Statistical Science
2023, Vol. 38, No. 4, 621–639
https://doi.org/10.1214/23-STS904

Replication Success Under Questionable Research Practices—a Simulation Study

Francesca Freuli, Leonhard Held and Rachel Heyard

Abstract. Increasing evidence suggests that the reproducibility and replicability of scientific findings is threatened by researchers employing questionable research practices (QRPs) in order to achieve statistically significant results. Numerous metrics have been developed to determine replication success but it has not yet been investigated how well those metrics perform in the presence of QRPs. This paper aims to compare the performance of different metrics quantifying replication success in the presence of four types of QRPs: cherry picking of outcomes, questionable interim analyses, questionable inclusion of covariates, and questionable subgroup analyses. Our results show that the metric based on the version of the sceptical p-value that is recalibrated in terms of effect size performs better in maintaining low values of overall type-I error rate, but often requires larger replication sample sizes compared to metrics based on significance, the controlled version of the sceptical p-value, meta-analysis or Bayes factors, especially when severe QRPs are employed.

Key words and phrases: Questionable research practices, replication success, simulation study, type-I error rate, power, rejection ratio.

1. INTRODUCTION

Large-scale replication projects in psychology, cancer biology and other fields continue to report low replicability rates (Open Science Collaboration, 2015, Errington et al., 2021). A possible reason for these low rates is that the original studies that were replicated have been impacted by so-called questionable research practices (QRPs). Researchers may engage in QRPs to increase their chance of achieving a positive result which, in return, increases the chance of getting their results published (Simmons, Nelson and Simonsohn, 2011, Nosek, Spies and Motyl, 2012). Examples of QRPs are manifold and they differ depending on which "researcher degrees of freedom" (Wicherts et al., 2016) were exploited in order to obtain statistically significant results. It has been well documented that such practices can increase the probability of false positive results substantially, potentially making them unreliable (Simmons, Nelson and Simonsohn, 2011, Roettger, 2019). The success of a replication of an original study with suspected QRPs might therefore be compromised, especially since QRPs are likely not recorded nor reported. There is evidence suggesting that QRPs are frequently employed. Between 39% and 51% of researchers admit already having applied at least one of them (Wolff, Baumann and Englert, 2018, Gopalakrishna et al., 2022), while they are often not aware that such practices are problematic (Bishop, 2019, Rabelo et al., 2020). Some recent studies showed that young researchers and students had applied QRPs because they received pressure from their supervisors (Moran et al., 2022, Christian et al., 2021).

Replications are essential to establish the validity of research findings. We define a scientific finding as *replicable* if consistent results are obtained across studies aimed at answering the same scientific question, each of which collecting its own data (National Academies of Sciences, Engineering, Medicine, 2019). Schmidt (2009) differentiates between direct and conceptual replications. We will focus on direct replications, as conceptual replications go further in testing hypotheses in a slightly different experimental set-up. As replications of scientific studies are becoming more and more common, metrics to assess

Francesca Freuli is Ph.D. Student, Department of Psychology and Cognitive Science, University of Trento, Trento, Italy (e-mail: francesca.freuli@unitn.it). Leonhard Held is Professor, Center for Reproducible Science, Department of Biostatistics, Epidemiology, Biostatistics and Prevention Institute, University of Zurich, Zurich, Switzerland (e-mail: leonhard.held@uzh.ch). Rachel Heyard is Postdoctoral Fellow, Center for Reproducible Science, Department of Biostatistics, Epidemiology, Biostatistics and Prevention Institute, University of Zurich, Zurich, Switzerland (e-mail: rachel.heyard@uzh.ch).

whether a replication was successful started to emerge (Anderson and Maxwell, 2016). There is no universally agreed-up on criterion for replication success. Therefore, most large replication projects used a whole set of metrics. The Reproduciblity Project Psychology (Open Science Collaboration, 2015), for example, used significance and p-values, effect sizes, subjective assessment of replication teams, and meta-analyses of effect sizes to evaluate the replicability of the original studies.

Using standard significance as an indicator for replication success, that is, declaring a replication successful if both the original and the replication studies yield a significant result (in the same direction), has long been custom in drug development where it is referred to as the "two-trials rule" (Senn, 2021). This criterion however ignores the effect size of the original and the replication studies and has other shortcomings (Simonsohn, 2015). In contrast, the Q-test assesses compatibility of the original and replication effect sizes without considering the corresponding p-values (Hedges and Schauer, 2019). Meta-analytic approaches use the effect sizes and their uncertainty of the original and the replication studies and summarise them into an overall effect size estimate. An often discussed shortcoming of meta-analytic approaches is that they ignore the successive nature of original and replication studies. A more recently developed metric, the sceptical p-value (Held, 2020, Held, Micheloud and Pawel, 2022) combines significance of the original and replication studies together with their effect sizes. It is important to note that all these metrics address slightly different aspects of replicability (Anderson and Maxwell, 2016).

In an attempt to find the best metric to quantify replication success in terms of frequentist operating characteristics, Muradchanian et al. (2021) conducted a simulation study to compare the performance of a variety of metrics in the presence of different levels of publication bias. The authors compared standard replication success metrics based on statistical significance or meta-analysis with more recently developed approaches, like the Small Telescopes by Simonsohn (2015) or the sceptical p-value and Bayesian approaches (as described in Verhagen and Wagenmakers (2014)). There was no single metric which performed best for all levels of publication bias, while the sceptical p-value and the Bayes factor approach slightly outperformed the more standard frequentist metrics, that is, meta-analysis and standard significance.

Little is known on how the different replication success metrics behave in the presence of QRPs. As the list of potential QRPs is long, we focus on a subset that are often referred to as "p-hacking", defined as "any measure that a researcher applies to render a previously nonsignificant p-value significant" (Stefan and Schönbrodt, 2023). For the present simulation study, we took inspiration from the QRPs considered in Simmons, Nelson and Simonsohn (2011). We consider the following:

- We simulate a specific form of outcome reporting bias (Kirkham et al., 2010, Kirkham et al., 2018) assuming that a researcher considers several outcomes for the same research hypothesis and only reports the outcome with the most favorable result, defined as the outcome yielding the smallest p-value. This QRP is a common form of p-hacking (Head et al., 2015, Moran et al., 2022) and has been referred to as *cherry picking* of outcomes (Mayo-Wilson et al., 2017).
- We simulate *questionable interim analyses* (Pocock, 1977, Sagarin, Ambler and Lee, 2014), where the researcher performs multiple statistical analyses during the data collection phase and stops adding new observations once a statistically significant result is observed. More than half of the researchers participating in surveys declared to have collected more data after checking for significance of results (John, Loewenstein and Prelec, 2012, Agnoli et al., 2017).
- We simulate a QRP where different covariates are added one-by-one to a regression model in order to find a significant effect (Wicherts et al., 2016, Wang et al., 2017). We will refer to this QRP as *questionable inclusion of covariates*.

These three first QRPs are directly derived from Simmons, Nelson and Simonsohn (2011), while we decided against simulating their fourth QRP the flexible reporting of subsets of experimental conditions. This QRP is difficult to simulate under the alternative hypothesis as it requires specification of several effect sizes, not just one. Indeed, if we wanted to simulate three conditions (e.g., high, medium and low) under the alternative, we would need to specify the effect estimate of two comparisons of conditions (e.g., high vs. low and medium vs. low). As this quickly becomes intricate with more conditions we decided to simulate an alternative, but closely connected QRP:

- We refer to this alternative QRP as *questionable subgroup analyses* (Rosenkranz, 2019, Chapter 3), where binary characteristics are used instead of experimental conditions. In this setting we assume that multiple subgroup analyses are performed based on binary characteristics (gender, seniority, ...) and only the most favorable result is reported, defined as the subgroup yielding the smallest p-value (Brookes et al., 2004). The frequency of questionable subgroup analyses has not yet been directly investigated, but the multiplicity problem inherent in subgroup analyses has often been described (Matthews, 2006, Chapter 9).

Even if the positive effect of QRPs on type-I error (T1E) rate, that is, the false positive rate, has already been intensively investigated (Simmons, Nelson and Simonsohn, 2011, Nosek, Spies and Motyl, 2012, Roettger, 2019), their influence on replication success has not. In

a simulation study, Ulrich and Miller (2020) investigated the effect of p-hacking on the probability of a successful replication using only the two-trials rule as a metric. Their general conclusion suggested that the effects of p-hacking "are unlikely to be massive". The aim of the simulation study presented in this paper is to study the characteristics of different replication success metrics when QRPs are suspected to be present in the original study. The metrics used are described in detail in Section 2.1. The design of the simulation study is outlined in Section 2.2 with separate sections for the original studies with QRP in Section 2.2.1 and the replication studies in Section 2.2.2. In order to assess how well the metrics perform, we need clear measures of comparison which are defined in Section 2.3. The results are outlined in Section 3 and the paper ends with a discussion of our results, our study's limitations and some recommendations.

2. METHODS

While planning our simulation study, we followed the recommendations outlined in Morris, White and Crowther (2019). We wrote a simulation study protocol which we preregistered on the Open Science Framework (doi: 10.17605/OSF.IO/YDBSH) before writing the code as suggested in Burton et al. (2006). The next sections will reiterate the most important steps of the methodology used, while we refer to the protocol for more details. We only consider continuous outcomes in the simulation of all the scenarios and apply throughout one-sided (one- or two-sample) t-tests which take into account the direction of the effect estimate.

2.1 Metrics for Replication Success

We will now introduce and define the five replication success metrics to be compared in detail. The standard significance level for one-sided hypothesis testing in drug development of $\alpha = 0.025$ (Senn, 2021, Chapter 12.2.5) will be used.

- The first metric is the two-trials rule, which is based on standard statistical significance. The two-trials rule has long been custom in drug development where a drug's efficiency needs to be demonstrated in two independent trials (Senn, 2021). According to the two-trials rule, a replication is marked as successful if the replication shows a statistically significant effect in the same direction as the significant original study. Let us assume that p_o refers to the one-sided p-value in the original study and p_r is the corresponding replication p-value; then, the two-trials rule marks a replication as successful if

$$\max\{p_o, p_r\} < \alpha = 0.025.$$

  Thresholding both, original and replication, p-values at α ensures that we control the overall T1E rate (*i.e.*, rate of false positive replication success) at α^2, and this is what we also try to achieve with alternative metrics (Rosenkranz, 2023).
- The second metric to quantify replication success is based on meta-analysis. According to the meta-analytical metric, a replication is successful if the effect estimate of a fixed effects meta-analysis combining the original and the replication study is significant in the anticipated direction, at a one-sided significance level α^2. If p_{MA} is the meta-analytical one-sided p-value, we flag replication success if

$$p_{\text{MA}} < \alpha^2 = 0.025^2 = 0.000625.$$

  The α^2 threshold ensures the same overall T1E control as for the two-trials rule. Original and replication studies are assumed to be exchangeable. Stouffer's method which is based on the Z-scores[1] of the original and replication study, is used to compute the meta-analytical p-values. We will use the weighted version of Stouffer's method as described in Cousins (2007) with weights $w_o = 1$ for the original and $w_r = \sigma_o/\sigma_r$ for the replication study, where σ_o and σ_r are the standard errors of the effect in the original and replication study, respectively. The resulting p-value is related to the "pooled trials rule" which is equivalent to investigating whether the overall effect of a fixed-effect meta-analysis is significant (Senn, 2021, Section 12.2.8).
- The next two metrics are based on the sceptical p-value, a method that combines a reverse-Bayes approach with a prior-data conflict assessment (Held, 2020, Held et al., 2022). Using the data from the original study, the method first determines a so-called sceptical prior for the underlying effect size which would deem the resulting posterior no longer significant. The sceptical p-value p_{S} then quantifies the conflict between the replication data and the sceptical prior. Replication success is achieved if

$$p_{\text{S}} < \alpha = 0.025.$$

  A necessary but not sufficient condition for success of the original proposal by Held (2020) is significance of both the original and the replication study. The original sceptical p-value will therefore flag success less often than the two-trials rule and has a smaller overall T1E rate. In the following, we will consider two recalibrations that have been proposed to make the sceptical p-value less stringent:
  - The golden sceptical p-value, our third metric, is based on a recalibration to ensure that replication success of borderline significant original studies

[1] The Z-score of study i is computed through $Z_i = \Phi^{-1}(1 - p_i)$ where Φ^{-1} is the quantile function of the standard normal distribution and p_i is the p-value.

($p_o \approx \alpha$) is possible, but only if there is no shrinkage of effect size (Held, Micheloud and Pawel, 2022). We have shrinkage whenever the replication effect estimate is smaller than the original effect estimate. The golden sceptical p-value has overall T1E rate smaller than α^2 as long as the variance ratio is larger than 1.
  - The controlled sceptical p-value, our fourth method, is a recently proposed extension that guarantees exact overall T1E control at level α^2, the overall T1E rate of the two-trials rule (Micheloud, Balabdaoui and Held, 2023).
- Our last and fifth metric to assess replication success is the replication Bayes factor (BF) (Verhagen and Wagenmakers, 2014). A BF quantifies the evidence for the null versus the alternative hypothesis (BF_{01}). The complete BF is based on both the original and replication data and can be written as the product of the replication BF and the BF based on the original data only (Ly et al., 2018). Exact overall T1E control of a Bayesian procedure is difficult (Grieve, 2016), but a suitable threshold for the replication BF can be obtained by a transformation of the one-sided 0.025 p-value threshold to the BF scale. To do so we utilize Equation 11 in Held and Ott (2018) to obtain the corresponding Bayes factor threshold $1/3.989 \approx 1/4$, assuming that the power to detect the original effect estimate was chosen to be 85%, the value we use in our simulation study. Replication success with the replication Bayes factor is achieved if

$$\text{replication } \text{BF}_{01} < 1/4.$$

Note that the overall T1E rate control of the metrics based on p-values is only valid in the absence of QRPs and publication bias. The two-trials rule represents our benchmark as it is the approach most commonly used in large reproducibility projects (e.g., Open Science Collaboration, 2015). Meta-analytical approaches have been reported to outperform this commonly used method in terms of frequentist operating characteristics, while in the presence of publication bias the sceptical p-value was found to perform particularly well (Muradchanian et al., 2021). As also other Bayesian metrics were found to perform well in Muradchanian et al. (2021), we additionally added the replication BF as a Bayesian success metric.

2.2 Design of the Simulation Study

Before describing the simulation of each QRP in detail, we introduce some common choices and parameters. We consider different levels of severity $k \in \{0, \ldots, 9\}$ for each QRP. This level of severity is interpreted differently depending on the QRP. Level $k = 0$ represents the absence of QRP. Original studies are reported (i.e., published) only if they yield a positive and significant result, leading to 100% publication bias. Replication studies are simulated based on the published original results, but they themselves do not include any QRPs as replication studies tend to be preregistered and conducted more rigorously. We simulate under both hypotheses, the null (H_0) and the alternative (H_1). The effect size under the alternative is fixed to $\theta = 0.34$ to achieve a power of $1 - \beta = 85\%$ with a sample size of $n_o = 80$ in the original study with a one-sample t-test, and $n_o = 157$ per group if a two-sample test is used. Under the null hypothesis of no effect we have $\theta = 0$. As outlined in our protocol, to ensure that the Monte Carlo error of our proportions of interest stays below 0.5%, the number of simulations of original studies was set to $N = 400{,}000$ among which $N \cdot \alpha = 10{,}000$ would result in a significant result and a replication. The simulation procedure includes five main steps: simulation of the original study, extraction of the significant results, estimation of the replication study sample size (based on the published original results), simulation of the replication study, and estimation of the rates of replication success using the metrics described above.

2.2.1 *Simulating original studies with QRPs*. The QRPs considered and described in the following were simulated separately without any combinations of QRPs.

Cherry picking. We simulate this QRP for each $k \in \{0, \ldots, 9\}$, where k represents the number of additional outcomes that are analysed, additional to the first one. We draw, for each individual $i \in \{1, \ldots, n_o\}$, a set of $k + 1$ outcomes from a multivariate normal distribution with mean $\boldsymbol{\theta}$ and correlation matrix $\boldsymbol{\Sigma}$ of size $(k+1) \times (k+1)$. The correlation matrix has standard deviation 1 on the diagonal and correlation $\rho = 0.5$ on the off-diagonal following Simmons, Nelson and Simonsohn (2011) and $\boldsymbol{\theta}$ is a vector of length $k + 1$ with elements θ. Let us assume $\mathbf{Y}$ represents the simulated data set, then

$$\mathbf{Y} = \begin{bmatrix} \mathbf{y}_1^{\mathrm{T}} \\ \vdots \\ \mathbf{y}_{n_o}^{\mathrm{T}} \end{bmatrix} = \begin{bmatrix} y_{1,1} & \cdots & y_{1,k+1} \\ \vdots & & \vdots \\ y_{n_o,1} & \cdots & y_{n_o,k+1} \end{bmatrix},$$

where

$$\mathbf{y}_i \sim N_{k+1}(\boldsymbol{\theta}, \boldsymbol{\Sigma}).$$

Note that the rows $\mathbf{y}_i^{\mathrm{T}}$ of $\mathbf{Y}$ are independent and identically distributed (i.i.d.). Next, a one-sample t-test is applied on each of the $k + 1$ columns and $k + 1$ one-sided p-values $p_0, \ldots, p_k$ are retained. A researcher practicing cherry picking reports only the smallest p-value as the p-value of the original study:

$$p_o = \min\{p_0, \ldots, p_k\}.$$

No multiplicity correction is performed. It is further assumed that only those simulated studies indicating a significant positive effect with $p_o < \alpha$ are published and will be replicated. We will simulate data for all $k = 0, \ldots, 9$ severity levels, for H_0 where $\boldsymbol{\theta} = \mathbf{0}$ and for H_1 where $\boldsymbol{\theta} = (\theta, \ldots, \theta)^{\mathrm{T}}$.

Questionable interim analyses. For this QRP, we assume that the researchers planned to recruit $n_o = 80$ individuals for their original study. However, for a specific $k \in \{0, \ldots, 9\}$, they decide to do k unplanned and therefore questionable interim analyses. The number of new participants per interim analysis is defined as $m = n_o/(k+1)$. A noninteger value of m is rounded up for a suitable number of interim analyses while rounded down for the remaining ones to ensure that the total sample size is still n_o (see the online protocol for more details). To simulate questionable interim analyses with $k \geq 1$, we first draw a sample $\mathbf{y}_1 = (y_1, \ldots, y_m)^{\mathrm{T}}$ from a normal distribution with mean θ and variance 1, $\mathbf{y}_1 \sim N_m(\theta, 1)$. We now assume that the researcher tests for a positive effect using a one-sample t-test leading to a one-sided p-value p_1. A significant result with $p_1 < \alpha$ leads to a replication study and we move to the next simulation. Otherwise and $k \geq 2$, we assume that m more individuals are recruited and simulate $\mathbf{y}_2 = (y_{m+1}, \ldots, y_{2m})^{\mathrm{T}} \sim N_m(\theta, 1)$. The next p-value p_2 is achieved through a t-test performed on the combination of both samples $(\mathbf{y}_1, \mathbf{y}_2)$ with sample size $2m$. If the null hypothesis is rejected at this stage, a replication is designed and performed based on the published original study of sample size $2m$. Otherwise a next sample of size m is drawn until either a significant result is observed or the total sample size reaches the maximum n_o. Note that we again simulate data for all $k = 0, \ldots, 9$ levels of severity and both hypotheses, H_0 with $\theta = 0$ and H_1 with $\theta = 0.34$.

Questionable inclusion of covariates. To simulate questionable inclusion of covariates we need to consider two samples (Simmons, Nelson and Simonsohn, 2011, Roettger, 2019), for example, two different treatment groups. Therefore, a larger original sample size is required to achieve the same power of 85% given an effect size of $\theta = 0.34$ under the alternative. For both (treatment) groups, we simulate two data matrices, $\mathbf{Y}_a$ and $\mathbf{Y}_b$, each with $n_o^a = n_o^b = 157$ rows (observations) and $k+1$ columns. The first column represents the outcomes $\mathbf{y}_a$ and $\mathbf{y}_b$ respectively, and the remaining k columns represent the covariates which are assumed to be unrelated to the outcome and treatment. $\mathbf{Y}_a$ and $\mathbf{Y}_b$ are drawn from a multivariate normal distribution with respective means $\boldsymbol{\theta}_a$ and $\boldsymbol{\theta}_b$ and correlation matrix $\boldsymbol{\Sigma}$ of size $(k+1) \times (k+1)$ (with standard deviation 1 on the diagonal and correlation $\rho = 0.5$ on the off-diagonal). Under the null hypothesis, $\theta_a = \theta_b = 0$ and the means of the distributions are defined as $\boldsymbol{\theta}_a = \boldsymbol{\theta}_b = (\theta_a, \mathbf{0}) = (0, \mathbf{0})$ (where $\mathbf{0}$ is a vector of size $k-1$). Under the alternative, we have $\theta_a = 0$ and $\theta_b = 0.34$. The mean for $\mathbf{Y}_a$ is $\boldsymbol{\theta}_a = (\theta_a, \mathbf{0}) = (0, \mathbf{0})$ and the mean for $\mathbf{Y}_b$ is $\boldsymbol{\theta}_b = (\theta_b, \mathbf{0}) = (0.34, \mathbf{0})$. To obtain a set of k binary covariates, the negative elements of the covariate columns will be transformed to 0, and the positive element will be transformed to 1. Note that we test the one-sided alternative hypothesis $H_1: \theta_b > \theta_a$.

We now follow Wang et al. (2017) and assume that the researcher wants to test for a positive treatment effect on the outcome $\mathbf{y} = (\mathbf{y}_a, \mathbf{y}_b)$ and applies a simple linear model with the treatment indicator as sole independent variable, without any additional covariates. This result yields a first one-sided p-value p_0 for the treatment effect. If $p_0 < \alpha$ the researcher publishes the significant result as such and a replication study can be designed and performed. Otherwise and if $k \geq 1$, k covariates are added to the model in a sequential way. Every time a new covariate is added to the model the researcher assesses whether the resulting p-value for the treatment effect is smaller than α. If yes those results are published. Otherwise another covariate is added. This practice is repeated until a significant treatment effect can be reported or all k covariates are included in the final model. The data are simulated for each $k = 0, \ldots, 9$ and both hypotheses, H_0 and H_1.

Questionable subgroup analyses. To simulate this practice for each $k = 0, \ldots 9$ under H_0 and H_1, we draw a data matrix $\mathbf{Y}$ with $n_o = 80$ rows and $k+1$ columns from a multivariate normal distribution with mean $\boldsymbol{\theta}$ and correlation matrix $\boldsymbol{\Sigma}$ with standard deviation 1 on the diagonal and correlation $\rho = 0$ on the off-diagonal (the columns of the matrix will not be correlated). As for questionable inclusion of covariates, the first column of $\mathbf{Y}$ represents the outcome $\mathbf{y}$ and the remaining k columns represent the covariates used for subgroup splitting. Under the null hypothesis we have $\boldsymbol{\theta} = \mathbf{0}$ and under H_1 we have $\boldsymbol{\theta} = (0.34, \mathbf{0})$. First, a one-sample t-test is applied on the outcome $\mathbf{y}$, resulting in a first one-sided p-value p_0. A significant result leads to a replication study. Without a significant result and for $k \geq 1$, the outcome $\mathbf{y}$, will be randomly split k times, following the sign of the k covariates obtaining $2k$ subgroups. For instance, if $k = 3$, $\mathbf{y}$ is randomly split two times and we obtain $2(k-1) = 4$ subgroups. Each of the subgroups might have a different sample size m_{s_j} with $j = 1, \ldots, 4$. Note that the sample is not split into four parts, but two times into two parts. As an example, we can imagine that a researcher used two binary covariates, gender and age (young vs. old), and first considers gender (men vs. women) to split the sample and then age (young vs. old). Each subgroup is analyzed separately with a one-sample t-test resulting in $2k$ one-sided p-values (e.g., a one-sample t-test is applied on all the women, on all the men, on all the young participants, and on all the old participants, respectively). If the lowest p-value is smaller than α, only the results of the subgroup with lowest p-value would be published, and a replication study can be designed and conducted.

2.2.2 *Planning and simulating replication studies without QRPs*. Whenever the simulated original study with QRP yields a significant result, a replication study is designed based on the published original results. The published sample size n_o' depends on the QRP investigated

and, in the presence of cherry picking and questionable inclusion of covariates, is simply the original sample size n_o regardless of which level of severity k was employed. For the other QRPs the published sample size $n'_o \leq n_o$ depends on which level yielded a significant result. For questionable interim analyses, if a significant result was found after the jth interim analysis, then $n'_o = j \cdot m$. For questionable subgroup analyses the published sample size is n_o if a significant result was found on the whole sample, and m_{s_j} if the smallest significant p-value is observed in subgroup j with sample size m_{s_j}.

For each significant original study, a sample size calculation will be performed to define the relative sample size $c = n_r/n'_o$, where n_r is the sample size of the replication study. In particular, the replication of each significant original study is designed in two ways: c will either be fixed at $c = 1$ as this is what replication researchers have intuitively been doing, or chosen adaptively based on the original study result and the designated replication success metric. The adaptive sample size calculation is of particular interest as the design of replication studies should ideally match the type of analysis (Anderson and Kelley, 2022). Specifically, assuming that the reported effect size in the original study is the true effect size, we will compute:

- the required relative sample size to achieve a significant positive effect in the replication study,
- the required relative sample size to obtain a meta-analytical p-value $p_{\text{MA}} < \alpha^2$,
- the required relative sample size to achieve replication success according to the golden sceptical p-value,
- the required relative sample size to achieve replication success according to the controlled sceptical p-value,
- and the required relative sample size to obtain a replication Bayes factor $< 1/4$.

All are based on standard normality assumptions aiming to achieve a power of 85% to detect the published effect estimate from the original study. Further details on the different sample size calculations are described in the relevant literature (Micheloud and Held, 2022, Held, 2020, Held, Micheloud and Pawel, 2022, Micheloud, Balabdaoui and Held, 2023, Pawel, Consonni and Held, 2023). When the meta-analytic metric for replication success is used a very convincing original study with p_o much smaller than α^2 would almost certainly lead to success regardless of the results of the replication study (Micheloud and Held, 2022, Section 2.2.3) and the replication study would actually not be required. Therefore, one would expect the relative sample size to converge to 0 with decreasing p_o. In this case, to ensure large enough replication sample size in order to perform the required tests, we fix c to 0.1 whenever $p_o < 0.0037$, which is the original p-value for which the required relative sample size to obtain $p_{\text{MA}} < \alpha^2$ is $c = 0.1$. Further, since the relative sample size c might be noninteger valued, the resulting replication sample size $n_r = c \cdot n'_o$ is rounded to the next integer. We further include an upper bound of $c \leq 100$ to ensure the replication study does not become unpractically large and a lower bound of $n_r \geq 2$, as otherwise no tests can be performed. A relative sample size of $c = 100$ may not be feasible in most disciplines, although a sample size increase by almost 20 ($n_o = 75$, $n_r = 1488$) has been used in one of the replication studies in Open Science Collaboration (2015).

The replication study is simulated following the same procedure as for the original studies with $k = 0$. However, for original studies with questionable inclusion of covariates the same number of covariates as reported in the original study is used because the replication authors would exactly follow the reported original studies. Figure 1 represents a diagram of the design of the simulation study.

2.3 Measures of Comparison

For each QRP, each level k and each design for the relative sample size, we compute the average relative effect size (under H_1), defined as the average of the ratio between effect estimate of the replication studies and the effect estimate of the original study. A relative effect size smaller than 1 means that there is shrinkage of the effect. Then, to investigate the performance of the five metrics under different levels of QRP we will follow Muradchanian et al. (2021) and compute the proportion of successful replications using the different metrics. We compute two different proportions: one based on the total number of simulations and one based on the number of significant original studies, that is, the number of replication studies. The proportions based on the total number of simulations correspond to the estimated overall type-I error (T1E) rate or the project power depending on whether the data were simulated under the null or the alternative hypothesis. The overall T1E rate of the two-trials rule in the absence of QRPs is the squared nominal T1E rate $\alpha^2 = 0.025^2 = 0.000625$ and the project power of the two-trials rule in the absence of QRPs would be the squared nominal power: $(1 - \beta)^2 = 0.85^2 = 0.7225$. The proportions calculated for all the replications correspond to estimated T1E rate and power, respectively. In theory, the overall and standard false-positive rate should be kept low, while the project and standard power should be high.

It is important not to investigate power and T1E rate in isolation. An increase of power with k could be interpreted as a benefit, but at the same time we may also observe an increase of T1E rate, which in turn should cause concern. To combine T1E rate and power in one measure, Bayarri et al. (2016) suggested the pre-experimental rejection ratio

$$R_{\text{pre}} = \frac{\text{Power}}{\text{T1E rate}}.$$

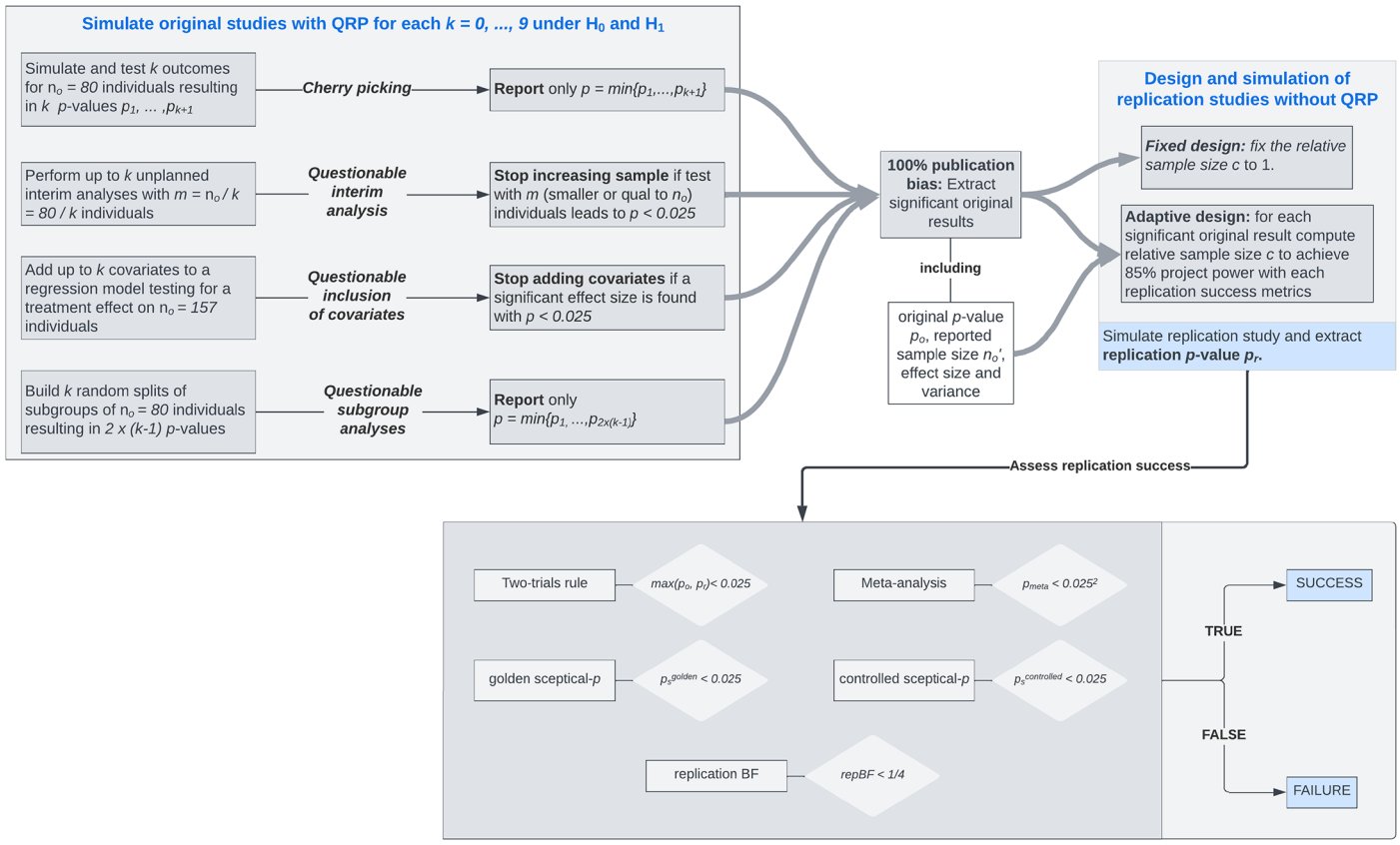

FIG. 1. *A diagram explaining the main steps in our simulation study.*

It can be interpreted as the odds of a correct rejection of the null hypothesis to an incorrect rejection of the null. The higher this ratio, the better the performance of the metric in correctly classifying replication success. Using overall T1E rate and project power in the above equation leads to the overall rejection ratio. The overall R_{pre} is $(1-\beta)^2/\alpha^2 = 34^2 = 1156$.

3. RESULTS

We first investigate the effect of the (different levels of) QRPs on the original studies. The strong positive effect of QRPs on T1E rate was already described elsewhere (Simmons, Nelson and Simonsohn, 2011, Roettger, 2019), but in order to fully understand what this means for replication success we start by investigating the influence on the original studies.

3.1 Original Studies with QRP

The T1E rate for different severity levels k are shown in Figure 2.A. In the absence of QRP ($k = 0$) the T1E rate in the original studies is, as expected, equal to $\alpha = 0.025$. Already weak QRP ($k = 1, 2$) doubles the T1E rate for cherry picking, questionable interim and subgroup analyses. Only the questionable inclusion of covariates does not increase the proportion of false positives as quickly. Figure 2.B shows the proportion of significant results under the alternative hypothesis (H_1), that is, the power, depending on the severity level k. In the absence of QRP ($k = 0$) the power is equal to 0.85. Then, with increasing k the chance of finding a true effect quickly increases. We observe the fastest increase for cherry picking and the lowest for questionable interim analyses.

We show the pre-experimental rejection ratio of H_1 to H_0 in Figure 2.C. The ordering of the different practices with respect to T1E rate is reversed for the pre-experimental rejection ratio. Questionable subgroup analyses, has the lowest rejection ratio for all levels k: for very severe questionable subgroup analyses ($k = 9$) we observe around one false rejection for every five true rejections of the null hypothesis.

Turning to the estimated effect size, Figure 3. A shows the original effect size observed in the studies with significant results, depending on the QRP and its level of severity, under the alternative hypothesis. The average effect size with $k = 0$ of those studies is larger than the true effect $\theta = 0.34$, illustrating the increase of effect size—also called the "winner's curse" (van Zwet and Cator, 2021)—caused by the fact that we induced 100% publication bias. The practice of questionable interim analyses has the strongest (positive) impact on the effect size in the original study. On the other hand, the questionable inclusion of covariates negatively affects the effect size, as the additional covariates absorb some of the effect of interest.

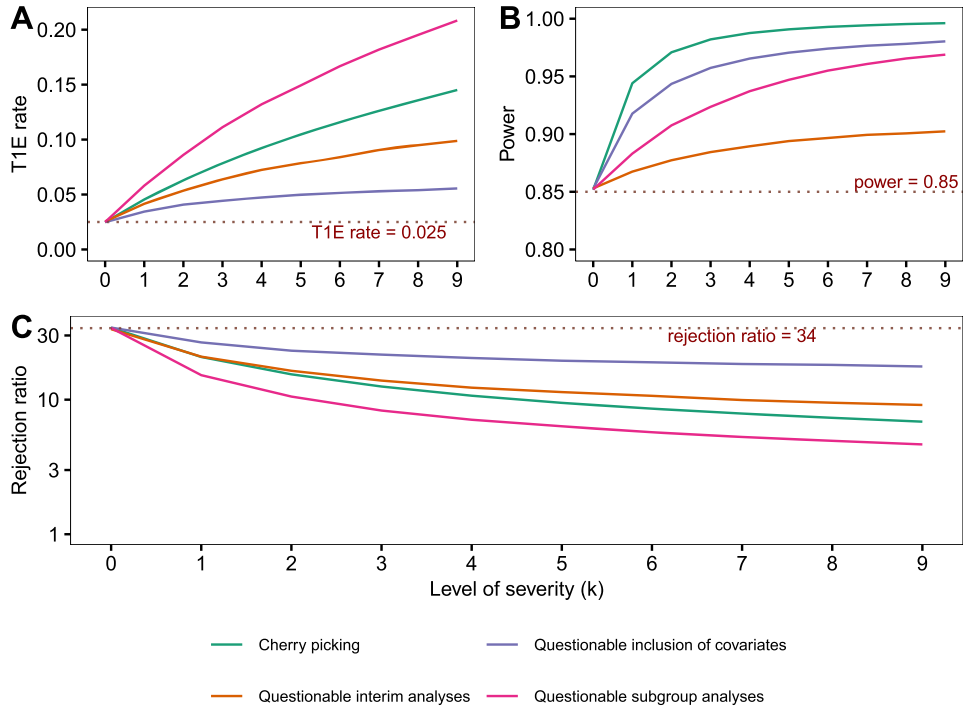

FIG. 2. *Original studies*: *The estimated T1E rate* (*A*), *the power* (*B*) *and the pre-experimental rejection ratio* (*C*) *depending on the level of severity k and the QRP.*

The practice of questionable subgroup analyses leaves the effect size almost unaffected.

As previously mentioned, for the questionable interim and subgroup analyses the reported sample size n_o' of the original study can be smaller than $n_o = 80$. Figure 3.B shows the reduction of average sample size of the original studies with significant effect induced by those QRPs, under the alternative hypothesis. The average published sample size for questionable interim analyses for $k = 2$ is only 53. The larger the severity k of questionable interim analyses, the more the published sample size of the original study drops under the alternative hypothesis. It decreases less fast for questionable subgroup analyses, where the researcher first tests for an effect on the full sample of $n_o = 80$ and only starts splitting the sample if no significant effect could be found. The same quantities as in Figure 3, but under the null hypothesis, can be found in Figure S.2 in the Supplementary Material (Freuli, Held and Heyard, 2023).

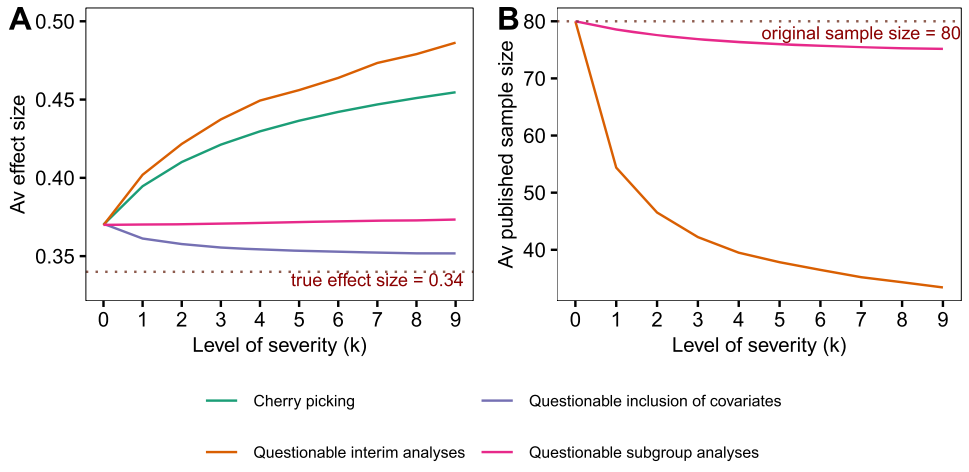

FIG. 3. *Average effect size in the significant original study depending on the QRP and the level of severity k*, *under the alternative hypothesis* (*A*); *and the average published sample size of the significant original studies for questionable interim analyses and subgroup analyses depending on the level of severity k*, *under the alternative hypothesis* (*B*). *For cherry picking and questionable inclusion of covariates*, *the published sample size stays equal to the originally defined sample size of* $n_o = 80$ *and* $n_o = 157$, *respectively.*

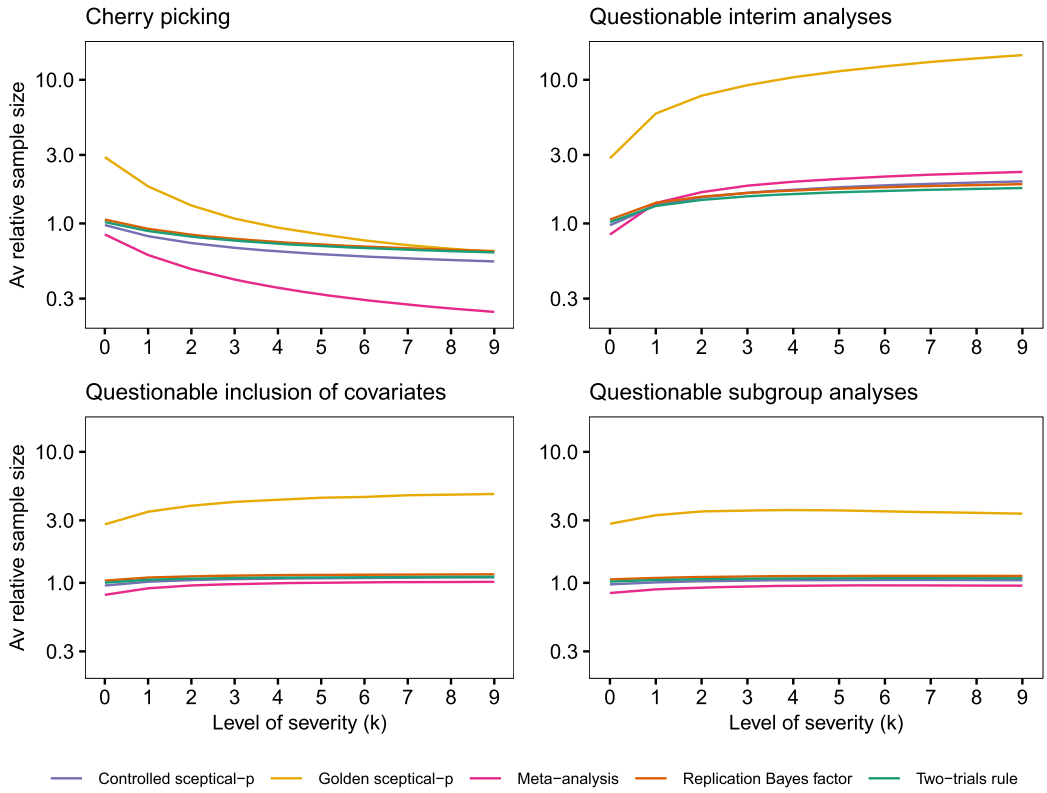

FIG. 4. *The average relative sample size c depending on the adaptive design chosen to compute the sample size for all QRPs and level of severity, under the alternative hypothesis. If the design is fixed c is always* 1.

3.2 Design of Replication Studies

For each original study with a significantly positive effect estimate, a replication study is designed based on the published results (i.e., the effect size, reported sample size and p-value). As described in Section 2.2.1, we used different designs to calculate the sample size of the replication studies (fixed and adaptive design).

If computed adaptively, the average relative sample size $c = n_r/n_o'$ (averaged over all designed replications) depends on the QRP, its level of severity and the chosen metric for replication success, as shown in Figure 4 under the alternative hypothesis. For all QRPs but cherry picking higher severity levels increase the relative sample size at least slightly. The adaptive relative sample size is based on the p-value of the original study which—on average—increases with higher levels of severity for these same QRPs (see Figure S.1 in the Supplementary Material). When adaptively designing c, larger p-values in the original study tend to require a larger replication sample size. Using the golden sceptical p-value to design c leads to larger relative sample sizes when k and the original p-values increase. Especially for questionable interim analyses, c based on the golden sceptical p-value becomes sometimes unreasonably large. On the contrary, since the original p-values decrease with k for cherry picking, the original studies are interpreted as very convincing and smaller sample sizes are sufficient for successful replications (c decreases with k for cherry picking for all metrics). Note that for cherry picking, c chosen using meta-analysis becomes very small as with a very convincing original study, a replication study would no longer be required for success. The corresponding Figure of the results under the null hypothesis can be found in the Supplementary Material (Figure S.3).

Finally, after all replication studies are designed, they are simulated without QRP. Figure 5 shows the average of the relative effect sizes, depending on the QRP and level of severity. This figure shows the commonly observed shrinkage effect when the researcher engages in cherry picking or questionable interim analyses: the replication effect size is smaller than the original effect size due to bias in the original study. In the presence of questionable subgroup analyses, the relative effect size stays close to 1 for all k as this QRP does not inflate the original effect size as much. The original effect size under questionable inclusion of covariates decreases and the relative effect size increases with k. We will refer to this phenomenon as "inverse shrinkage".

3.3 Replication Success

The next sections will investigate the differences in performance of the replication success metrics in the presence of different (levels of) QRPs, under both, the null

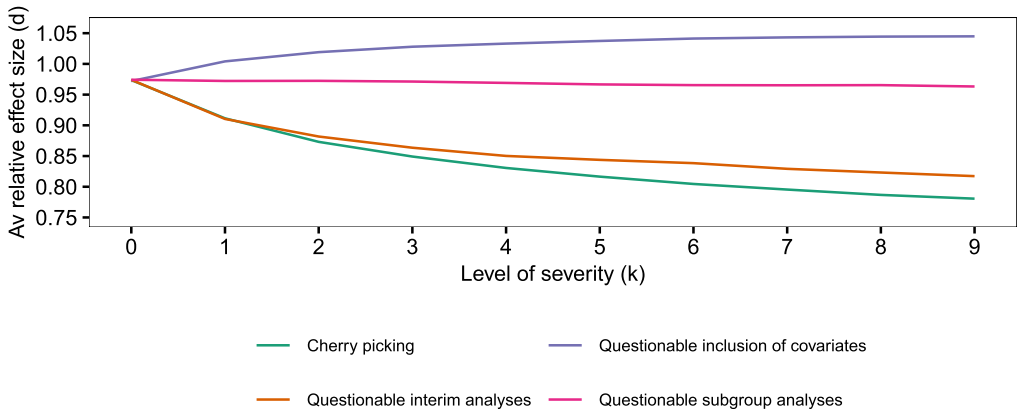

FIG. 5. *The average relative effect size depending on the QRP and the level of severity k, under the alternative hypothesis. For this Figure we only show the results with fixed sample size where $c = 1$.*

and alternative hypotheses. The results for all metrics with fixed replication sample size ($c = 1$) are presented together with the results with adaptive design. We will separately analyse and discuss the effects of QRPs on the overall T1E rate, the project power and the overall rejection ratio. Figures S.5–S.7 in the Supplementary Material show the corresponding quantities when the proportions are computed for all conducted replications. Those proportions represent the T1E rate of the replication, the replication power and pre-experimental rejection ratio of the replication.

3.3.1 *Overall type-I error.* Firstly, Figure 6 closely examines how varying levels of QRP affect the overall T1E rate when defining replication success using different metrics and either designing the replication adaptively or not.

In the absence of QRPs ($k = 0$) and apart from the replication BF, all metrics are defined in a way to control the overall T1E rate approximately at level $\alpha^2 = 0.000625$. The results for the two-trials rule at $k = 0$ show perfect overall T1E rate control at α^2. The overall T1E rate at $k = 0$ for the controlled and golden sceptical p-values and the meta-analysis are slightly smaller than α^2 because of the 100% publication bias in our simulation study. The controlled sceptical p-value is defined as to exactly control this overall T1E rate, so if nonsignificant results were replicated, the overall T1E rate would be equal to α^2 when $k = 0$. Still at $k = 0$, the golden sceptical p-value combined with an adaptive design results in an even lower overall T1E rate which comes from the fact that the adaptive design leads to very large—sometimes even unreasonably large—relative sample sizes (see Figure S.3 in the Supplementary Material). From Equation 5 in Held, Micheloud and Pawel (2022), we know that increasing the replication sample size leads to a more stringent requirement for replication success which in turn, under the null hypothesis, lowers the chance of a replication study that is convincing enough. The overall T1E rate of the replication BF is not bounded at α^2 as can be seen for $c = 1$ and $k = 0$.

Regardless of the QRP investigated, the false-positive rate increases with k because it is influenced by the increase in the number of false-positive original results (as seen in Figure 2). This increase is most pronounced when the original studies suffer from questionable subgroup analyses followed by the setting with cherry picked original results. Questionable inclusion of covariate, on the other hand, has virtually no effect on the overall T1E rate as it also did not increase the T1E rate of the original studies much. For all k and QRPs, defining replication success with the golden sceptical p-value and designing the replication adaptively leads to the lowest overall T1E rate. It also increases less fast with k as compared to the other metrics which can be explained by the fact that the golden sceptical p-value penalizes shrinkage (see Figure 5). This metric flags a replication as successful only if both studies, the original and the replication, are convincing by themselves. When applying a fixed design, the ordering in overall T1E rate observed for the different metrics is consistent for all QRPs: the replication BF results in the largest overall T1E rate followed by the two-trials rule for all k. The smallest overall T1E rate is observed with either the golden sceptical p-value or the meta-analysis as metrics. Comparing the influence of the QRPs on the performance of the metrics is easier with fixed c. Applying cherry picking leads to a lot of shrinkage with increasing k which is penalized by the golden sceptical p-value, explaining the good performance of the metric. For questionable interim analyses and questionable inclusion of covariates combined with a fixed design, the meta-analysis performs slightly better, while the overall T1E for the golden sceptical p-value and the meta-analysis with fixed design are very similar for questionable subgroup analyses. The potential superior performance of the metric based on meta-analysis disappears with an adaptive design, especially for cherry-picking. As discussed

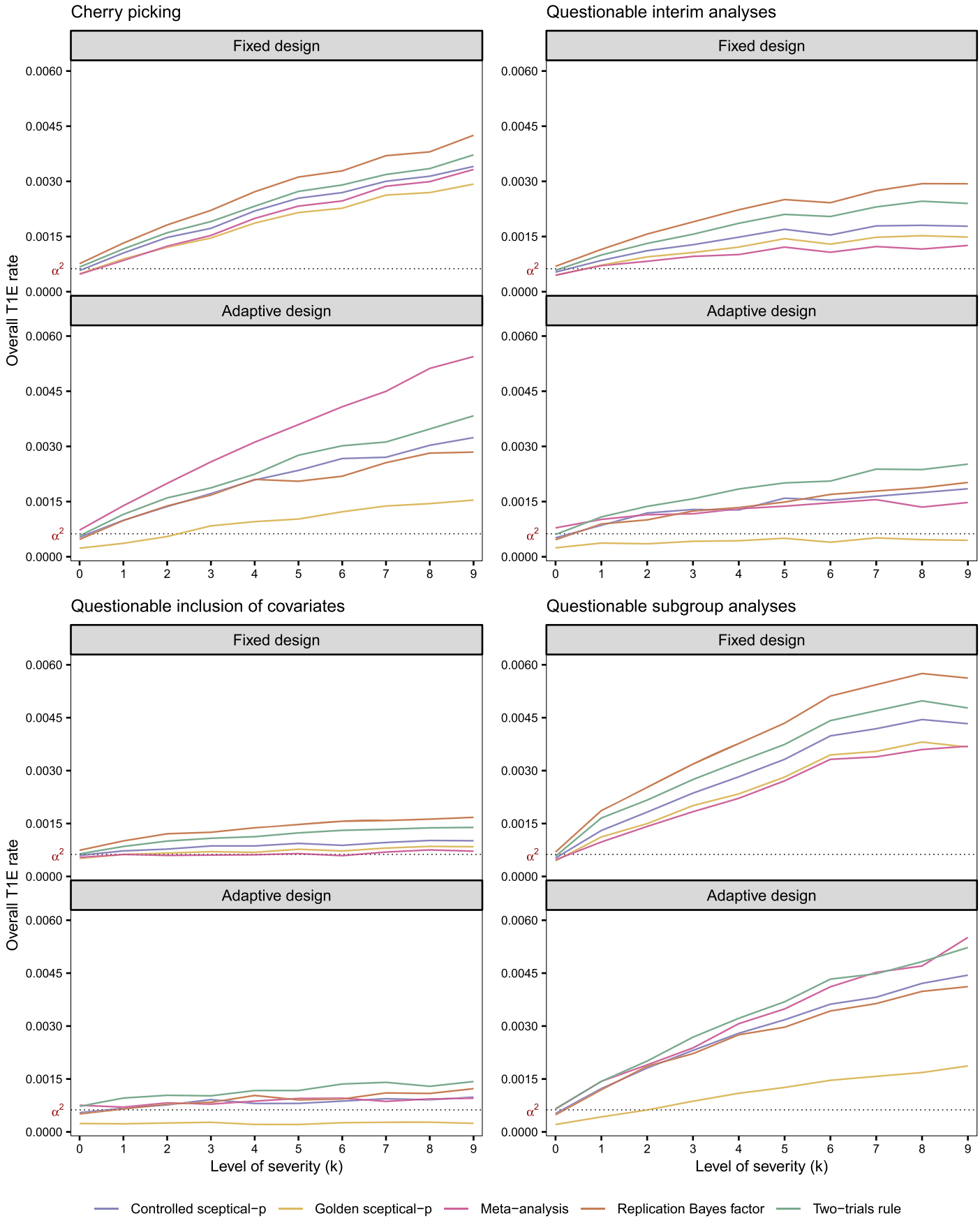

FIG. 6. *The estimated overall T1E rate for increasing severity level of different QRPs, depending on the metric for replication success used and the design. The overall T1E rate of the two-trials rule in the absence of QRPs and publication bias is* $\alpha^2 = 0.000625$.

above, severe cherry-picking decreases the original p-value which increases the chance of defining replication success with a meta-analysis. This is particularly the case with adaptive design where c is always larger than 2 with meta-analysis as criterion (see Figure S.3 in the Supplementary Material) as this might in addition decrease the replication p-values. The jump in overall T1E from fixed to adaptive design with the meta-analysis is similar for questionable subgroup analyses, but less strong for the remaining two QRPs. The original p-values are less affected by severe questionable subgroup analyses and stay relatively small, while they increase with questionable inclusion of covariates and interim analyses.

The adaptive replication BF usually performs better than the fixed version (with lower overall T1E rate) because, under H_0, the adaptive c tends to be larger than 1 which decreases the chance of a convincing replication study. As the replication BF needs the replication study to be convincing and cares less about the evidence from the original study this leads to a lower share of replication success under H_0. The metric based on the controlled sceptical p-value results in an overall T1E always lower to the one with the two-trials rule. This result is certainly influenced by the simulation of 100% publication bias, but the gap between the overall T1E rate of the two metrics increases with k.

The T1E rate using the metric based on meta-analysis would be inflated even more if nonsignificant original results were replicated, as it is possible to have the meta-analysis flag a replication successful even if the original result was not convincing with a large p-value, whenever the estimated effect in the replication study is very strong and vice versa (as can be inferred from Figure 9). Interpretation for the replication BF in comparison to the metrics based on p-values is not straightforward. However, the overall T1E rate inflation would be even more pronounced for the replication BF if nonsignificant original studies would have been replicated. Indeed, the replication BF does not directly depend on how much evidence the original study provides against the null and can therefore flag replication success with an unconvincing original study as long as the replication result is compelling (Pawel and Held, 2022).

3.3.2 *Project power.* Next, Figure 7 shows the influence of varying levels of QRPs on the project power when defining replication success using different metrics with an adaptive and fixed design.

Starting with the interpretation of project power at $k = 0$ we observe that only the two-trials rule with fixed design leads to a project power of approximately $(1 - \beta)^2$. To design the replication studies adaptively we assumed that the original effect size was the truth. However, due to the 100% publication bias, the original effect size is—on average—larger than the truth. Therefore, the adaptive version of the project power for the two-tirals rule with $k = 0$ is smaller than $(1 - \beta)^2$.

Higher severity of cherry picking positively influences the overall power when using the meta-analytical metric to quantify replication success, or combine a fixed design of the relative sample size with any of the other metrics. The metrics other than meta-analysis with adaptive design lead to a decreased project power once $k \geq 2$. As previously seen (in Figure 4), the average relative sample size from an adaptive design decreases with k as the effect size of the original significant result to be replicated increases. This has a direct effect on the project power.

When questionable interim analyses are applied we observe an interesting phenomenon where the ordering of the metrics by project power is reversed once $k \geq 2$. In general, severe levels of questionable interim analyses have a detrimental effect on the project power, especially with fixed design or golden sceptical p-value. This can be explained by the fact that this QRP induces important shrinkage and leads to small original sample sizes, which in turn, make the extreme original effect hard to replicate with $c = 1$. Adaptively designing the replication study seems to preserve a higher project power for all metrics but the golden sceptical p-value. For the latter metric the questionable interim analyses lead to very large c which leads to more stringent success requirements. In the presence of inverse shrinkage as for questionable inclusion of covariates, an inflation of project power is observed also for the adaptive version of the golden sceptical p-value. Questionable subgroup analyses have a less strong effect on the project power and we observe less obvious differences between the metrics.

3.3.3 *Overall rejection ratio*. Finally, the overall rejection ratio in Figure 8 provides a summary of the previous results.

The overall rejection ratio confirms the results observed under the null hypothesis: for all QRPs, we observe higher ratios estimated for the adaptive version of the golden sceptical p-value which indicates the largest number of true rejections for each false rejection of replication success. The overall rejection ratios decrease with k for all metrics and QRP while they are almost stable for questionable inclusion of covariates since the overall T1E rate for this QRP and publication bias does not increase much. Also, for all practices and large values of k, the overall rejection ratio is smaller than 1156, the ratio for the two-trials rule in the absence of QRP, which indicates that the presence of QRP in the original study render false rejections of replication success more likely, regardless of the replication success metric used. Only the adaptive version of the sceptical p-value and the fixed version of the meta-analysis stay above this value for all k in the presence of questionable inclusion of covariates.

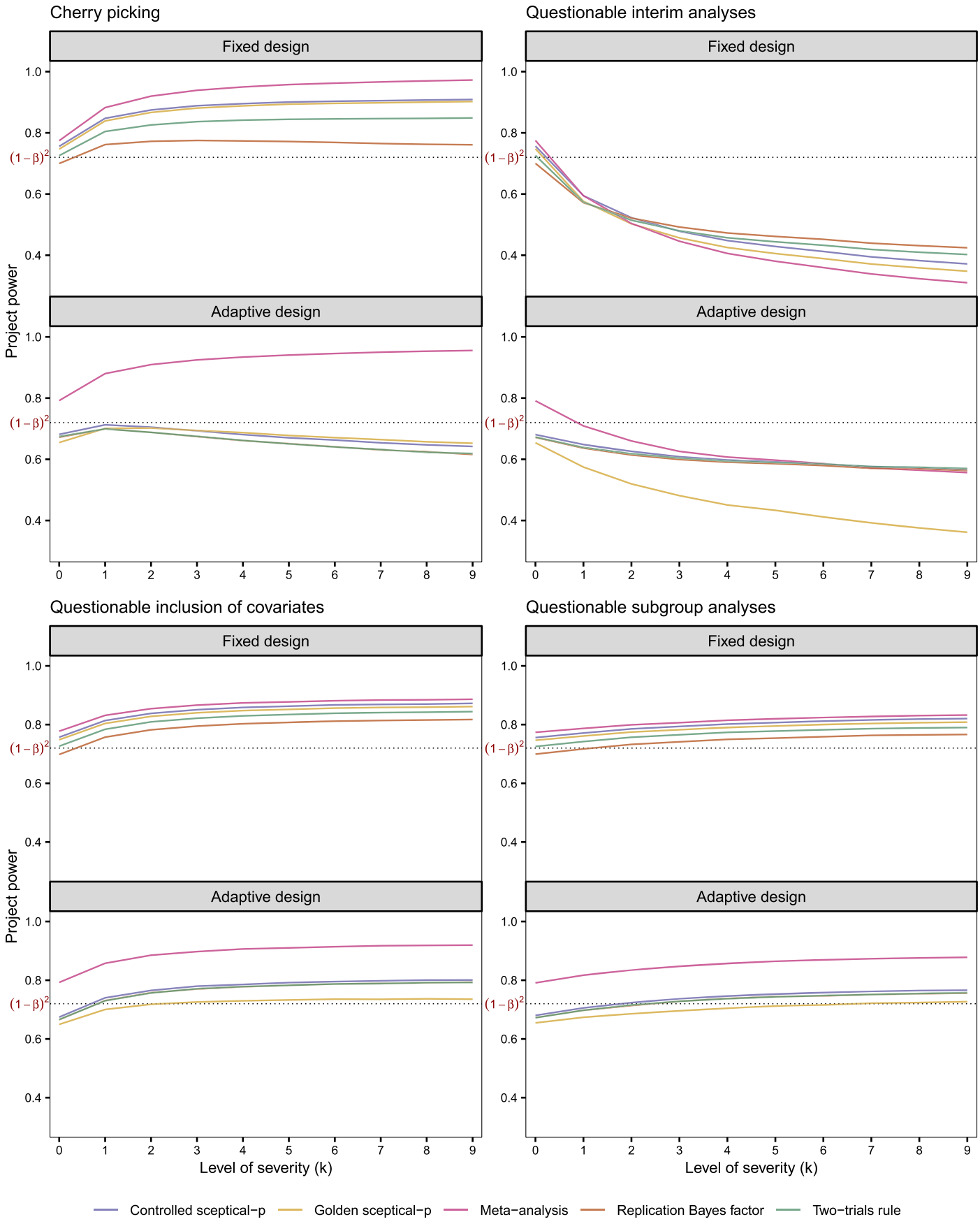

FIG. 7. *The estimated project power for increasing severity level of different QRPs, depending on the metric for replication success used and the design. The project power of the two-trials rule in the absence of QRPs and publication bias is* $(1-\beta)^2 = 0.7225$.

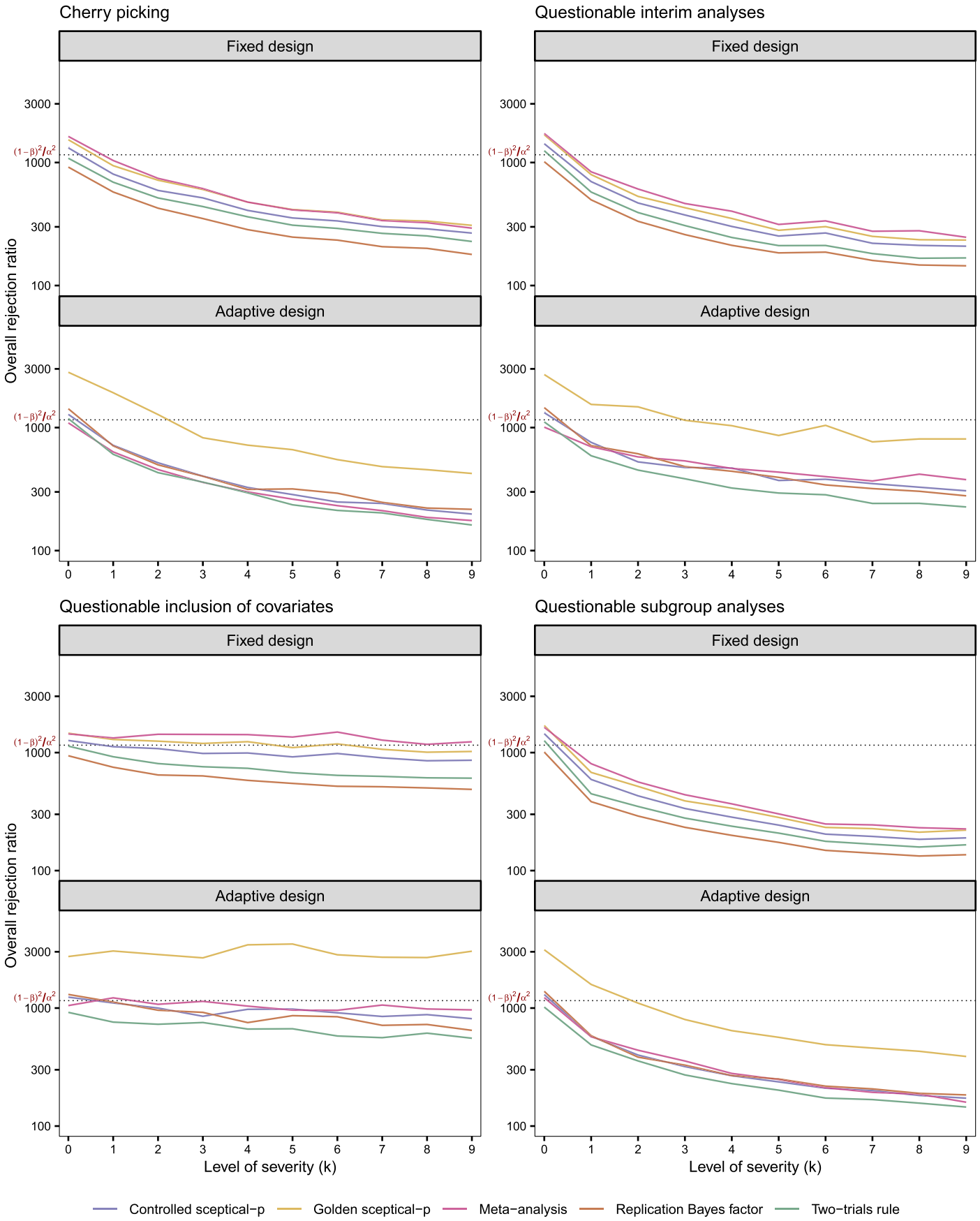

FIG. 8. *The estimated overall rejection ratio for increasing severity level of different QRPs, depending on which metric defines replication success and which design was used. The ratio for the two-trials rule in the absence of QRPs and publication bias is* $(1-\beta)^2/\alpha^2 = 1156$.

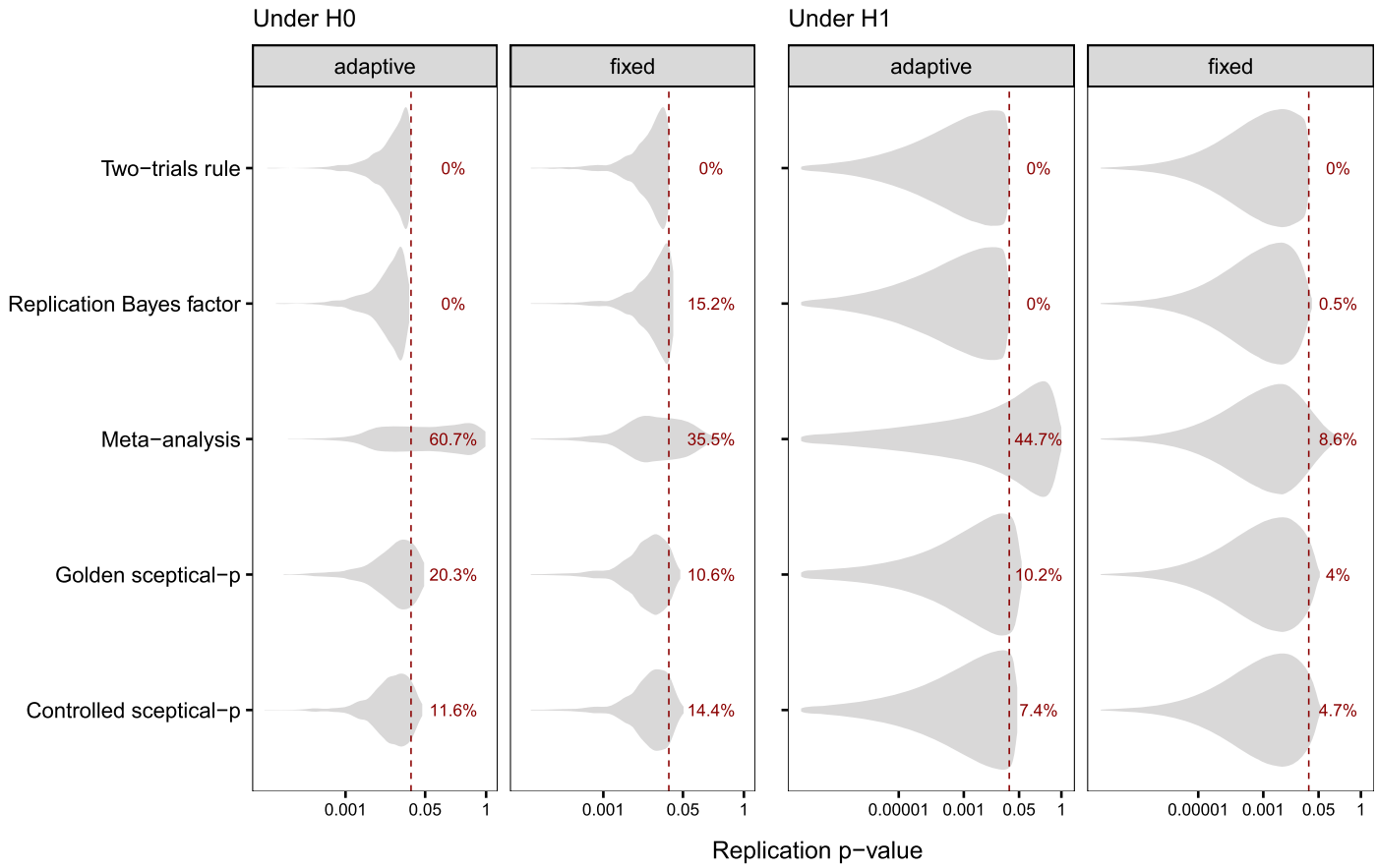

FIG. 9. *Violin plots of the replication p-values (larger than* 10^{-8}*) of those replication studies that were judged successful by the different metrics (on the y-axis), depending on whether c was fixed to one or estimated adaptively under the null and the alternative hypotheses. The results are presented for* $k = 0$*, under both the null and the alternative hypothesis for all QRPs pooled together. The red dashed lines represent the significance level* $\alpha = 0.025$*. The percentage of values larger than* α *are reported. Note that by design all original p-values were significant.*

Regardless of the practice studied, the meta-analytical metric to quantify replication success performs relatively well (i.e., large overall rejection ratios which mean a high number of true rejections for each false rejection of replication success). To understand why, we refer to Figure 9. Here we see for each hypothesis and design, the p-values computed in those replication studies that were successful depending on the definition of replication success. With the meta-analytical metric, we can obtain replication success even if the replication p-value is large. The meta-analysis allows such large replication p-values whenever the original study was very convincing. This metric does not require both studies to be "convincing on its own", in contrast to the common understanding of a successful replication. This figure also shows how, for the golden and controlled sceptical p-value to flag success the replication p-value does not have to be smaller than α, but the study has to be convincing. The replication BF seldomly allows the p-values of the replication to be larger than α, and if it does, this p-value is still very close to α. Hence, it is crucial not to investigate one operating characteristic in isolation, but rather weight them against each other and inspect replication success case-by-case.

4. DISCUSSION

In this simulation study, we compared the performance of different replication success metrics in the presence of QRPs. The simulations were performed under both the null and the alternative hypotheses. Only the significant original results were replicated since we assumed 100% publication bias, where only the studies with significant results would get published. The replication studies were designed based on the published results. Diverse metrics were proposed to quantify replication success, and we compared the performance of the following metrics: standard significance, often referred to as two-trial rule, the meta-analytical approach, two versions of the sceptical p-value, with "golden" or "controlled" recalibration, respectively, and replication BF. In addition, we allowed for increasing levels of severity k for each of the four QRPs studied: cherry picking, questionable interim analyses, questionable inclusion of covariates and questionable subgroup analyses. To compare the performance of the replication success metrics, we estimated the overall T1E

rate, the project power and the overall rejection ratio. The design of our simulation study was preregisted on OSF.

We found that applying QRPs in the original study has a strong effect on the operating characteristics of the original study, typically leading to an increase in T1E rate, and a decrease in power and rejection ratio. Furthermore, QRPs affected the original effect sizes substantially, producing a strong shrinkage effect for cherry picking and questionable interim analyses and inverse shrinkage for questionable inclusion of covariates. These effects in turn influenced the replicability of the original results. Using the golden sceptical p-value to define replication success and computing the replication sample size led to the smallest values of overall T1E rate for all severity levels k, QRPs. This could potentially be explained by the fact that the sceptical p-values requires both studies to be convincing enough with respect to the p-values but also the relative effect size. On the other hand, other metrics, such as the two-trial rule, might declare replication success even if there is substantial shrinkage of the replication effect size, for example, caused by QRPs in the original study. This is especially likely if the relative sample size is larger.

Regardless of the metric, the overall rejection ratio decreases with the severity level k for all practices except for questionable inclusion of covariates where it is constant. Also for this ratio the golden sceptical p-value performs best (i.e., highest values) for all QRPs and severity levels, while the controlled sceptical p-value is consistently better than the two-trials rule and the replication BF when comparing the same designs, but worse than the meta-analytical approach and at a similar level.

Interestingly, with respect to its operating characteristics, the meta-analysis performed strikingly well in the simulation study: low overall T1E rate and high project power. However, this observation is only linked to the fact that we only simulated the replications of those original studies that yield a significant result. The behavior of the meta-analytical approach is much different from the two-trial rule and sceptical p-value as it can lead to replication success if the replication study is not significant and shows substantial shrinkage. Such replication results cannot be successful when using the golden sceptical p-value as it penalizes shrinkage. Even though the golden sceptical p-value performs well, it also requires larger relative sample size. Hence, it is a trade-off that has to be considered, as the required relative sample size may be very large rendering a reasonably powered replication unfeasible. The replication BF does not consider both studies of equal importance, but can rather be very small and lead to success if the replication is very convincing even if the original study is not. This scenario is not simulated here, leading to largly underestimated overall T1E rates.

This is the first study investigating the performance of different replication success metrics in the presence of a set of QRPs. The obtained results show interesting perspectives for future studies. First of all, we did not investigate the effect of combinations of different QRPs, as done in Simmons, Nelson and Simonsohn (2011). In addition, it is necessary to emphasize that in our study we simulated the QRPs following one of the multiple descriptions reported in the literature (Wang et al., 2017). More comparisons, and even neutral comparison studies (Boulesteix, Lauer and Eugster, 2013), of the golden sceptical p-value, which was the most promising in our study, with other replication success metrics are needed. Finally, in-depth analyses of the implications of the designs to determine the relative sample size could give insight into and recommendations on which designs should be used in which situation. Another interesting extension in our simulation study could be to allow varying base rates of true effects as did Ulrich and Miller (2020), because they found low base rates being an important contributor of low replicability.

Our study is not without limitations. We only focused on a subset of four QRPs which were reasonably straightforward to simulate. We are aware that also simulation studies can be subject to QRPs and made an effort to avoid them by preregistering our study protocol (Pawel, Kook and Reeve, 2023). We only designed and simulated a replication study if the original study showed a significant result. This does not affect the overall T1E rate of the two-trials rule, but it does reduce the overall T1E rate of other methods. Specifically, the sceptical p-value in both the golden and controlled version avoids the "double dichotomisation" of the two-trials rule and can flag replication success even if the p-value of the original study is somewhat larger than α. A restriction to significant studies only will hence reduce both overall T1E rate and project power (Held, Micheloud and Pawel, 2022, Section 3). The meta-analytical approach and replication BF may even flag replication success if one of the studies is not convincing at all and the restriction to significant original studies again reduces overall T1E rate and project power. We made this choice in the assumption that a researcher performs QRPs only to get a significant result that can easily be published. Furthermore, conducting replication studies of nonsignificant original studies will increase the costs of large-scale replication projects in practice. However, it would be interesting to assess the performance of the replication success metrics considering all original results. Finally, the simulation study could be extended to "many-to-one" replication designs (Klein et al., 2014).

It is important to note that in our simulation study, the replication followed exactly the same (simple) design as the original study (without QRPs). In real world replications this is not guaranteed and differences in study design

might influence replication success. Further, in our simulation setup we are interested in the proportion of replication successes (under the null and the alternative hypotheses) recorded as a binary measure of success. This could be seen as a limitation and it would be of interest to also analyse replication success from a quantitative perspective.

5. SOFTWARE, DATA, AND SOURCE FILES

All materials related to this paper are available from gitlab.uzh.ch/rachel.heyard/qrpsimulations and OSF (osf.io/ydbsh/). This paper can be reproduced using the `Rmarkdown` version of the document. The scripts used for the simulations are also included in the gitlab repository. The entire study was conducted in `R` (version 4.2.3) and the simulation was designed using the `SimDesign` package. The methodology based on the sceptical p-value is implemented in the `R` package `ReplicationSuccess` available from `CRAN`. We used the replication BF functionality from the package `BayesRep` available on gitlab (gitlab.uzh.ch/samuel.pawel/BayesRep) and `BayesRepDesign` available from `CRAN`.

CONFLICT OF INTEREST

LH is the inventor of the sceptical p-value. FF and RH declare no conflicts of interest.

ACKNOWLEDGMENTS

We thank Samuel Pawel for helpful comments on our manuscript.

SUPPLEMENTARY MATERIAL

Supplement to "Replication Success Under Questionable Research Practices—a Simulation Study" (DOI: 10.1214/23-STS904SUPP; .pdf). Figure S.1: significant original p-values. This Figure shows violin plots of the p-values of the significant original studies under H_1. Figure S.2: The average significant original effect sizes and the published original sample size under the null hypothesis. The first part of the Figure (A) shows the average effect size of all those original studies yielding a positive significant results under the null hypothesis, depending on the QRP and the level of severity employed. The figure shows how large the bias of the published results under the null is already without QRP ($k = 0$), and how it is affected by the QRP. For questionable interim analyses and subgroup analyses the average published sample size of the significant original studies under the null is represented in the second part of the Figure (B). Figure S.3: The average relative sample size c under the null hypothesis. The figure presents the relative sample size c averaged over all designed replications under the null hypothesis for different QRPs and different levels of severity k. Compared to the average relative sample size under the alternative presented in the main manuscript, c is less affected by the QRPs and their level of severity. Figure S.4: The proportion of significant original results per 1000 simulated studies. The proportion of significant original results per 1000 simulated studies in both, the null and the alternative, are shown in this figure depending on the QRP employed and the level of severity. The representations directly relate to the T1E rate and the power (of the original studies). Figure S.5–S.7: Performance of replication success metrics for different QRPs. These figures present the T1E rate, the power, and the pre-experimental rejection ratios. Unlike in the main paper, these quantities are computed as the share (or the ratio of the shares) of successful replications among all replications or original significant results, under the null and the alternative hypothesis respectively.

REFERENCES

AGNOLI, F., WICHERTS, J. M., VELDKAMP, C. L. S., ALBIERO, P. and CUBELLI, R. (2017). Questionable research practices among Italian research psychologists. *PLoS ONE* **12** e0172792. https://doi.org/10.1371/journal.pone.0172792

ANDERSON, S. F. and KELLEY, K. (2022). Sample size planning for replication studies: The devil is in the design. *Psychol. Methods*. https://doi.org/10.1037/met0000520

ANDERSON, S. F. and MAXWELL, S. E. (2016). There's more than one way to conduct a replication study: Beyond statistical significance. *Psychol. Methods* **21** 1–12. https://doi.org/10.1037/met0000051

BAYARRI, M. J., BENJAMIN, D. J., BERGER, J. O. and SELLKE, T. M. (2016). Rejection odds and rejection ratios: A proposal for statistical practice in testing hypotheses. *J. Math. Psych.* **72** 90–103. MR3506028 https://doi.org/10.1016/j.jmp.2015.12.007

BISHOP, D. (2019). Rein in the four horsemen of irreproducibility. *Nature* **568** 435–435.

BOULESTEIX, A.-L., LAUER, S. and EUGSTER, M. J. A. (2013). A plea for neutral comparison studies in computational sciences. *PLoS ONE* **8** e61562. https://doi.org/10.1371/journal.pone.0061562

BROOKES, S. T., WHITELY, E., EGGER, M., SMITH, G. D., MULHERAN, P. A. and PETERS, T. J. (2004). Subgroup analyses in randomized trials: Risks of subgroup-specific analyses; power and sample size for the interaction test. *J. Clin. Epidemiol.* **57** 229–236. https://doi.org/10.1016/j.jclinepi.2003.08.009

BURTON, A., ALTMAN, D. G., ROYSTON, P. and HOLDER, R. L. (2006). The design of simulation studies in medical statistics. *Stat. Med.* **25** 4279–4292. MR2307592 https://doi.org/10.1002/sim.2673

CHRISTIAN, K., JOHNSTONE, C., LARKINS, J.-A., WRIGHT, W. and DORAN, M. R. (2021). A survey of early-career researchers in Australia. *eLife* **10**. https://doi.org/10.7554/eLife.60613

OPEN SCIENCE COLLABORATION (2015). Estimating the reproducibility of psychological science. *Science* **349**.

COUSINS, R. D. (2007). Annotated bibliography of some papers on combining significances or p-values.

ERRINGTON, T. M., MATHUR, M., SODERBERG, C. K., DENIS, A., PERFITO, N., IORNS, E. and NOSEK, B. A. (2021). Investigating the replicability of preclinical cancer biology. *eLife* **10**. https://doi.org/10.7554/eLife.71601

FREULI, F., HELD, L. and HEYARD, R. (2023). Supplement to "Replication success under questionable research practices—a simulation study." https://doi.org/10.1214/23-STS904SUPP

GOPALAKRISHNA, G., RIET, G. T., VINK, G., STOOP, I., WICHERTS, J. M. and BOUTER, L. M. (2022). Prevalence of questionable research practices, research misconduct and their potential explanatory factors: A survey among academic researchers in The Netherlands. *PLoS ONE* **17** e0263023. https://doi.org/10.1371/journal.pone.0263023

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

HEAD, M. L., HOLMAN, L., LANFEAR, R., KAHN, A. T. and JENNIONS, M. D. (2015). The extent and consequences of p-hacking in science. *PLoS Biol.* **13** e1002106. https://doi.org/10.1371/journal.pbio.1002106

HEDGES, L. V. and SCHAUER, J. M. (2019). More than one replication study is needed for unambiguous tests of replication. *J. Educ. Behav. Stat.* **44** 543–570.

HELD, L. (2020). A new standard for the analysis and design of replication studies. *J. Roy. Statist. Soc. Ser. A* **183** 431–448. MR4052785

HELD, L., MATTHEWS, R., OTT, M. and PAWEL, S. (2022). Reverse-Bayes methods for evidence assessment and research synthesis. *Res. Synth. Methods* **13** 295–314. https://doi.org/10.1002/jrsm.1538

HELD, L., MICHELOUD, C. and PAWEL, S. (2022). The assessment of replication success based on relative effect size. *Ann. Appl. Stat.* **16** 706–720. MR4438808 https://doi.org/10.1214/21-aoas1502

HELD, L. and OTT, M. (2018). On p-values and Bayes factors. *Annu. Rev. Stat. Appl.* **5** 393–422. MR3774753 https://doi.org/10.1146/annurev-statistics-031017-100307

JOHN, L. K., LOEWENSTEIN, G. and PRELEC, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychol. Sci.* **23** 524–532. https://doi.org/10.1177/0956797611430953

KIRKHAM, J. J., ALTMAN, D. G., CHAN, A.-W., GAMBLE, C., DWAN, K. M. and WILLIAMSON, P. R. (2018). Outcome reporting bias in trials: A methodological approach for assessment and adjustment in systematic reviews. *BMJ* **362** k3802. https://doi.org/10.1136/bmj.k3802

KIRKHAM, J. J., DWAN, K. M., ALTMAN, D. G., GAMBLE, C., DODD, S., SMYTH, R. and WILLIAMSON, P. R. (2010). The impact of outcome reporting bias in randomised controlled trials on a cohort of systematic reviews. *BMJ* **340** c365. https://doi.org/10.1136/bmj.c365

KLEIN, R. A., RATLIFF, K. A., VIANELLO, M., ADAMS, R. B., BAHNÍK, Š., BERNSTEIN, M. J., BOCIAN, K., BRANDT, M. J., BROOKS, B. et al. (2014). Investigating variation in replicability. *Soc. Psychol.* **45** 142–152.

LY, A., ETZ, A., MARSMAN, M. and WAGENMAKERS, E.-J. (2018). Replication Bayes factors from evidence updating. *Behav. Res. Methods* **51** 2498–2508.

MATTHEWS, J. N. S. (2006). *Introduction to Randomized Controlled Clinical Trials*, 2nd ed. *Texts in Statistical Science Series*. CRC Press/CRC, Boca Raton, FL. MR2261274 https://doi.org/10.1201/9781420011302

MAYO-WILSON, E., LI, T., FUSCO, N., BERTIZZOLO, L., CANNER, J. K., COWLEY, T., DOSHI, P., EHMSEN, J., GRESHAM, G. et al. (2017). Cherry-picking by trialists and meta-analysts can drive conclusions about intervention efficacy. *J. Clin. Epidemiol.* **91** 95–110.

MICHELOUD, C., BALABDAOUI, F. and HELD, L. (2023). Assessing replicability with the sceptical p-value: Type-I error control and sample size planning. *Stat. Neerl.*.

MICHELOUD, C. and HELD, L. (2022). Power calculations for replication studies. *Statist. Sci.* **37** 369–379. MR4444372 https://doi.org/10.1214/21-sts828

MORAN, C., RICHARD, A., WILSON, K., TWOMEY, R. and COROIU, A. (2022). I know it's bad, but I have been pressured into it: Questionable research practices among psychology students in Canada. *Can. Psychol.*

MORRIS, T. P., WHITE, I. R. and CROWTHER, M. J. (2019). Using simulation studies to evaluate statistical methods. *Stat. Med.* **38** 2074–2102. MR3937487 https://doi.org/10.1002/sim.8086

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. *R. Soc. Open Sci.* **8** 201697.

NATIONAL ACADEMIES OF SCIENCES, ENGINEERING, MEDICINE (2019). *Reproducibility and Replicability in Science*. The National Academies Press. Washington, DC.

NOSEK, B. A., SPIES, J. R. and MOTYL, M. (2012). Scientific utopia: II. Restructuring incentives and practices to promote truth over publishability. *Perspect. Psychol. Sci.* **7** 615–631. https://doi.org/10.1177/1745691612459058

PAWEL, S., CONSONNI, G. and HELD, L. (2023). Bayesian approaches to designing replication studies. *Psychol. Methods*, Accepted.

PAWEL, S. and HELD, L. (2022). The sceptical Bayes factor for the assessment of replication success. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **84** 879–911. MR4460579

PAWEL, S., KOOK, L. and REEVE, K. (2023). Pitfalls and potentials in simulation studies: Questionable research practices in comparative simulation studies allow for spurious claims of superiority of any method. *Biom. J.* e2200091. https://doi.org/10.1002/bimj.202200091

POCOCK, S. J. (1977). Group sequential methods in the design and analysis of clinical trials. *Biometrika* **64** 191–199.

RABELO, A. L., FARIAS, J. E., SARMET, M. M., JOAQUIM, T. C., HOERSTING, R. C., VICTORINO, L., MODESTO, J. G. and PILATI, R. (2020). Questionable research practices among Brazilian psychological researchers: Results from a replication study and an international comparison. *Int. J. Psychol.* **55** 674–683.

ROETTGER, T. B. (2019). Researcher degrees of freedom in phonetic research. *Lab. Phonol.* **10**.

ROSENKRANZ, G. (2019). *Exploratory Subgroup Analyses in Clinical Research*. Wiley, New York.

ROSENKRANZ, G. K. (2023). A generalization of the two trials paradigm. *Ther. Innov. Regul. Sci.* **57** 316–320. https://doi.org/10.1007/s43441-022-00471-4

SAGARIN, B. J., AMBLER, J. K. and LEE, E. M. (2014). An ethical approach to peeking at data. *Perspect. Psychol. Sci.* **9** 293–304.

SCHMIDT, S. (2009). Shall we really do it again? The powerful concept of replication is neglected in the social sciences. *Rev. Gen. Psychol.* **13** 90–100.

SENN, S. (2021). *Statistical Issues in Drug Development*, 3rd ed. Wiley, New York.

SIMMONS, J. P., NELSON, L. D. and SIMONSOHN, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychol. Sci.* **22** 1359–1366.

SIMONSOHN, U. (2015). Small telescopes: Detectability and the evaluation of replication results. *Psychol. Sci.* **26** 559–569. https://doi.org/10.1177/0956797614567341

STEFAN, A. M. and SCHÖNBRODT, F. D. (2023). Big little lies: A compendium and simulation of p-hacking strategies. *R. Soc. Open Sci.* **10**.

ULRICH, R. and MILLER, J. (2020). Questionable research practices may have little effect on replicability. *eLife* **9**. https://doi.org/10.7554/eLife.58237

VAN ZWET, E. W. and CATOR, E. A. (2021). The significance filter, the winner's curse and the need to shrink. *Stat. Neerl.* **75** 437–452. MR4374073 https://doi.org/10.1111/stan.12241

VERHAGEN, J. and WAGENMAKERS, E.-J. (2014). Bayesian tests to quantify the result of a replication attempt. *J. Exp. Psychol. Gen.* **143** 1457–1475. https://doi.org/10.1037/a0036731

WANG, Y. A., SPARKS, J., GONZALES, J. E., HESS, Y. D. and LEDGERWOOD, A. (2017). Using independent covariates in experimental designs: Quantifying the trade-off between power boost and type I error inflation. *J. Exp. Soc. Psychol.* **72** 118–124.

WICHERTS, J. M., VELDKAMP, C. L. S., AUGUSTEIJN, H. E. M., BAKKER, M., VAN AERT, R. C. M. and VAN ASSEN, M. A. L. M. (2016). Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid. *Front. Psychol.* **7** 1832. https://doi.org/10.3389/fpsyg.2016.01832

WOLFF, W., BAUMANN, L. and ENGLERT, C. (2018). Self-reports from behind the scenes: Questionable research practices and rates of replication in ego depletion research. *PLoS ONE* **13** e0199554. https://doi.org/10.1371/journal.pone.0199554