Importantly, sample size calculation is now possible even for nonsignificant original studies, as shown in the top axis of Figure 7 for conditional power values of 80%, 90%, and 95%. Note that the required relative sample size can become quite large if $p_o > \alpha$.

14679574, 2023, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/stan.12312 by Eth Zürich, Wiley Online Library on [17/10/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

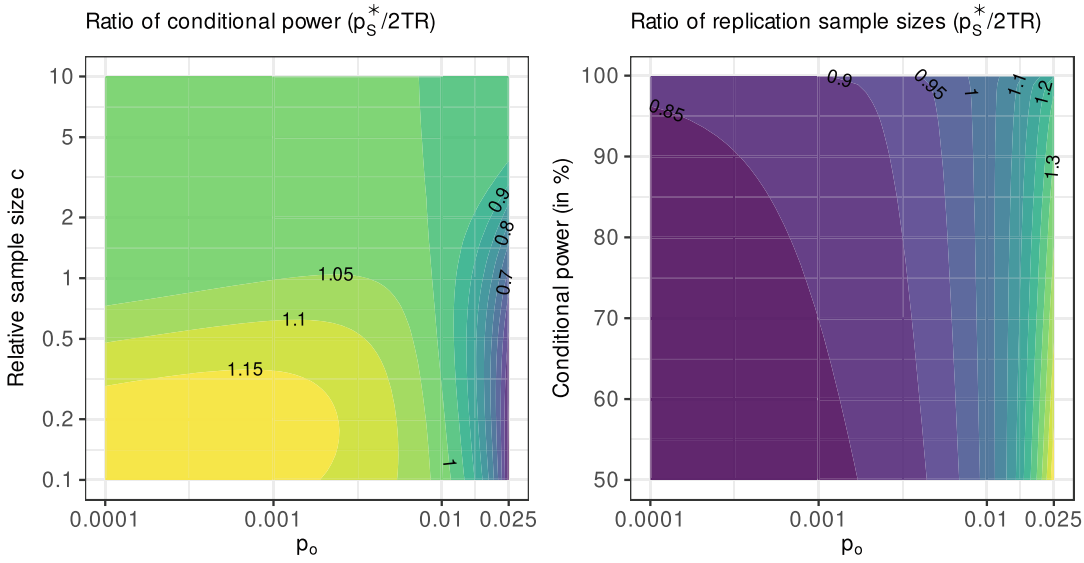

FIGURE 6 Ratio of conditional power (left) and replication sample sizes (right) calculated with the sceptical p-value (p_S^*) versus the two-trials rule (2TR) as a function of the original p-value p_o and the relative sample size c (left) or the conditional power (right) for $\alpha = .025$.

Figure 6 (right panel) shows the ratio of the replication sample sizes calculated with the sceptical p-value versus the two-trials rule. This is done only for significant original studies, as this is required for success with the two-trials rule. The sceptical p-value requires less samples than the two-trials rule for already convincing original studies ($p_o < .007$). Similar results are obtained when predictive power is used instead of conditional, see Supporting Material G.

The T1E rate of the sceptical p-value can also be considered under H_0^r, conditional on the original study result. First, the relative sample size c to reach a certain power for a fixed original z-value z_o is calculated. These values of z_o and c are then used in (1) to derive a lower bound for the replication z-value z_r to achieve replication success:

$$z_r \geq z_{\gamma_c}\sqrt{1 + c/(z_o^2/z_{\gamma_c}^2 - 1)}. \tag{16}$$

Subsequent transformation of the right hand side of (16) to the corresponding upper bound for p_r gives the conditional T1E rate. Note that the conditional T1E rate of the two-trials rule is constant at α as long as $p_o \leq \alpha$. Figure 7 shows the conditional T1E rate of the sceptical p-value as a function of the p-value p_o from the original study. The relative sample size c in (16) is calculated with the sceptical p-value method to reach a conditional power of 80%, 90%, and 95% respectively with $\gamma_c(\alpha = .025)$. The conditional T1E rate is larger than $\alpha = 2.5\%$, the conditional T1E rate of the two-trials rule, for $p_o < .008$ but bounded by 4.3%, 4.5% and 4.7% for a power of 80%, 90% and 95%, respectively. The conditional T1E rate is hence never larger than $2\alpha = 5\%$, and this also holds for other values of α provided that they are not very small, see Supporting Material G. If $p_o > .008$, the conditional T1E rate of the sceptical p-value is smaller than 2.5% in all three cases. This illustrates that the conditional T1E rate is sufficiently bounded if the replication sample size is computed based on standard power values to detect the observed effect from the original study.

14679574, 2023, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/stan.12312 by Eth Zürich, Wiley Online Library on [17/10/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

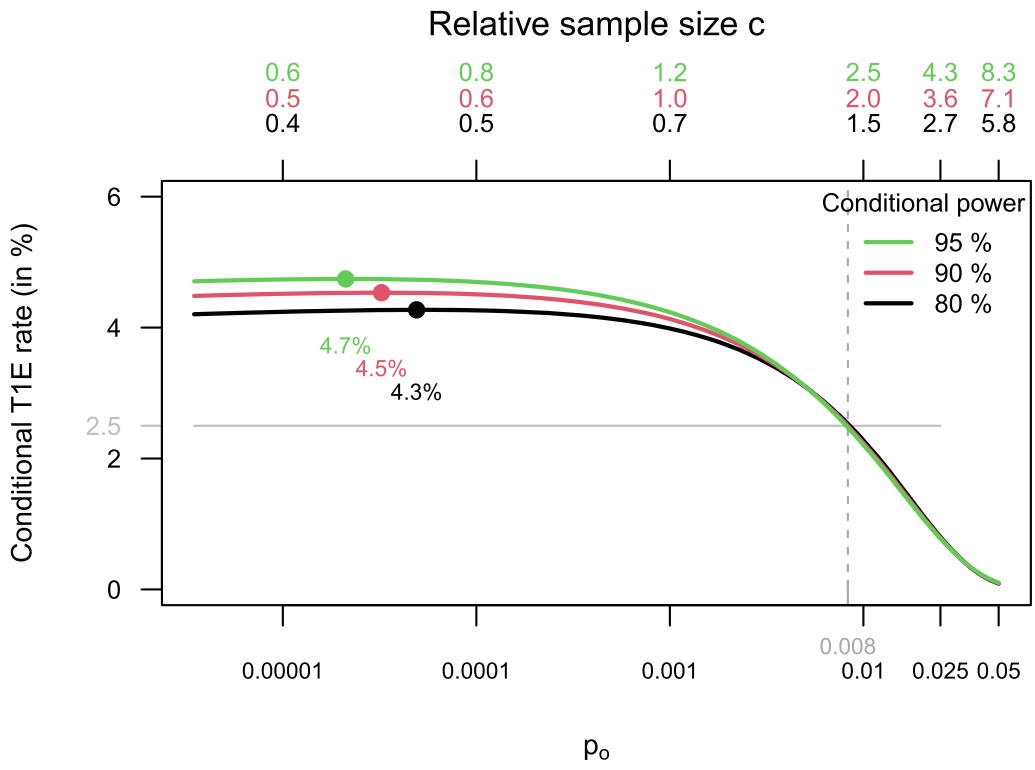

FIGURE 7 Conditional T1E rate of the sceptical p-value as a function of the original p-value p_o. The relative sample size (top axis) is calculated with the sceptical p-value method to reach a conditional power of 80%, 90%, and 95% at $\alpha = .025$. Each dot represents the upper bound for the conditional T1E rate with the respective power. The gray horizontal line indicates the T1E rate of the two-trials rule.

A similar pattern can be seen if the conditional T1E rate is based on predictive rather than conditional power, see Supporting Material G, where we also show the conditional T1E rate of the nominal and golden sceptical p-values.

4 | APPLICATION

We now illustrate the proposed methodology using all 18 pairs of studies from the Experimental Economics Replication Project (Camerer et al., 2016, EERP). The different effect estimates were all transformed to correlation coefficients, where Fisher's z-transformation achieves asymptotically normal effect estimates $\hat{\theta}_i$ with known standard errors. Table 2 summarizes the results for each of the 18 studies.

4.1 | Comparison of two-trials rule versus sceptical p-value

The last two columns in Table 2 show the p-value $p_{\max}$ from the two-trials rule and the sceptical p-value p_S^*, respectively. There is generally a good agreement, with $p_S^* < p_{\max}$ for 12 out of the 18 studies.

14679574, 2023, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/stan.12312 by Eth Zürich, Wiley Online Library on [17/10/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 2 Studies from the Experimental Economics Replication Project.

Study	$\hat{\theta}_o$	$\hat{\theta}_r$	p_o	p_r	Power (%)	c	$c^\star$	$p_{\max}$	p_S^*
de Clippel et al.	0.12	0.27	0.0005	<0.0001	91.0	1.0	0.8	0.0005	<0.0001
Kogan et al.	0.34	0.31	<0.0001	0.0005	93.9	0.7	0.6	0.0005	0.0002
Fudenberg et al.	0.31	0.34	0.0003	<0.0001	93.9	1.0	0.9	0.0003	0.0003
Dulleck et al.	0.91	0.93	<0.0001	0.0004	90.8	0.7	0.6	0.0004	0.0003
Friedman and Oprea	0.76	0.47	<0.0001	0.002	99.6	0.5	0.4	0.002	0.0003
Bartling et al.	0.91	0.79	0.003	0.0006	96.3	1.9	1.7	0.003	0.003
Kessler and Roth	0.53	0.36	<0.0001	0.008	94.6	0.2	0.1	0.008	0.003
Charness and Dufwenberg	0.40	0.38	0.005	0.001	89.0	1.6	1.5	0.005	0.005
Kirchler et al.	0.80	0.60	0.008	0.005	93.7	2.1	2.1	0.008	0.011
Fehr et al.	0.49	0.32	0.006	0.013	92.3	1.8	1.7	0.013	0.015
Ambrus and Greiner	0.32	0.23	0.027	0.006	93.3	3.2	4.2	0.027	0.024
Ericson and Fuster	0.22	0.12	0.015	0.027	92.3	2.4	2.7	0.027	0.032
Huck et al.	1.19	0.39	0.0002	0.082	99.1	1.4	1.2	0.082	0.045
Abeler et al.	0.18	0.08	0.023	0.08	90.7	2.7	3.4	0.08	0.074
Chen and Chen	1.23	0.17	0.017	0.28	98.3	3.7	4.1	0.28	0.24
Ifcher and Zarghamee	0.29	−0.01	0.016	0.53	90.7	2.3	2.6	0.53	0.53
Duffy and Puzzello	1.00	−0.12	0.007	0.66	95.0	2.2	2.1	0.66	0.69
Kuziemko et al.	0.29	−0.12	0.035	0.92	93.1	3.6	5.3	0.92	0.92

Note: Shown are the original and replication effect estimates ($\hat{\theta}_i$) and p-values (p_i), the power of the replication study with the actual relative sample size c, the corresponding relative sample size $c^\star$ calculated with the sceptical p-value method, $p_{\max}$ and p_S^*. The references of the studies are available in Camerer et al. (2016).

The studies of Ericson and Fuster (2011) and Ambrus and Greiner (2012) deserve closer scrutiny. While $p_{\max} = .027$ is the same for both studies, the values of p_S^* differ. In Ericson and Fuster (2011), $p_o = .015$ and $p_r = .027$ are relatively close to each other and the red line in Figure 3 (left panel) at $c = 1/2.4$ (because we have to reverse the role of p_o and p_r) explains why $p_S^* = .032 > p_{\max}$. In contrast, there is a larger difference in p-values ($p_o = .027$, $p_r = .006$) in Ambrus and Greiner (2012), and thus $p_S^* < p_{\max}$, see the green line in Figure 3 (left) at $c = 3.2$. For this combination of p-values, the sceptical p-value is smaller than $p_{\max}$ for every value of the relative sample size c. Of note, in Ambrus and Greiner (2012) the two-trials rule fails because of the dichotomization at $\alpha = .025$, but replication success with the sceptical p-value is achieved ($p_S^* = .024$).

4.2 | Sample size calculation

In the EERP, the replication sample sizes were calculated to reach "at least 90% power […] to detect the original effect size at the [two-sided] 5% significance level" (Camerer et al., 2016, p. 1434). We recomputed the exact power values to then calculate the required relative sample

14679574, 2023, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/stan.12312 by Eth Zürich, Wiley Online Library on [17/10/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

size $c^\star$ with the sceptical p-value (see Table 2). The sceptical p-value would require a smaller sample size than the two-trials rule for 12 out of 18 studies in this dataset. Two original studies were nonsignificant at the $\alpha = .025$ level (Ambrus & Greiner, 2012; Kuziemko, Buell, Reich, & Norton, 2014). Sample size calculation based on the sceptical p-value is also possible for those two studies, where we obtain $c^\star > c$. Note that for these two studies, the sample size c is calculated to achieve significance of the replication study, but the two-trials rule will fail anyway.

5 | DISCUSSION

We have described a novel statistical framework for the assessment of replicability, stemming from a recently proposed reverse-Bayes approach to assess replication success (Held, 2020a). The resulting controlled sceptical p-value p_S^* has exact overall T1E rate of α^2 and additionally ensures that the conditional and partial T1E rates are sufficiently bounded. The two-trials rule can be seen as a special case of the formulation for $c \downarrow 0$, where p_S^* converges to the maximum of the two study-specific p-values. The success region of the sceptical p-value is smooth, shifting gradually away from the squared one of the two-trials rule for increasing c, thus avoiding the "double dichotomization" and offering larger project power. Used in the design of the replication study, the new approach requires a smaller sample size than the two-trials rule for already convincing original studies. In contrast to the golden version, the controlled sceptical p-value allows sample size calculation for borderline significant and even nonsignificant original studies.

As the p-value $p = (p_S^*)^2$ is a proper p-value with exact linear T1E control under the intersection null hypothesis, a p-value function (Fraser, 2019) could be computed and a "sceptical" confidence interval could be calculated. This would address an important point raised by Diggle (2020) about the need to accompany the sceptical p-value with suitable estimation procedures to assess the relevance of the observed effects. We plan to consider this in future work.

However, exact overall T1E control comes at a certain price: the explicit penalization of small relative effect sizes in the nominal or golden versions of the sceptical p-value is lost and replication success may occur even for large shrinkage of the replication effect estimate, if the relative sample size c is large enough. Our conclusion is that exact overall T1E control and penalization of small effect sizes are two competing goals that cannot be achieved by a single criterion. It would therefore be interesting to extend the recently proposed dual-criterion for replication studies (Rosenkranz, 2021), which simultaneously requires significance and relevance, to the sceptical p-value.

This paper considers the situation where each original study only has one replication. Further work will extend the proposed methodology to the analysis of multiple replications per original study. A natural approach is to perform a meta-analysis of the replication studies and use the resulting combined effect estimate (possibly allowing for heterogeneity) as the replication effect estimate.

AUTHOR CONTRIBUTIONS

Charlotte Micheloud contributed substantially to research, analysis, coding, and writing. Fadoua Balabdaoui derived the required null distribution of the sceptical p-value, added further proofs and contributed to writing. Leonhard Held designed, performed and supervised research and analysis, wrote parts of the code and drafts of the paper.

14679574, 2023, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/stan.12312 by Eth Zü/Rich, Wiley Online Library on [17/10/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

ACKNOWLEDGMENT

LH thanks the University of Zurich for granting a sabbatical leave that made this research possible. CM and LH acknowledge support by the Swiss National Science Foundation (Project # 189295). We appreciate helpful comments by Rachel Heyard and Samuel Pawel. Open access funding provided by Universitat Zurich.

DATA AVAILABILITY STATEMENT

Software and data are available in the R-package `ReplicationSuccess` available from `CRAN`. The data is originally from https://osf.io/pnwuz/, see Pawel and Held (2020, supplement S1) for details on data preprocessing. The code to reproduce the analysis and figures is available at https://gitlab.uzh.ch/charlotte.micheloud/framework-for-replicability.

REFERENCES

Ambrus, A., & Greiner, B. (2012). Imperfect public monitoring with costly punishment: An experimental study. *American Economic Review*, *102*(7), 3317–3332.

Anderson, S. F., & Kelley, K. (2022). Sample size planning for replication studies: The devil is in the design. *Psychological Methods*. https://doi.org/10.1037/met0000520

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.

Camerer, C. F., Dreber, A., Forsell, E., Ho, T.-H., Huber, J., Johannesson, M., … Wu, H. (2016). Evaluating replicability of laboratory experiments in economics. *Science*, *351*(6280), 1433–1436.

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., Johannesson, M., … 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.

Casella, G., & Berger, R. L. (2002). *Statistical inference*. Belmont, CA: Duxbury Press.

Diggle, P. 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*, 450.

Ericson, K. M. M., & Fuster, A. (2011). Expectations as endowments: Evidence on reference-dependent preferences from exchange and valuation experiments. *The Quarterly Journal of Economics*, *126*(4), 1879–1907.

Errington, T. M., Mathur, M., Soderberg, C. K., Denis, A., Perfito, N., Iorns, E., & Nosek, B. A. (2021). Investigating the replicability of preclinical cancer biology. *eLife*, *10*, e71601.

Fisher, R. A. (1958). *Statistical methods for research workers* (13th ed.(rev.) ed.). Edinburgh: Oliver & Boyd.

Fraser, D. A. S. (2019). The *p*-value function and statistical inference. *The American Statistician*, *73*(supp 1), 135–147.

Grimmett, G. R., & Stirzaker, D. R. (2001). *Probability and random processes* (3rd ed.). Oxford, UK: Oxford University Press.

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*, *69*(3), 697–708.

Held, L., Matthews, R., Ott, M., & Pawel, S. (2022). Reverse-Bayes methods for evidence assessment and research synthesis. *Research Synthesis Methods*, *13*(3), 295–314.

Held, L., Micheloud, C., & Pawel, S. (2022). The assessment of replication success based on relative effect size. *The Annals of Applied Statistics*, *16*, 706–720.

Heller, R., Bogomolov, M., & Benjamini, Y. (2014). Deciding whether follow-up studies have replicated findings in a preliminary large-scale omics study. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(46), 16262–16267.

Kuziemko, I., Buell, R. W., Reich, T., & Norton, M. I. (2014). Last-place aversion: Evidence and redistributive implications. *The Quarterly Journal of Economics*, *129*(1), 105–149.

Ly, A., Etz, A., Marsman, M., & Wagenmakers, E.-J. (2019). Replication Bayes factors from evidence updating. *Behavior Research Methods*, *51*, 2498–2508.

14679574, 2023, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/stan.12312 by Eth Zuerich, Wiley Online Library on [17/10/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Maca, J., Gallo, P., Branson, M., & 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*. New York: Chapman and Hall/CRC.

Matthews, R. A. J. (2018). Beyond "significance": principles and practice of the analysis of credibility. *Royal Society Open Science*, *5*(1), 171047.

Micheloud, C., & Held, L. (2022). Power calculations for replication studies. *Statistical Science*, *37*(3), 369–379.

National Academies of Sciences, Engineering, and Medicine. (2019). *Reproducibility and replicability in science*. Washington, DC: The National Academies Press.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*, aac4716.

Pawel, S., & Held, L. (2020). Probabilistic forecasting of replication studies. *PLoS ONE*, *15*(4), e0231416.

Pawel, S., & Held, L. (2022). The sceptical Bayes factor for the assessment of replication success. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *84*(3), 879–911.

Pearson, K. (1933). On a method of determining whether a sample of size *n* supposed to have been drawn from a parent population having a known probability integral has probably been drawn at random. *Biometrika*, *25*(3-4), 379–410.

Pillai, N. S., & Meng, X.-L. (2016). An unexpected encounter with Cauchy and Lévy. *Annals of Statistics*, *44*, 2089–2097.

Rosenkranz, G. (2002). Is it possible to claim efficacy if one of two trials is significant while the other just shows a trend? *Drug Information Journal*, *36*(1), 875–879.

Rosenkranz, G. (2021). Replicability of studies following a dual-criterion design. *Statistics in Medicine*, *40*(18), 4068–4076.

Rosenkranz, G. (2023). A generalization of the two trials paradigm. *Therapeutic Innovation & Regulatory Science*, *57*(2), 316–320.

Royal Netherlands Academy of Arts and Science. (2018). *Replication studies—Improving reproducibility in the empirical science*. Amsterdam: KNAW.

Senn, S. (2007). *Statistical issues in drug development* (second ed.). Chichester, U.K: John Wiley & Sons.

Spiegelhalter, D. J., & Freedman, L. S. (1986). A predictive approach to selecting the size of a clinical trial, based on subjective clinical opinion. *Statistics in Medicine*, *5*(1), 1–13.

Stouffer, S. A., Suchman, E. A., Devinney, L. C., Star, S. A., & Williams, R. M., Jr. (1949). *The American soldier: Adjustment during army life. (Studies in social psychology in World War II)*. Princeton: Cambridge University Press, Princeton University Press.

Zhan, S. J., Kunz, C. U., & Stallard, N. (2023). Should the two-trial paradigm still be the gold standard in drug assessment? *Pharmaceutical Statistics*, *22*(1), 96–111.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Micheloud, C., Balabdaoui, F., & Held, L. (2023). Assessing replicability with the sceptical *p*-value: Type-I error control and sample size planning. *Statistica Neerlandica*, *77*(4), 573–591. https://doi.org/10.1111/stan.12312

14679574, 2023, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/stan.12312 by Eth Zürich, Wiley Online Library on [17/10/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License