Freuli, Francesca; Held, Leonhard; Heyard, Rachel

Working Paper

Replication Success under Questionable Research Practices - A Simulation Study

I4R Discussion Paper Series, No. 2

Provided in Cooperation with:
The Institute for Replication (I4R)

Suggested Citation: Freuli, Francesca; Held, Leonhard; Heyard, Rachel (2022) : Replication Success under Questionable Research Practices - A Simulation Study, I4R Discussion Paper Series, No. 2, Institute for Replication (I4R), s.l.

This Version is available at:
https://hdl.handle.net/10419/265252

INSTITUTE for REPLICATION

No. 2
I4R DISCUSSION PAPER SERIES

Replication Success under Questionable Research Practices - A Simulation Study

Francesca Freuli

Leonhard Held

Rachel Heyard

October 2022

INSTITUTE for REPLICATION

I4R DISCUSSION PAPER SERIES

I4R DP No. 2

Replication Success under Questionable Research Practices – A Simulation Study

Francesca Freuli[1], Leonhard Held[2], Rachel Heyard[2]

[1] *University of Trento/Italy, Dept. of Psychology and Cognitive Science*

[2] *University of Zurich/Switzerland, Center for Reproducible Science, Dept. of Biostatistics, Epidemiology, Biostatistics and Prevention Institute*

OCTOBER 2022

E-Mail: joerg.peters@rwi-essen.de
RWI – Leibniz Institute for Economic Research
Hohenzollernstraße 1-3
45128 Essen/Germany
www.i4replication.org

ISSN: 2752-1931

Replication success under questionable research practices – a simulation study

Francesca Freuli* Leonhard Held† Rachel Heyard‡

Abstract

Increasing evidence suggests that the reproducibility and replicability of scientific findings is threatened by researchers employing questionable research practices (QRP) in order to achieve publishable, positive and significant results. Numerous metrics have been developed to determine replication success but it has not yet been established 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 different types of QRPs: cherry picking, questionable interim analyses, questionable inclusion of covariates, and questionable subgroup analyses. Our results show that the metric based on the golden sceptical p-value does better in maintaining low values of overall type-I error rate, but often needs larger replication sample sizes, especially when severe QRPs are employed.

1 Introduction

The replicability of research findings in various fields has long been threatened by so-called questionable research practices (QRP). Researchers may engage in QRPs to increase their chances of achieving a positive result which, in return, increases the chance of getting their results published [Simmons et al., 2011, Nosek et al., 2012]. Examples of QRPs are manifold and they differ depending on which of the "researcher degrees of freedom" [Wicherts et al., 2016] was 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 et al., 2011, Roettger, 2019]. The success of a replication of a study with suspected QPRs might therefore be compromised, especially since QRPs are likely not recorded nor reported. The researcher might not even be aware of the consequences [Bishop, 2019]. QRPs are so rooted in the scientific landscape that between 39% and 51% of researchers admit already having applied at least one of those practices [Wolff et al., 2018, Gopalakrishna et al., 2022], considering its use defensible [Rabelo et al., 2020]. Some recent studies showed that young researchers and students had conducted QRPs because they received pressure from their supervisors [Moran et al., 2022, Christian et al., 2021].

As replications of scientific studies are becoming more and more common, metrics to assess whether a replication was successful started to emerge [Anderson and Maxwell, 2016]. There is no universally agreed on criterion for replication success. Therefore, the large replication projects did not use one single metric but rather a 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. In one of the most recent projects, the Cancer Biology Reproducibility Project, Errington et al. [2021] present seven different methods to assess replication success, which are mainly based on direction of effect, effect size and significance. Using

*Francesca Freuli is Ph.D. Student, Department of Psychology and Cognitive Science, University of Trento, Trento, Italy, 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, 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, rachel.heyard@uzh.ch

standard significance as an indicator for replication success, *i.e.* 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 often referred to as "two-trials rule" [Senn, 2007]. 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. This approach flags replication success if the original study and the meta-analysis yield consistent results [Camerer et al., 2018]. 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 et al., 2022c] combines significance of the original and replication studies together with their effect sizes. In an attempt to find the best metric to quantify replication success, 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.

Little is known on how the different replication success metrics behave in the presence of QRP. 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 non-significant p-value significant" [Stefan and Schönbrodt, 2022]. For the present simulation study, we took inspiration from the four different QRPs considered in Simmons et al. [2011] to come up with the following scenarios A to D. For all scenarios, we assume that the researcher is interested in a positive effect and therefore computes one-sided p-values.

In **scenario A** we simulate a form of outcome reporting bias [Kirkham et al., 2010, 2018] when we assume 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 one-sided p-value. This QRP has been referred to as cherry picking [CP, Mayo-Wilson et al., 2017]. Our **scenario B** relates to the employment of questionable interim analyses [QIA, Pocock, 1977, Sagarin et al., 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. Another common practice is **scenario C**, where different covariates are added one-by-one to a simple regression model in order to get a significant result [Wicherts et al., 2016, Wang et al., 2017]. We will refer to this QRP as questionable inclusion of covariates (QIC). Scenarios A to C are derived from Simmons et al. [2011], while we decided against simulating the fourth QRP described in Simmons et al. as this practice, the flexible reporting of subsets of experimental conditions, is difficult to simulate under the alternative hypothesis: it would require specification of several effect sizes, not just one. Instead, we include another QRP which we will refer to as questionable subgroup analyses (QSA). In this **scenario D** we assume that multiple subgroup analyses are performed based on certain binary characteristics of the individuals included (gender, seniority, ...) and only the most favorable result is published, defined as the subgroup yielding the smallest one-sided p-value [Brookes et al., 2004].

Even if the effect of some QRPs on type-I error rate, *i.e* the false positive rate, has already been analysed [Simmons et al., 2011, Nosek et al., 2012, Roettger, 2019], their influence on replication success has not. 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 goal is further to investigate which is the best metric to detect replication success in the presence of different QRPs. 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 decide which metric performs best, 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.

2 Methods

A simulation study is used to compare the characteristics of four replication success metrics in the presence of QRPs. While planning our simulation study, we followed the recommendations outlined in Morris et al. [2019]. We wrote a simulation study protocol which we preregistered on the Open Science Framework 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. Note that we only consider continuous outcomes in the simulation of all the scenarios and apply throughout one-sided (one- or two-sample) t-tests.

2.1 Metrics for replication success

We will now introduce and define the four replication success (RS) metrics to be compared in detail. We will use a significance level for a one-sided hypothesis test of $\alpha = 0.025$. Note that the effect direction is taken into account by using one-sided p-values.

- The first metric is the two-trials rule (TTR), which is based on standard statistical significance. The TTR has long been custom in drug development where a drug's efficiency needs to be proven in two independent trials [Senn, 2007]. 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 TTR marks a replication as successful if

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

- The second metric to quantify replication success is a meta-analysis [MA, as used in Camerer et al., 2018]. According to the MA 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, the type-I error rate of the TTR. If p_{MA} is the meta-analytical (one-sided) p-value then we flag replication success if

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

  Original and replication studies are assumed to be exchangeable. Stouffer's method [Cousins, 2007] is used to compute the meta-analytical p-values since it is equivalent to investigating whether the overall effect of a fixed-effect meta-analysis is significant [Senn, 2007, Section 12.2.8].

- Finally, the other two metrics investigated are based on the sceptical p-value, a method that combines a reverse-Bayes approach with a prior-predictive assessment of conflict [Held, 2020, Held et al., 2022b]. The method establishes a sufficiently sceptical prior that would achieve a state in which the original result would no longer be significant. The sceptical p-value p_{S} then quantifies the conflict between the replication data and the sufficiently sceptical prior. Replication success is achieved if

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

  The sceptical p-value depends not only on the two p-value p_o and p_r, but also on the relative sample size c. The methodology has been implemented in the `R` package `ReplicationSuccess` [Held et al., 2022c]. There are two versions of the sceptical p-value which we will consider here:

  - The golden sceptical p-value, the third metric, is based on a recalibration to ensure that replication success of borderline significant original studies ($p_o \approx \alpha$) is possible, but only if there is no shrinkage of effect size [Held et al., 2022c]. This is the default method in `ReplicationSuccess`.
  - The controlled sceptical p-value, our fourth method, is a recently proposed extension that guarantees exact type-I error control [Held et al., 2022a] (in the absence of QRP).

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 the standard methods, while in the presence of publication bias the sceptical p-value performs particularly well [Muradchanian et al., 2021]. This is why we include both these metrics (the MA metric and the golden respectively controlled sceptical p-value) in our simulation study.

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 the QRP. This level of severity is interpreted differently depending on the practice. Level $k = 0$ represents the absence of any QRP. Original studies are simulated for different severities of the QRPs considered and reported ("published") only if they yield a positive and significant result. Replication studies are simulated based on the published original results, but they themselves do not include any questionable research practices. 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 400'000. The data was simulated under the assumption that only original studies with a positive and significant effect were published and later replicated. 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 four metrics described above.

2.2.1 Simulating originial studies with QRP

The QRPs considered and described in the following were simulated separately.

Simulating original studies with cherry picking - Scenario A

CP, or outcome reporting bias, is a very common QRP and a form of p-hacking [Head et al., 2015, Moran et al., 2022]. It occurs when a researcher considers several outcomes to answer a certain research question and only reports "the cherry", *i.e.* the outcome that yields the lowest p-value without mentioning the other outcomes analysed nor applying a correction for multiple testing. We will simulate this practice 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 $(k + 1) \times (k + 1)$ (with standard deviation 1 on the diagonal and correlation $\rho = 0.5$ on the off-diagonal following Simmons et al. [2011]). Let us assume $\mathbf{y}$ represents the simulated data set, then:

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

where

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

with $\boldsymbol{\theta}$ being a vector of length k with elements θ.

Note that the rows $\mathbf{y}_i$ of $\mathbf{y}$ are independent and identically distributed (iid). Next, a one-sided one-sample t-test is applied on each of the $k + 1$ columns and $k + 1$ 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:

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

We further assume that only those simulated studies indicating a significant positive effect with $p_o < \alpha$ are published and will be replicated.

Simulating original studies with questionable interim analyses - Scenario B

QIA, also called data peeking [Sagarin et al., 2014], is another commonly used QRP. More than half of the researchers participating in surveys declared to have collected more data after checking the significance of results [John et al., 2012, Agnoli et al., 2017]. For our specific scenario, we assume that the researcher 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 non-integer 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 $\mathbf{y}_1 = (y_1, \ldots, y_m)$ 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 and one-sided t-test leading to a p-value p_1. A significant result with $p_1 < \alpha$ leads to a replication study, as results would be published and we move to the next simulation. Otherwise, we assume that m more individuals are recruited and simulate $\mathbf{y}_2 = (y_{m+1}, \ldots, y_{2m}) \sim N_m(\theta, 1)$. 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 $\theta = 0.34$.

Simulating original studies with questionable inclusion of covariates - Scenario C

When the decision to include or exclude covariates in a statistical model depends on the significance of the observed result the researcher engages in questionable inclusion of covariates. To simulate QIC we need to consider two samples [Simmons et al., 2011, Roettger, 2019], *e.g.* two different treatment groups. For this particular QRP we need a larger original sample size to achieve the same power of 85% given an effect size of $\theta = 0.34$ under the alternative. For both 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 columns will be the outcomes $\mathbf{y}_a$ and $\mathbf{y}_b$ and the remaining k columns will be the covariates. $\mathbf{Y}_a$ and $\mathbf{Y}_b$ are drawn from a multivariate normal 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 (a vs. b) on the outcome $\mathbf{y} = (\mathbf{y}_a, \mathbf{y}_b)$ and will therefore apply a simple linear model with the treatment indicator as sole independent variable. This results yields a first one-sided p-value, p_0. If $p_0 < \alpha$ the researcher publishes the significant result as such and a replication study can be designed and performed. Otherwise, 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 is smaller than α. If yes those results are published. Otherwise the remaining covariates are added 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 hyptheses, H_0 and H_1.

Simulating original study with questionable subgroup analyses - Scenario D

The frequency of QSA in the literature has not yet been directly investigated, but the multiplicity problem inherent in subgroup analyses has been often described [Matthews, 2006, Chapter 9]. 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

scenario C, the first column of $\mathbf{Y}$ will be 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 one-sided t-test is applied on the outcome $\mathbf{y}$, resulting in a first p-value p_0. A significant result leads to a replication study, as results would be published. 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 with each one having sample size m_{s_j} with $j = 1, \ldots, 2k$. For instance, if $k = 3$, $\mathbf{y}$ is randomly split two times and we obtain $2k = 4$ 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, e.g. gender and age (young vs. old), and first considers the one covariate (men vs. women) to split the sample and then the other one (young vs. old). Each subgroup is analyzed separately with a one-sample, one-sided t-test resulting in $2k$ p-values. If the lowest p-value is less than α, those results would be published, and a replication study can be designed and conducted. If not, the next simulated data set is simulated.

2.2.2 Planning and simulation replication studies without QRP

Whenever the simulated original study with the questionable research practice yields a positive and significant result, a replication study is designed based on the published original results. Those published original results depend on the QRP investigated. The published sample size n'_o in scenarios A and C is simply the original sample size n_o regardless of which level of severity k was employed. For scenarios B and D, the published sample size $n'_o \leq n_o$ depends on which level yielded a significant result. For QIA, if a significant result was found after the j^{th} interim analysis, then $n'_o = j \cdot m$. For QSA 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}. Then, when replicating the originial studies with QIC, we also need to consider the number of covariates included in the published model. For each significant original study, five different strategies are employed to compute the relative sample size $c = n_r/n'_o$, where n_r is the sample size of the replication study. In particular c will either be fixed at $c = 2$ or chosen adaptively based on the original study result and the designated RS metric, as the design of replication studies should match the type of analysis [Anderson and Kelley, 2022]. Specifically, we will compute

- the required relative sample size c_{TTR} to achieve a significant positive effect in the replication study,
- the required relative sample size c_{MA} to obtain a meta-analytical p-value $p_{\text{MA}} < \alpha^2$,
- the required relative sample size $c_{\text{RS}}^{\text{golden}}$ to achieve replication success according to $p_{\text{S}}^{\text{golden}}$,
- and the required relative sample size $c_{\text{RS}}^{\text{controlled}}$ to achieve replication success according to $p_{\text{S}}^{\text{controlled}}$.

All are based on standard normality assumptions aiming to achieve a power of 85% to detect the estimate $\hat{\theta}_o$ 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 et al., 2022c,a]. Since the relative sample size c might be non-integer, the resulting replication sample size $n_r = c \cdot n'_o$ has to be rounded to the next integer. We further had to include an upper bound of $c \leq 100$ to ensure the replication study does not get unrealistically large and a lower bound of $n_r \geq 2$, as otherwise no tests can be performed.

The replication study is simulated following the same procedure as for the original study with $k = 0$. For further details and code we refer to our simulation study protocol available from the Open Science Framework.

2.3 Measures of comparison

For each QRP, each level k and each strategy for c, we start by computing the average relative effect size $\bar{d}$ (under H_1), defined as the average of the ratio between effect estimate of the replication studies $\hat{\theta}_r$ and the

effect estimate of the original study $\hat{\theta}_o$. A relative effect size smaller than 1 means that there is shrinkage of the effect. Then, to investigate which of the four metrics performs best under different levels of QRP we will follow Muradchanian et al. [2021] and compute the proportion of replication success 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, *i.e* the number of replication studies. The proportions based on the total number of simulations correspond to the overall false-positive rate or the overall power depending on whether we work under the null or the alternative hypothesis. The target value in the absence of QRP would be the squared nominal type-I error rate $\alpha^2 = 0.025^2 = 0.000625$ and the squared nominal power: $(1-\beta)^2 = 0.85^2 = 0.7225$. The proportions calculated for all the replications correspond to type-I error rate and power, respectively. In theory, the overall and standard false-positive rate should be kept low, while the overall 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 good thing, but at the same time we may also observe an increase of T1E rate, which in turn should cause concern. To combine type-I error rate and power in one measure, Bayarri et al. [2016] suggested the pre-experimental rejection ratio

$$R_{\text{pre}} = \frac{\text{Power}}{\text{Type-I error rate}}.$$

The higher this ratio, the better the performance of the metric in correctly classifying replication success. Gravestock and Held [2019] have used R_{pre} to compare different methods to incorporate historical data in clinical trials. It can be interpreted as the odds of correct rejection of the null hypothesis to an incorrect rejection of the null. The target overall R_{pre} is $(1-\beta)^2/\alpha^2 = 1'156$. Note that we will compare the results for all metrics computed using a fixed relative sample size $c = 2$. We also compare how those same metrics perform when combining them with their respective sample size calculation, *e.g.* combine the metric based on the controlled sceptical p-value with $c_{\text{RS}}^{\text{controlled}}$.

3 Results

We simulated 400'000 original studies for each of the selected questionable research practices with level of severity $k = 0, \ldots, 9$, under H_0 and H_1. Before looking at how the analysed replication success metrics perform in the presence of QRPs, we first investigate the effect the (different levels of) QRPs have on the original studies. Some of the QRPs and their effect on type-I error rate etc were already described elsewhere [Simmons et al., 2011, Roettger, 2019], but in order to fully understand the effect of the QRP on replication success we start by investigating their influence on the original studies.

3.1 Original studies with QRP

The type-I error (T1E) rate for different severity levels k are shown in Figure 1.A. This is computed as the proportion of significant original studies among all simulations under the null hypothesis. In the absence of any QRP ($k = 0$) the T1E rate in the original studies is, as expected, equal to $\alpha = 0.025$. As previously discussed, QRPs have an important effect on the T1E rate: already weak presence of QRPs ($k = 1, 2$) more than doubles the T1E rate for CP, QIA and QSA. Only the questionable inclusion of covariates does not increase the share of false positives as quickly. Figure 1.B shows the proportion of significant results under the alternative hypothesis (H_1), *i.e.* the power, depending on the severity 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 increases. We observe the fastest increase for CP and the lowest for QIA. The proportion of significant original results under the null and the alternative depending on the level of QRP in Figure S.4 in the supplement.

We show the pre-experimental rejection ratio of H_1 to H_0 in Figure 1.C. This ratio quantifies the trade-off between power and T1E rate and can be interpreted as the odds of a correct rejection of H_0 to an incorrect rejection of H_0. The ordering of the QRPs with respect to T1E rate is reversed for the pre-experimental

rejection ratio. The QRP with the strongest T1E rate increase, QSA, has the lowest rejection ratio for all levels k: for very severe QSA ($k = 9$) we observe around one false rejection for every five true rejections of the null hypothesis.

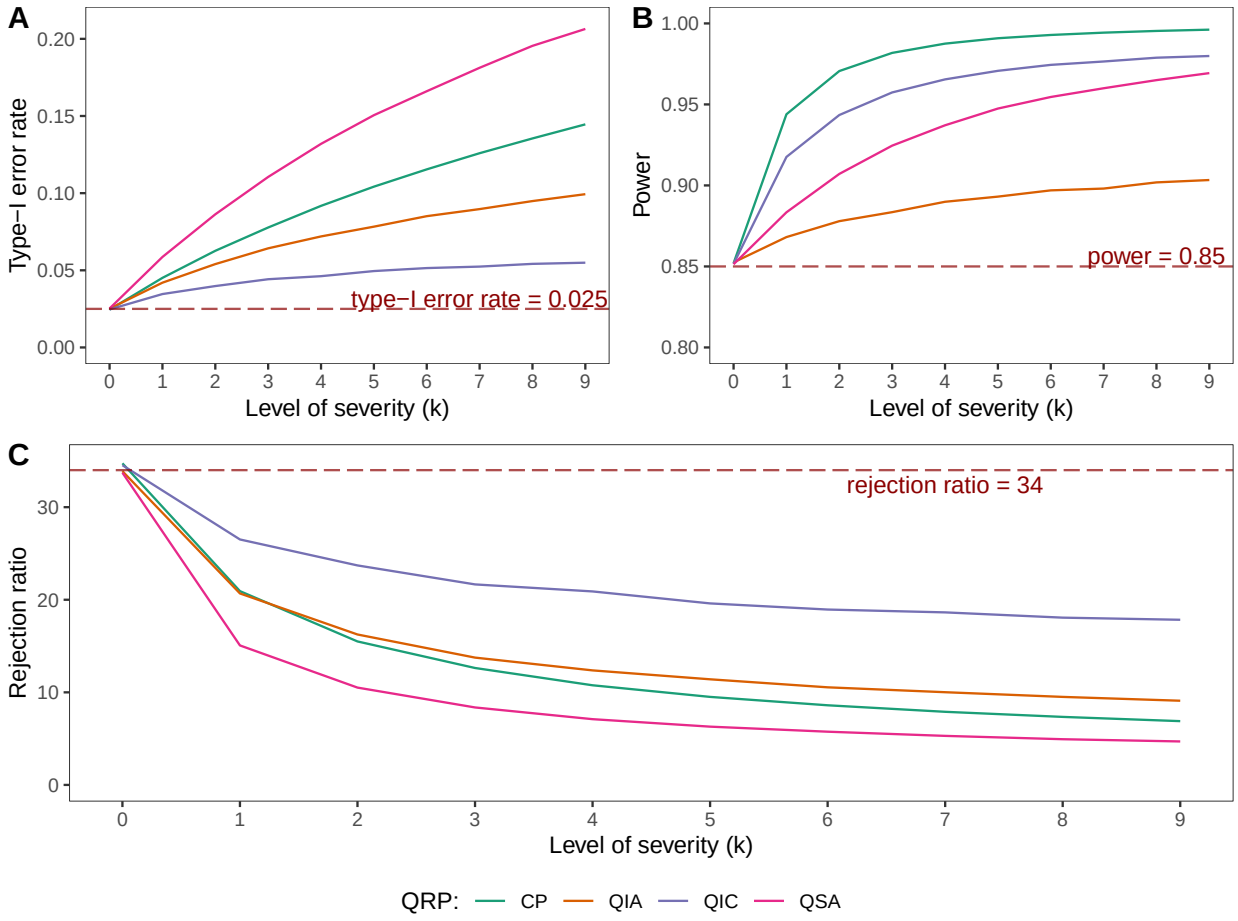

Figure 1: Original studies: The T1E rate (A), the power (B) and the pre-experimental rejection ratio (C) depending on the level of severity k and the QRP. The T1E rate is the proportion of significant results under the null hypothesis, the power is the proportion under the alternative hypothesis and the rejection ratio is their ratio.

Figure 2.A shows the original effect size observed in the studies with significant results, depending on the QRP and the level of severity, under the alternative hypothesis. The average effect size with $k = 1$ of those studies with significant results is larger than the true effect $\theta = 0.34$, illustrating the increase of effect size caused by publication bias, as we assume that only significant results are published. QIA has the strongest (positive) impact on the effect size in the original study. On the other hand, QIC negatively affects the effect size, as the additional covariates absorb some of the effect of interest. QSA leaves the effect size almost unaffected.

As previously mentioned, for QIA and QSA the published sample size of the original study can be smaller than $n_o = 80$. Figure 2.B shows the reduction of average sample size of the original studies with significant effect due to the QRPs, under the alternative hypothesis. Remember that for QIA with $k = 1$, a first test is performed on a sample of $n_o/2 = 40$. If this test turns out to be significant, we assume that the researcher stops data collection and reports and publishes a sample size of $n'_o = 40$. The average published sample size shown in Figure 2.B for $k = 2$ is around 53, a weighted average of $n'_o = 40$ and $n'_o = 80$. The published sample size of the original study drops further for QIA under the alternative hypothesis. It decreases less

fast for QSA, 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 2, but under the null hypothesis, can be found in Figure S.2 in the supplement.

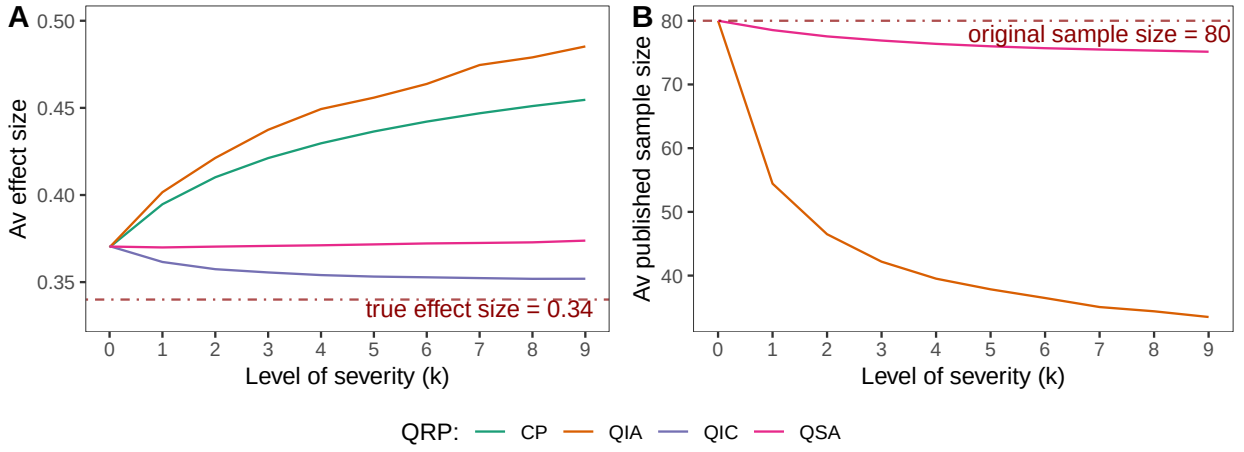

Figure 2: Average effect size in the significant original study depending on the QRP and the level of severity k, under the alternative hypothesis (A); and the average published sample size of the significant original studies for QIA and QSA depending on the level of severity k, under the alternative hypothesis (B). For CP and QIC, the published sample size stays equal to the originally defined sample size of $n_o = 80$ and $n_o = 157$ respectively.

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, sample size and p-value). As described in Section 2.2.1, we used five different strategies to calculate the sample size of the replication studies. The average relative sample size c (averaged over all designed replications) depends on the QRP, its level of severity and the chosen strategy, as shown in Figure 3 for the alternative hypothesis. We observe different implications of QRP on the replication sample sizes: more severe cherry picking leads to smaller replication sample size for all strategies except the one based on meta-analysis. The sample size calculation based on the meta-analytical criterion behaves differently here, as higher levels of cherry picking reduce the reported original p-values (see Figure S.1 in the supplement) which in turn increases the relative sample size. Larger severity levels k have a positive effect on the relative sample size c for the other QRPs when c is based on the golden sceptical p-value. Computing the replication sample size based on standard significance or the controlled sceptical p-value yields similar more constant results, while the average relative sample size calculated with the golden sceptical p-value is larger. Note that we specified an upper limit of 100 in our simulation study as c could potentially explode (*e.g.* for severe QIA the average goes up to 15). The corresponding Figure of the results under the null hypothesis can be found in the supplement (Figure S.3). Finally, after all replication studies are designed, they are simulated without QRP. Figure 4 shows the average of the relative effect sizes $d = \hat{\theta}_r/\hat{\theta}_o$, depending on the QRP and its level of severity. This Figure shows the commonly observed shrinkage effect in the CP and QIA scenario: the replication effect size is smaller than the original effect size due to bias in the original study induced by the QRP. In the QSA scenario, 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 QIC decreases and the relative effect size increases with k. We will refer to this phenomenon as "inverse shrinkage".

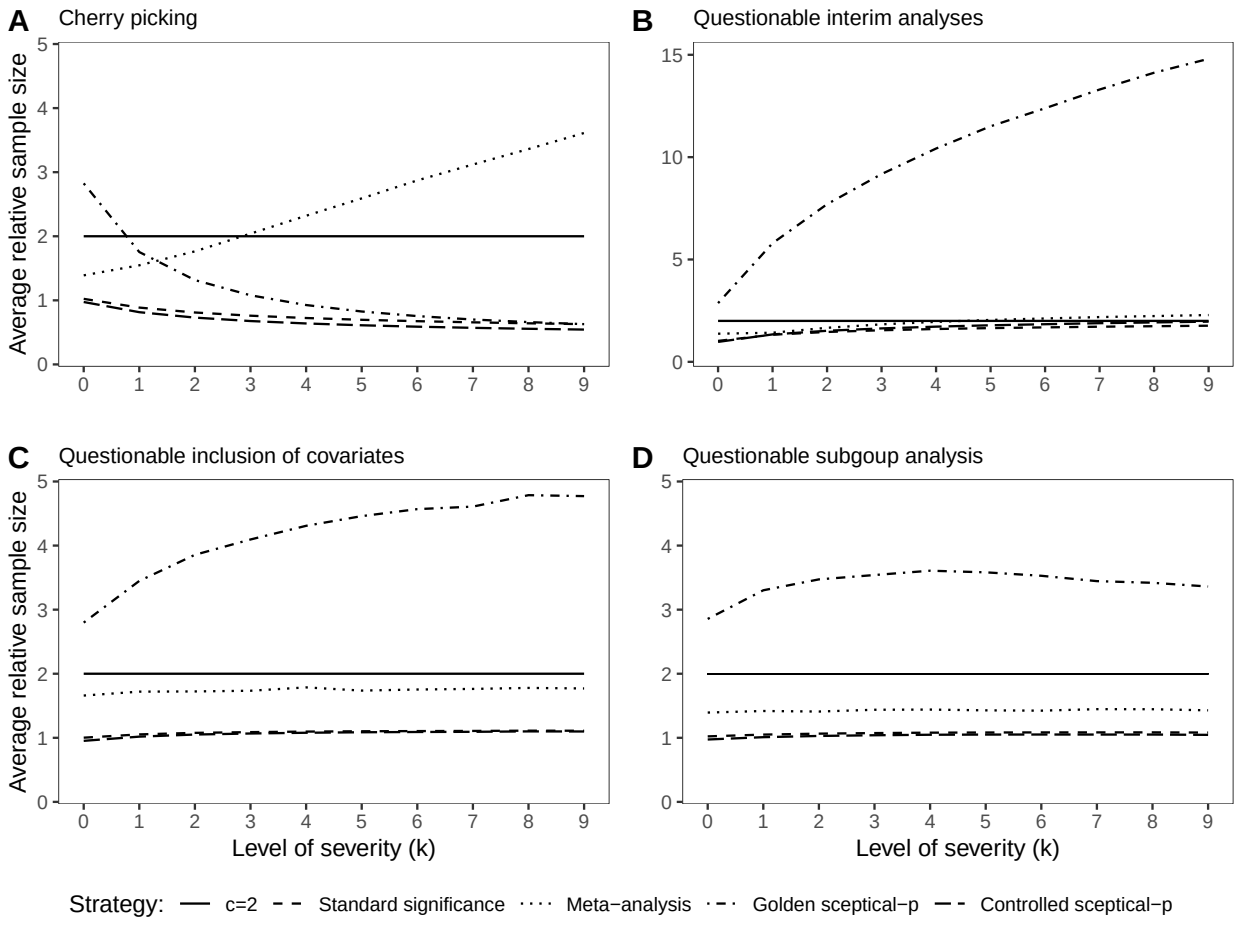

Figure 3: The average relative sample size c depending on the strategy chosen to compute the sample size for all QRP and level of severity, under the alternative hypothesis.

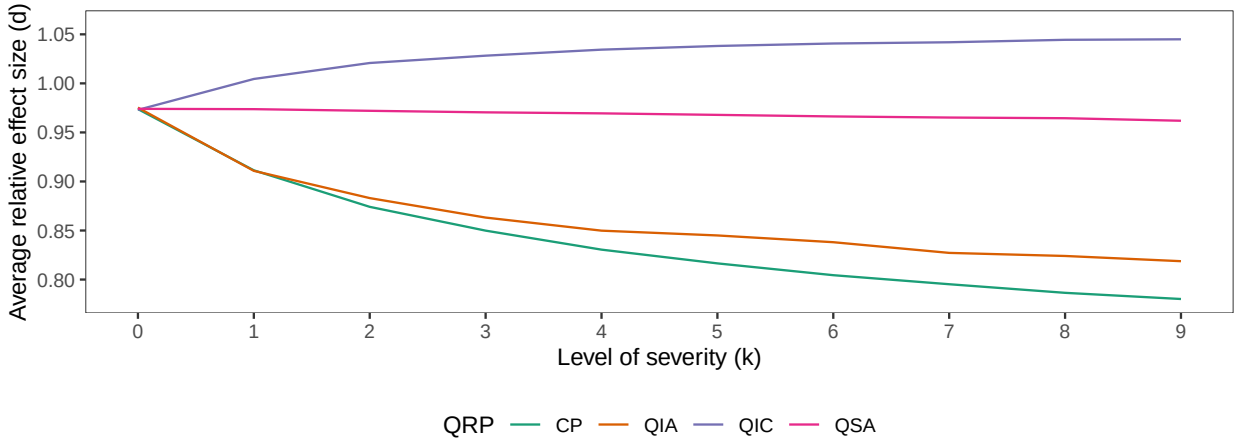

Figure 4: Average relative effect size depending on the QRP and the level of severity k, under the alternative hypothesis. For this Figure we only show the scenario with fixed $c = 2$.

3.3 Replication success

The next Section(s) will investigate which replication success metric performs better in the presence of different (levels of) QRP, in both, the null and alternative hypotheses. As discussed above, the results for all metrics with fixed replication sample size are presented together with the results where each metric is combined with its respective strategy for the computation of c (adaptive strategy, i.e. the TTR metric is used on the replication studies designed with standard significance). For each scenario the overall T1E rate, the overall power and the pre-experimental rejection ratio are shown depending on the QRP, its level of severity (k) and the strategy for c. The overall T1E rate and overall power are the rate of successful replications among all simulations ($N = 400'000$) obtained under the null (H_0) and the alternative hypothesis (H_1), respectively. Figures S.5 - S.8 in the supplement show the corresponding quantities when the proportions are computed for all replications conducted.

We reemphasize that given the design of our study only significant original studies can lead to replication success. The overall T1E rate, power and rejection ratio cannot be interpreted without considering the effect of the different QRPs on the original relative sample size, effect size T1E rate and power (as described in Sections 3.1 and 3.2).

3.3.1 Replication success in Scenario A (cherry picking)

Figure 5 shows the overall T1E rate (A); the overall power (B); and the pre-experimental rejection ratio (C) for all severity levels k and for both fixed and adaptive sample size estimation, the latter one matching the metric used in the RS assessment.

The overall false-positive rate increases with severity level k. As the overall T1E rate has been calculated over the entire simulation set, it is influenced by the increase in the number of studies that were replicated because they yield significant results. As seen in Figure 1, higher levels of k mean more false positive results and therefore more designed replication studies. Indeed, the conditional T1E rate is less affected by more severe levels of CP (see Figure S.5 in the supplement). The lowest false-positive rates are observed for all k levels, when defining replication success using $p_{\mathrm{S}}^{\mathrm{golden}}$. This result can be explained by the definition of the golden recalibration for the sceptical p-value, under which replication success cannot be observed if the relative effect size is too small. As illustrated in Figure 4 higher values of k have a negative effect on the relative effect size, *i.e* induce shrinkage. The gap in overall T1E rate computed using the golden recalibration and the remaining metrics increases with k. Regardless of the level of CP considered the two-trials rule shows the highest share of false-positives, followed by the controlled sceptical p-value, which was expected to behave similarly to the TTR, and the meta-analytical approach.

Higher severity of cherry picking positively influences the overall power (Figure 5.B) when using the meta-analytical metric to quantify replication success or the remaining metrics and doubling the sample size for the replication. The metrics other than MA with adaptive c lead to a decreased overall power once $k \geq 2$. As previously seen (in Figure 3), the average relative sample size decreases with k as the effect size of the original significant result, that should be replicated, increases. Overall, the TTR metric finds the lowest share of successful replications under H_1 when combined with c based on standard significance. Applying very severe cherry picking ($k = 9$) in the original study leads to a chance of finding a significant effect in the replication of only a bit more than 60% when defining RS using the TTR.

Finally, the pre-experimental rejection ratio (graph C of Figure 5) provides a summary of the previous results. The good performance of the golden sceptical p-value with low T1E rate is underlined by the higher R_{pre} estimated for all levels of k. In other words, for every false rejection, we observe a larger number of true rejections when the golden sceptical p-value is applied, especially when the relative sample size is equal to two. This ratio is always lower than the target value of 1'156 for $k \geq 5$, highlighting the effect of cherry picking on the replicability of studies.

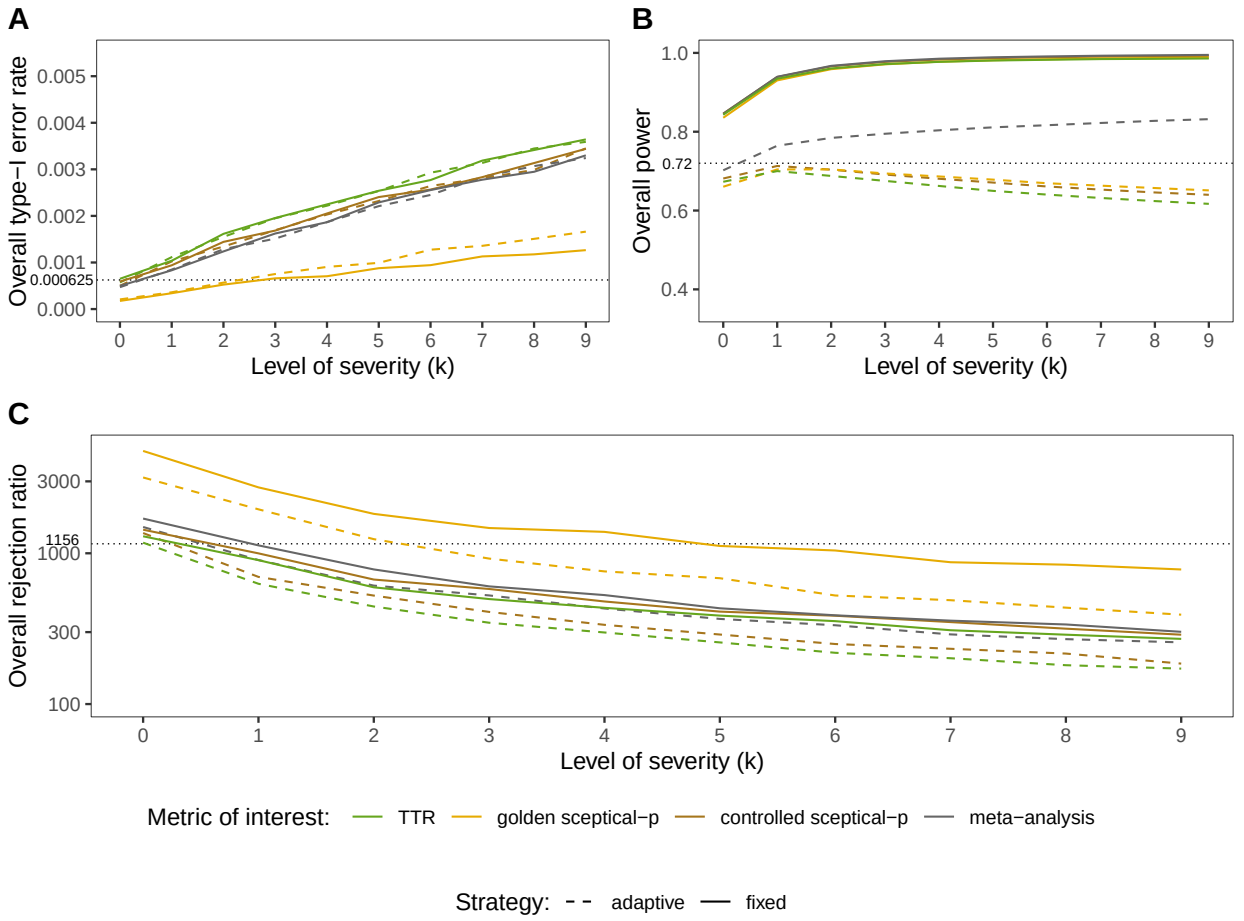

Figure 5: For **scenario A**, the overall T1E rate (A); the overall power (B); and the pre-experimental rejection ratio (C) are shown with increasing k, depending on the metric to quantify replication success and the corresponding strategy to compute the replication sample size. The target values expected given the simulation design are indicated through the dotted lines.

3.3.2 Replication success in Scenario B (questionable interim analyses)

We will now investigate the same quantities for scenario B: the overall T1E rate, the overall power, and the overall rejection ratio depending on the severity of questionable interim analyses k, the metric used to quantify replication success and the different strategies to compute the relative sample size, *i.e.* $c = 2$ ("fixed") as well as the c estimated with the strategy corresponding to the respective metric used ("adaptive").

Again, the overall T1E rate increases with the severity level k (Figure 6.A), while it increases less fast as in the presence of cherry picking. This increase in overall T1E rate is related to the increase of the false-positive rate observed for more severe QIA in the original study. We observe a clear ordering of overall T1E rate depending on the metrics for all k: the highest overall false positive rates are estimated for the TTR followed by $p_{\mathrm{S}}^{\mathrm{controlled}}$, MA, and $p_{\mathrm{S}}^{\mathrm{golden}}$. The lowest T1E rate is estimated for all k when defining replication success with the golden sceptical p-value. Again, as for cherry picking, those results might be due to the shrinkage of the effect sizes induced by the questionable interim analyses (see Figure 2.A), because the golden sceptical p-value penalizes shrinkage.

Turning to Figure 6.B a general decrease in overall power is observed for larger k for all metrics. QIA is the only QRP where a decrease in overall power is observed. As shown in Figure 2 the published original sample size (n_o') decreases rapidly with k, while the average significant effect size increases with k. Even though

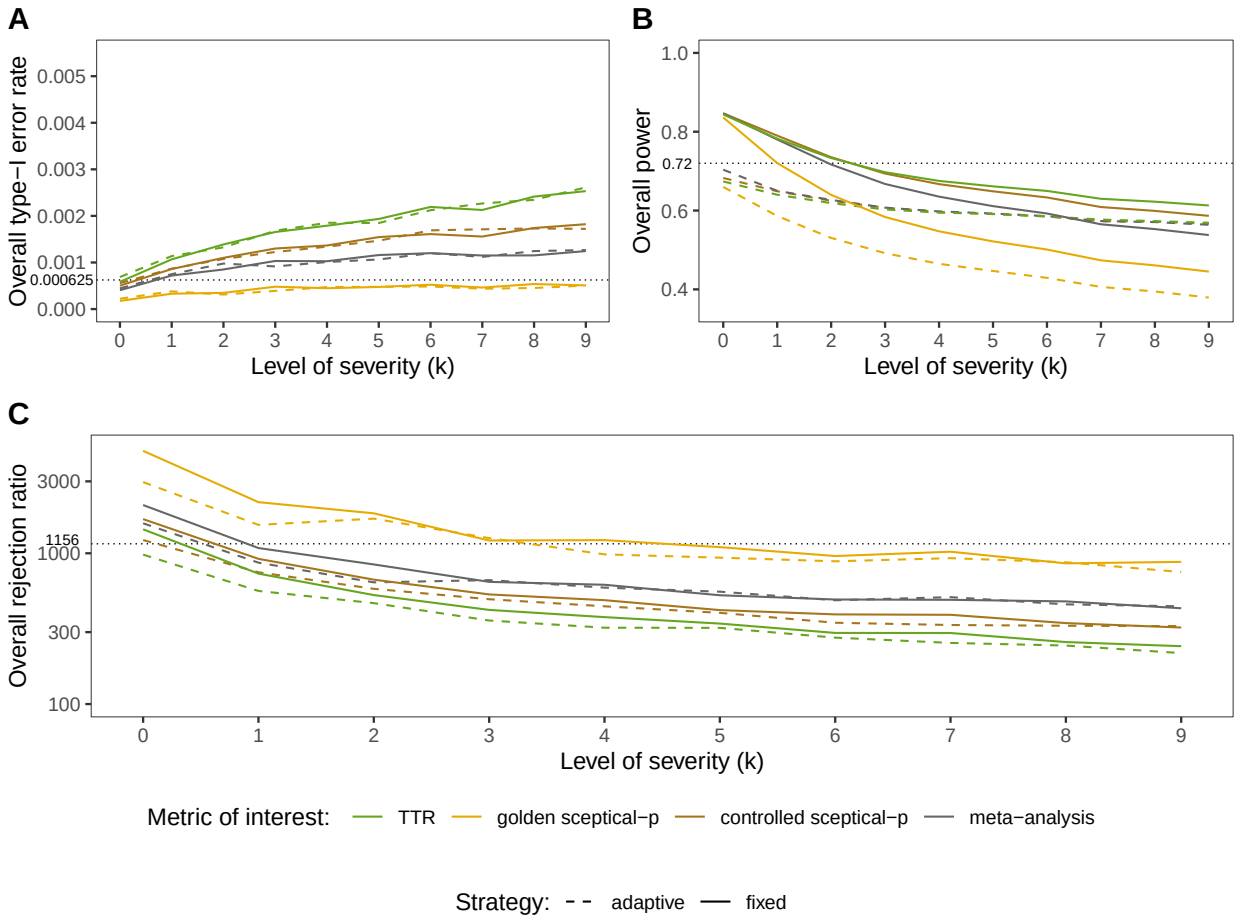

Figure 6: For **scenario B**, the overall T1E rate (A), overall power (B), and pre-experimental rejection ratio (C) for scenario B are shown with increasing k, depending on the metric used to quantify replication success and the correspondent strategy to calculate the replication sample size. The dotted lines represent the target values expected given our simulation design.

the replication studies are designed with sample sizes that are at least doubled for large k (see Figure 4), the replication studies often do not succeed to replicate the extreme events. The golden sceptical p-value penalises shrinkage and therefore produces the lowest overall power, especially with the adaptive strategy for c. For a high level of QIA, this metric estimates a less than 50% chance of observing a replication success under H_1. The pre-experimental rejection ratios (Figure 6.C) decrease with k for all metrics. As in scenario A, the highest rejection ratios are obtained using the golden sceptical p-value approach.

3.3.3 Replication success in Scenario C (questionable inclusion of covariates)

The overall T1E rate, the overall power, and the pre-experimental rejection ratio obtained in scenario C are shown in Figure 7. They will be explained depending on both the level of k, the metric used to define replication success and the strategy for c, which is either fixed to 2 or estimated with an approach corresponding to the metric used (adaptive).

The overall false-positive rate increases with k (Figure 7.A) but only very slowly. In Figure 1.A, the questionable inclusion of covariates had the slowest increase of T1E rate with k compared to the other QRP. Comparing the performance of the replication success metrics, p_S^{golden} estimates the lowest share of successful replication. The ordering is the same as for the previously discussed QRPs: golden sceptical p-value is

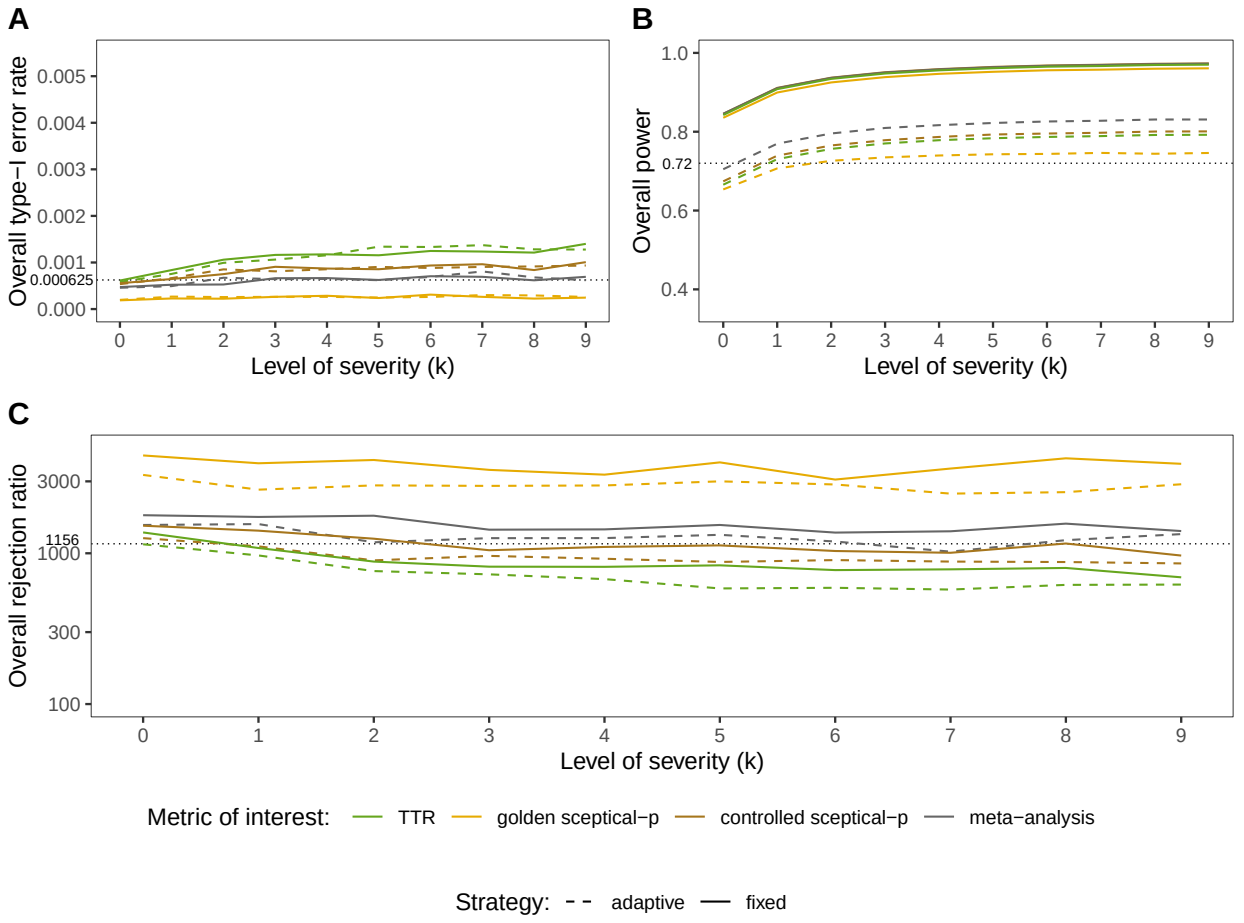

Figure 7: For **scenario C**, the overall T1E rate (A), the overall power (B), and the pre-experimental rejection ratio (C) with increasing k are shown. They depend on the metric to quantify replication success associated to the corresponding replication sample size strategy. The dotted line shows the target values expected given the simulation design.

followed by meta-analysis, then controlled sceptical p-value and TTR. Whether a fixed c level is used, or whether it is estimated with the corresponding method does not affect the ordering nor the T1E rate much. Working under the alternative hypothesis, a general increase of overall power can be observed with increased k (Figure 7.B). This result might be linked to the increase in relative effect size (see Figure 4), while the replication sample size is the same as the original sample size. Also, for this practice, the golden sceptical p-value as a metric for replication success produces the lowest overall power, if c is estimated using this same metric. The overall power quantified using the meta-analytical metric is larger than the expected target value of 0.72 for all k.

Finally, larger pre-experimental rejection ratios (graph C in figure 7) can be observed when golden sceptical p-value is applied. So, this metric ensures larger probability of estimating a true replication success than a false one. We also observe that the ratios estimated by all metrics do not decrease significantly (unlike for other practices) because the overall T1E rates are constant regardless of k level. The T1E rate for the golden sceptical p-value is and stays extremely low for all k. Under H_0 the effect size was not much influenced by k, while the relative sample size increases. The corresponding rejection ratio is very high for all k. This result is most likely due to the inverse shrinkage that is observed when applying QIC under the alternative.

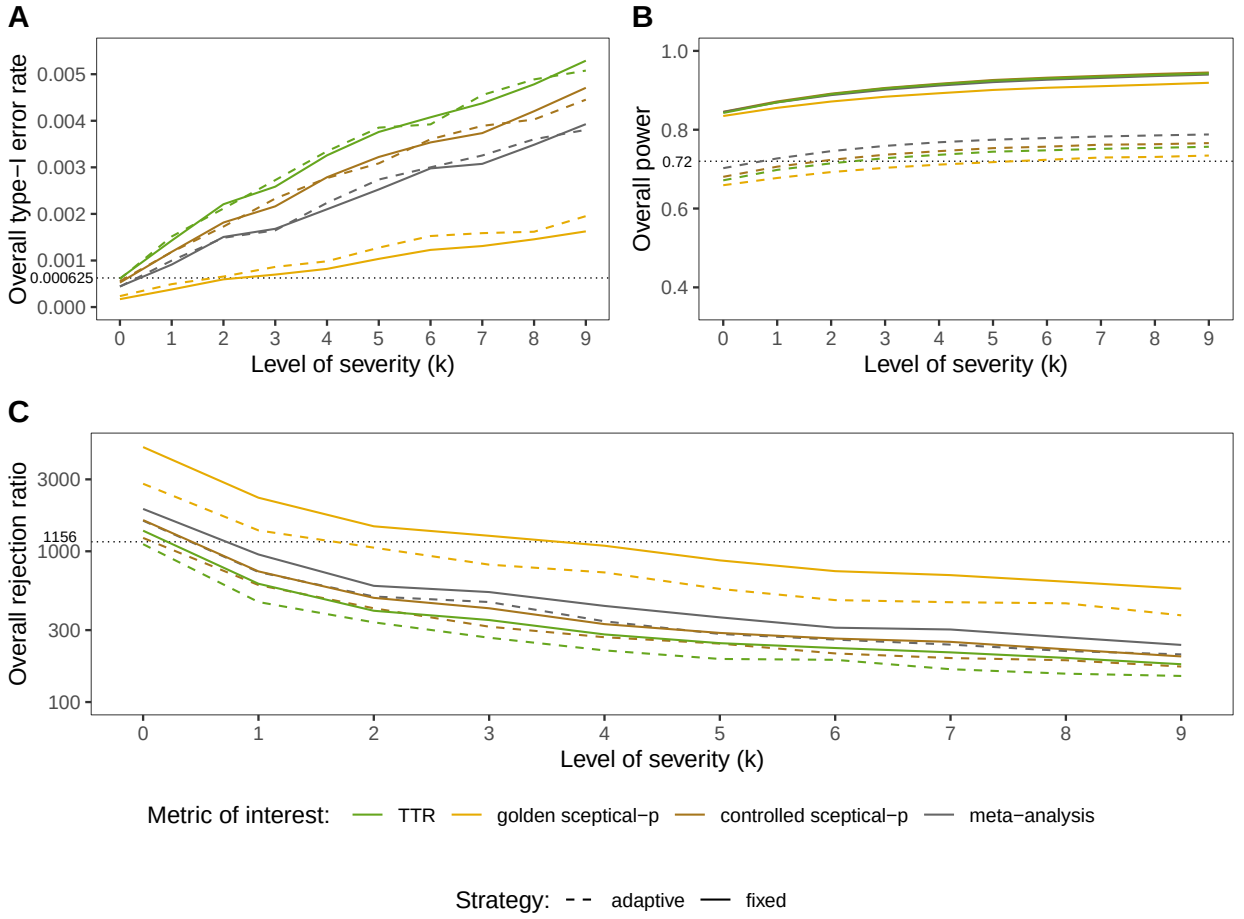

Figure 8: For **scenario D**, the overall T1E rate (A), the overall power (B), and pre-experimental rejection ratios (C) are shown for scenario D. The different levels of k and the metric to quantify replication success with the corresponding strategy to compute the replication sample size are considered. The target values expected from the simulation design are indicated through the dotted lines.

3.3.4 Replication success in Scenario D (questionable subgroup analyses)

Lastly, we will concentrate on the results obtained in scenario D. The overall T1E rate, the overall power and the overall rejection ratio depending on the level of k, the metric used to define replication success and the strategy for c (fixed to 2 or estimated with an approach corresponding to the metric used) are shown in Figure 8.

In scenario D, we observe results that are similar to the other scenarios: the false-positive rate increases with the level of k (Figure 8.A), but the increase is strongest for QSA, reflecting the large increase of the original false positive rate (in Figure 1.A). Again the ordering of the metrics by T1E rate did not change and the lowest overall false-positive rate is observed when defining replication success using the golden sceptical p-value (regardless of k). In this scenario, we also observe an increase in overall power (Figure 8.B) with increasing k linked both to the constant relative effect size for QSA (see Figure 4) and to the replication sample size equal or larger than the original sample size (see Figure 2.B). The largest overall power is observed when the relative sample size is equal to two, for all RS metrics. The overall rejection ratios (Figure 8.C) are the largest with the golden sceptical p-value.

3.3.5 Summary of results

To summarize, in all scenarios the false-positive rate increases with k because it is influenced by the increase in the number of false-positive original results. This increase is most pronounced in scenario D (QSA) followed by scenario A (CP). In all scenarios and for all k, the lowest overall T1E rate is observed when defining replication success using p_S^{golden}. The golden sceptical p-value is defined in a way that in order to flag a replication as successful both studies, the original and the replication, have to be convincing by themselves. The type of sample size calculation for the replication study does not seem to affect the false-positive rates much. However, the relative sample size c based on the golden sceptical p-value is sometimes very large and might not always be applicable.

Interestingly, one would expect the effective T1E rate with $k = 0$ to be equal to the nominal T1E rate α^2. This is indeed the case for all QRP with the TTR. The other metrics produce an effective T1E rate lower than α^2 as a replication study is only performed for original studies with significant results. The controlled sceptical p-value is defined as to exactly control this nominal T1E rate, so if non-significant results were replicated, the effective T1E rate would be equal to α^2 when $k = 0$. The T1E rate using the MA metric would be inflated if non-significant original results were replicated, as it is possible to have the MA metric flag a replication successful even if the original result was not convincing with a large p-value, if the estimated effect in the replication study is very strong (as can be inferred from Figure 9).

Under the alternative H_1, we observe shrinkage of effect size, for large k, especially when CP or QIA is applied. The sceptical p-value penalizes high levels of shrinkage. This is observed in the decrease of the overall power in scenarios A and B with this metric. In the presence of inverse shrinkage (as for QIC), on the other hand, an inflation of overall power is observed also for the golden sceptical p-value. For all QRPs but QIA, the overall power is larger when $c = 2$ as compared to a situation in which the relative sample size is estimated adaptively.

The pre-experimental rejection ratios confirm the results observed under the null hypothesis: in all scenarios, we observe higher ratios estimated by p_S^{golden} which indicates the largest number of true rejections for each false rejection (of replication success). In scenarios A, B, and D, the overall rejection ratio is smaller than the expected target value of 1'156 for large values of k, which indicates that the presence of those practices render false rejections of replication success more likely.

Regardless of the QRP studied, the meta-analytical metric to quantify RS performed relatively well, *i.e.* has high overall power. To understand why, we refer to Figure 9. Here we see for each scenario, the p-values computed in those replication studies that were successful depending on the metrics used to define RS. With the meta-analytical metric, we can obtain replication success even if the replication p-value is large. The MA metric allows such large replication p-values whenever the original study was very convincing. The MA criterion for RS does not require both studies to be "convincing on its own", in contrast to the common understanding of a successful replication.

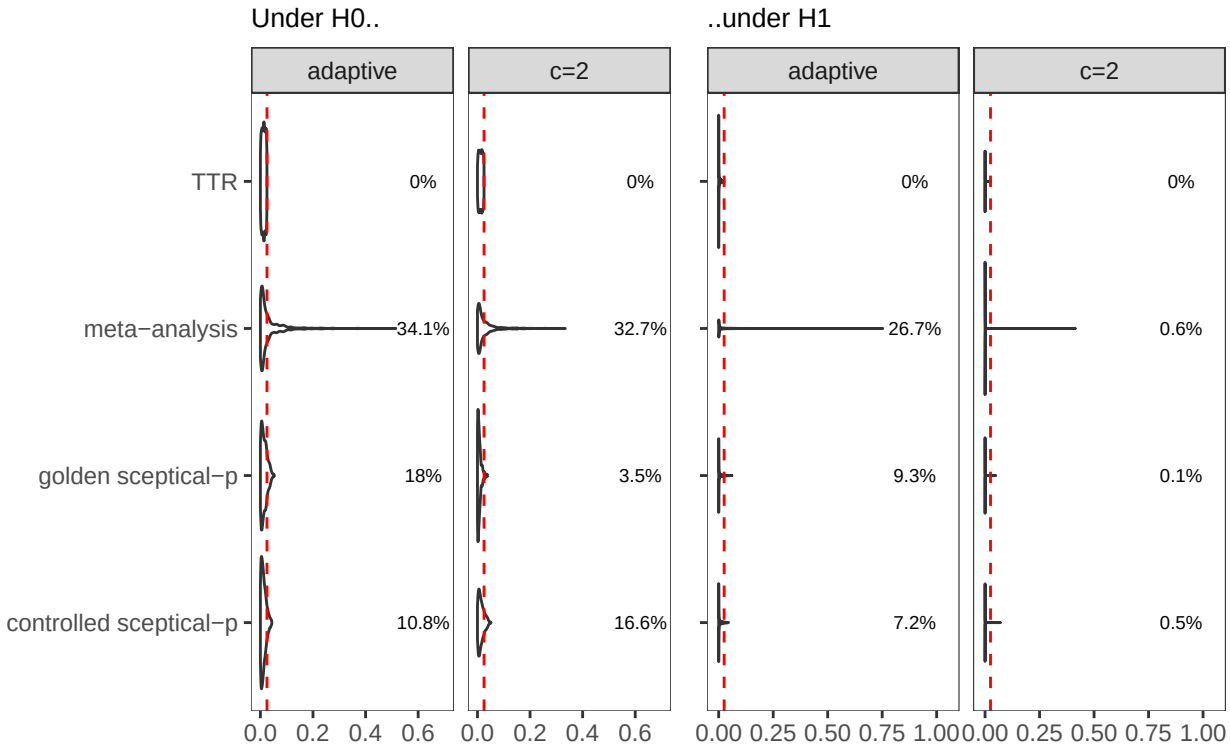

Figure 9: Violin plots of the replication p-values of those replication studies that were judged successful by the different metrics (on the y-axis), depending on whether c was fixed to 2 or estimated adaptively. The results are presented for $k = 0$, under both the null and the alternative hypothesis for all QPRs pooled together. The red dashed lines represent the significance level $\alpha = 0.025$. The percentage of values larger than α are reported.

4 Discussion

In this study, we simulated original and replication studies to compare the performance of different replication success metrics in the presence of questionable research practices. 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 are 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, and two versions of the sceptical p-value, with "golden" or "controlled" calibration, respectively. In addition, we allowed for increasing levels of severity k for each of the four questionable research practices 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 power and the rejection ratio. The design of our simulation study was preregisted on OSF.

The different (levels of) QRP have a strong effect on the operating characteristics of the original studies (T1E rate, power and rejection ratio). Also the average original effect size is affected a lot, producing a strong shrinkage effect for (severe) CP and QIA and inverse shrinkage for QIC. Therefore it is self-evident that the QRP must also influence the replicability of the original results. Using the golden sceptical p-value to define replication success leads to the smallest values of overall T1E rate for all k and QRPs and whether c is fixed to 2 or estimated adaptively. In order to have the sceptical p-value declare replication success, both studies have to be convincing enough, with respect to the original and replication p-values but also

to the relative effect size. On the other hand, the TTR might declare replication success even if a lot of shrinkage of the effect size is observed in the replication. This is especially likely if the relative sample size is larger. Interestingly, the meta-analysis performed strikingly well in the simulation study: low overall T1E rate and high overall power. This observation might be linked to the fact that we only simulated the replications of those original studies that yielded a significant result. It turns out that the meta-analytical metric flagged replications as successful that have very large p-values. If the original result is very convincing, the replication might be classified as successful, regardless of the actual result of the replication study; even if the replication presents a lot of shrinkage or a large p-value. Such replication results cannot be successful when using the golden sceptical p-value, as this metric penalizes shrinkage. Even though the golden sceptical p-value performs better, it also requires larger relative sample size. Hence, it is a trade-off that has to be considered, as c might be estimated very large making the appropriate replication unfeasible.

Regardless of the computed metric, the overall rejection ratios decrease with the severity k for CP, QIA, and QSA while it is constant in the presence of QIC. For this ratio again 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, but worse than the meta-analytic approach. This seems to be caused by the fact that the MA approach may even flag replication success if the replication study is not convincing at all.

It seems that most metrics, but in particular the golden sceptical p-value, are able to detect the effect that CP, QIA, and QSA practices have on replication success (when looking at overall power and rejection ratio). Instead, no metric can identify the effect of QIC (scenario C). Applying this practice decreases the original effect size and therefore the relative effect size increases with k, leading to inverse shrinkage with the levels of k. The golden sceptical p-value penalized shrinkage, but seems to perform less good when inverse shrinkage is observed.

This is the first study investigating the performance of different replication success metrics in the presence of a set of questionable research practices. 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 et al. [2011]. In addition, it is necessary to emphasize that in our study we simulated the QRPs, and especially QIC, following one of the multiple descriptions reported in the literature [Wang et al., 2017]. More comparisons, and even neutral comparison studies [Boulesteix et al., 2013], of the golden sceptical p-value, which was the most promising in our results, with other RS metrics are needed. Finally, in-depth analyses of the implications of the strategies to estimate the relative sample size could give insight into and recommendations on which strategy should be used in which situation.

Our study is not without limitations. We only designed and simulated a replication study if the original study showed a significant result. This does not affect the T1E rate of the two-trials rule, but it does reduce the T1E rate of all the 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 T1E rate and power [Held et al., 2022c, Section 3]. The MA approach may even flag replication success if one of the studies is not convincing at all, the restriction to significant original studies will hence also reduce T1E rate and power.

We made this choice in the assumption that a researcher performs questionable research practice only to get a significant result that can easily be published. Furthermore, conducting replication studies of non-significant 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. It would also be of interest to compare the coverage and width of the meta-analytic confidence interval with the one obtained by inverting the controlled sceptical p-value [Held et al., 2022a]. Finally, the simulation study could be extended to "many-to-one" replication designs [Klein et al., 2014].

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 4.2.0).

6 Conflict of interest

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

References

Franca Agnoli, Jelte M Wicherts, Coosje LS Veldkamp, Paolo Albiero, and Roberto Cubelli. Questionable research practices among italian research psychologists. *PloS one*, 12(3):e0172792, 2017.

S. F. Anderson and K. Kelley. Sample size planning for replication studies: The devil is in the design. *Psychological Methods*, 2022. Advance online publication.

Samantha F. Anderson and Scott E. Maxwell. There's more than one way to conduct a replication study: Beyond statistical significance. *Psychological Methods*, 21(1):1–12, 2016. URL https://doi.org/10.1037/met0000051.

M.J. Bayarri, Daniel J. Benjamin, James O. Berger, and Thomas M. Sellke. Rejection odds and rejection ratios: A proposal for statistical practice in testing hypotheses. *Journal of Mathematical Psychology*, 72:90–103, June 2016. doi: 10.1016/j.jmp.2015.12.007. URL https://doi.org/10.1016/j.jmp.2015.12.007.

Dorothy Bishop. Rein in the four horsemen of irreproducibility. *Nature*, 568(7753):435–435, April 2019. doi: 10.1038/d41586-019-01307-2. URL https://doi.org/10.1038/d41586-019-01307-2.

Anne-Laure Boulesteix, Sabine Lauer, and Manuel J. A. Eugster. A plea for neutral comparison studies in computational sciences. *PLoS ONE*, 8(4):e61562, April 2013. doi: 10.1371/journal.pone.0061562. URL https://doi.org/10.1371/journal.pone.0061562.

Sara T Brookes, Elise Whitely, Matthias Egger, George Davey Smith, Paul A Mulheran, and Tim J Peters. Subgroup analyses in randomized trials: risks of subgroup-specific analyses;: power and sample size for the interaction test. *Journal of Clinical Epidemiology*, 57(3):229–236, 2004.

Andrea Burton, Douglas G. Altman, Patrick Royston, and Roger L. Holder. The design of simulation studies in medical statistics. *Statistics in Medicine*, 25(24):4279–4292, 2006. doi: 10.1002/sim.2673. URL https://doi.org/10.1002/sim.2673.

Colin F Camerer, Anna Dreber, Felix Holzmeister, Teck-Hua Ho, Jürgen Huber, Magnus Johannesson, Michael Kirchler, Gideon Nave, Brian A Nosek, Thomas Pfeiffer, et al. Evaluating the replicability of social science experiments in nature and science between 2010 and 2015. *Nature Human Behaviour*, 2(9):637–644, 2018.

Katherine Christian, Carolyn Johnstone, Jo-Ann Larkins, Wendy Wright, and Michael R Doran. Research culture: A survey of early-career researchers in Australia. *Elife*, 10:e60613, 2021.

Robert D. Cousins. Annotated bibliography of some papers on combining significances or p-values, 2007. URL https://arxiv.org/abs/0705.2209. https://arxiv.org/abs/0705.2209.

Timothy M Errington, Maya Mathur, Courtney K Soderberg, Alexandria Denis, Nicole Perfito, Elizabeth Iorns, and Brian A Nosek. Investigating the replicability of preclinical cancer biology. *eLife*, 10, December 2021. doi: 10.7554/elife.71601. URL https://doi.org/10.7554/elife.71601.

Gowri Gopalakrishna, Gerben Ter Riet, Gerko Vink, Ineke Stoop, Jelte M Wicherts, and Lex M Bouter. Prevalence of questionable research practices, research misconduct and their potential explanatory factors: A survey among academic researchers in The Netherlands. *PloS one*, 17(2):e0263023, 2022.

Isaac Gravestock and Leonhard Held. Power priors based on multiple historical studies for binary outcomes. *Biometrical Journal*, 61(5):1201–1218, 2019. doi: https://doi.org/10.1002/bimj.201700246. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/bimj.201700246.

Megan L. Head, Luke Holman, Rob Lanfear, Andrew T. Kahn, and Michael D. Jennions. The extent and consequences of p-hacking in science. *PLOS Biology*, 13(3):e1002106, March 2015. doi: 10.1371/journal.pbio.1002106. URL https://doi.org/10.1371/journal.pbio.1002106.

Larry V Hedges and Jacob M Schauer. More than one replication study is needed for unambiguous tests of replication. *Journal of Educational and Behavioral Statistics*, 44(5):543–570, 2019.

L. Held, C. Micheloud, and F. Balabdaoui. A statistical framework for replicability. Technical report, 2022a. https://arxiv.org/abs/2207.00464.

Leonhard Held. A new standard for the analysis and design of replication studies. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 183(2):431–448, 2020.

Leonhard Held, Robert Matthews, Manuela Ott, and Samuel Pawel. Reverse-Bayes methods for evidence assessment and research synthesis. *Research Synthesis Methods*, 13(3):295–314, 2022b. doi: https://doi.org/10.1002/jrsm.1538. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/jrsm.1538.

Leonhard Held, Charlotte Micheloud, and Samuel Pawel. The assessment of replication success based on relative effect size. *The Annals of Applied Statistics*, 16:706–720, 2022c. URL https://doi.org/10.1214/21-AOAS1502.

Leslie K. John, George Loewenstein, and Drazen Prelec. Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, 23(5):524–532, April 2012. doi: 10.1177/0956797611430953. URL https://doi.org/10.1177/0956797611430953.

Jamie J Kirkham, Kerry M Dwan, Douglas G Altman, Carrol Gamble, Susanna Dodd, Rebecca Smyth, and Paula R Williamson. The impact of outcome reporting bias in randomised controlled trials on a cohort of systematic reviews. *BMJ*, 340, 2010. ISSN 0959-8138. doi: 10.1136/bmj.c365. URL https://www.bmj.com/content/340/bmj.c365.

Jamie J Kirkham, Douglas G Altman, An-Wen Chan, Carrol Gamble, Kerry M Dwan, and Paula R Williamson. Outcome reporting bias in trials: a methodological approach for assessment and adjustment in systematic reviews. *BMJ*, 362, 2018. ISSN 0959-8138. doi: 10.1136/bmj.k3802. URL https://www.bmj.com/content/362/bmj.k3802.

Richard A. Klein, Kate A. Ratliff, Michelangelo Vianello, Reginald B. Adams, Štěpán Bahník, Michael J. Bernstein, Konrad Bocian, Mark J. Brandt, Beach Brooks, Claudia Chloe Brumbaugh, Zeynep Cemalcilar, Jesse Chandler, Winnee Cheong, William E. Davis, Thierry Devos, Matthew Eisner, Natalia Frankowska, David Furrow, Elisa Maria Galliani, Fred Hasselman, Joshua A. Hicks, James F. Hovermale, S. Jane Hunt, Jeffrey R. Huntsinger, Hans IJzerman, Melissa-Sue John, Jennifer A. Joy-Gaba, Heather Barry Kappes, Lacy E. Krueger, Jaime Kurtz, Carmel A. Levitan, Robyn K. Mallett, Wendy L. Morris, Anthony J. Nelson, Jason A. Nier, Grant Packard, Ronaldo Pilati, Abraham M. Rutchick, Kathleen Schmidt, Jeanine L. Skorinko, Robert Smith, Troy G. Steiner, Justin Storbeck, Lyn M. Van Swol, Donna Thompson, A. E. van 't Veer, Leigh Ann Vaughn, Marek Vranka, Aaron L. Wichman, Julie A. Woodzicka, and Brian A. Nosek. Investigating variation in replicability. *Social Psychology*, 45(3):142–152, May 2014. doi: 10.1027/1864-9335/a000178. URL https://doi.org/10.1027/1864-9335/a000178.

John N.S. Matthews. *Introduction to Randomized Controlled Clinical Trials*. Chapman & Hall/CRC, second edition, 2006.

Evan Mayo-Wilson, Tianjing Li, Nicole Fusco, Lorenzo Bertizzolo, Joseph K Canner, Terrie Cowley, Peter Doshi, Jeffrey Ehmsen, Gillian Gresham, Nan Guo, et al. Cherry-picking by trialists and meta-analysts can drive conclusions about intervention efficacy. *Journal of Clinical Epidemiology*, 91:95–110, 2017.

C. Micheloud and L. Held. Power calculations for replication studies. *Statistical Science*, 37:369–379, 2022.

Chelsea Moran, Alexandra Richard, Kaitlin Wilson, Rosie Twomey, and Adina Coroiu. I know it's bad, but i have been pressured into it: Questionable research practices among psychology students in canada. *Canadian Psychology/Psychologie canadienne*, 2022.

Tim P. Morris, Ian R. White, and Michael J. Crowther. Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, 38(11):2074–2102, January 2019. URL https://doi.org/10.1002/sim.8086.

Jasmine Muradchanian, Rink Hoekstra, Henk Kiers, and Don van Ravenzwaaij. How best to quantify replication success? A simulation study on the comparison of replication success metrics. *Royal Society Open Science*, 8(5):201697, 2021.

Brian A Nosek, Jeffrey R Spies, and Matt Motyl. Scientific utopia: Ii. restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science*, 7(6):615–631, 2012.

Open Science Collaboration. Estimating the reproducibility of psychological science. *Science*, 349(6251), August 2015. doi: 10.1126/science.aac4716. URL https://doi.org/10.1126/science.aac4716.

Stuart J Pocock. Group sequential methods in the design and analysis of clinical trials. *Biometrika*, 64(2): 191–199, 1977.

André LA Rabelo, Jéssica EM Farias, Maurício M Sarmet, Teresa CR Joaquim, Raquel C Hoersting, Luiz Victorino, João GN Modesto, and Ronaldo Pilati. Questionable research practices among brazilian psychological researchers: Results from a replication study and an international comparison. *International Journal of Psychology*, 55(4):674–683, 2020.

Timo B Roettger. Researcher degrees of freedom in phonetic research. *Laboratory Phonology: Journal of the Association for Laboratory Phonology*, 10(1), 2019.

Brad J Sagarin, James K Ambler, and Ellen M Lee. An ethical approach to peeking at data. *Perspectives on Psychological Science*, 9(3):293–304, 2014.

Stephen Senn. *Statistical Issues in Drug Development*. John Wiley & Sons, Ltd, December 2007. URL https://doi.org/10.1002/9780470723586.

Joseph P Simmons, Leif D Nelson, and Uri Simonsohn. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, 22(11): 1359–1366, 2011.

Uri Simonsohn. Small telescopes: Detectability and the evaluation of replication results. *Psychological Science*, 26(5):559–569, 2015.

Angelika Stefan and Felix Schönbrodt. Big little lies: A compendium and simulation of p-hacking strategies. 2022. https://psyarxiv.com/xy2dk/.

Josine Verhagen and Eric-Jan Wagenmakers. Bayesian tests to quantify the result of a replication attempt. *Journal of Experimental Psychology: General*, 143(4):1457, 2014.

Y Andre Wang, Jehan Sparks, Joseph E Gonzales, Yanine D Hess, and Alison Ledgerwood. Using independent covariates in experimental designs: Quantifying the trade-off between power boost and type I error inflation. *Journal of Experimental Social Psychology*, 72:118–124, 2017.

Jelte M Wicherts, Coosje LS Veldkamp, Hilde EM Augusteijn, Marjan Bakker, Robbie Van Aert, and Marcel ALM Van Assen. Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid p-hacking. *Frontiers in Psychology*, 7:1832, 2016.

Wanja Wolff, Lorena Baumann, and Chris Englert. Self-reports from behind the scenes: Questionable research practices and rates of replication in ego depletion research. *PLoS One*, 13(6):e0199554, 2018.

Supplement for "Replication success under questionable research practices – a simulation study"

by F. Freuli, L. Held, R. Heyard

This supplementary material regroups the Figures which are displaying relevant information but were not included in the main manuscript.

Figure **S.1** shows violin plots of the p-values of the the significant original studies under H_1.

The first part of Figure **S.2** (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 QIA and QSA the average published sample size of the significant original studies under the null is represented in the second part of Figure **S.2** (B).

Figure **S.3** presents the relative sample size 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 questionable research practices and their level of severity.

The count of significant original results in both, the null and the alternative, are shown in Figure **S.4** depending on the QRP employed and the level of severity. The representations directly relate to the type-I error and the power (of the original studies).

Then, Figures **S.5** to **S.8** present the replication type-I error (A), power (B), and pre-experimental rejection rates (C). 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. Each scenario (*i.e* QRP) has its own Figure while different levels of severity k and of the relative sample size c are considered.

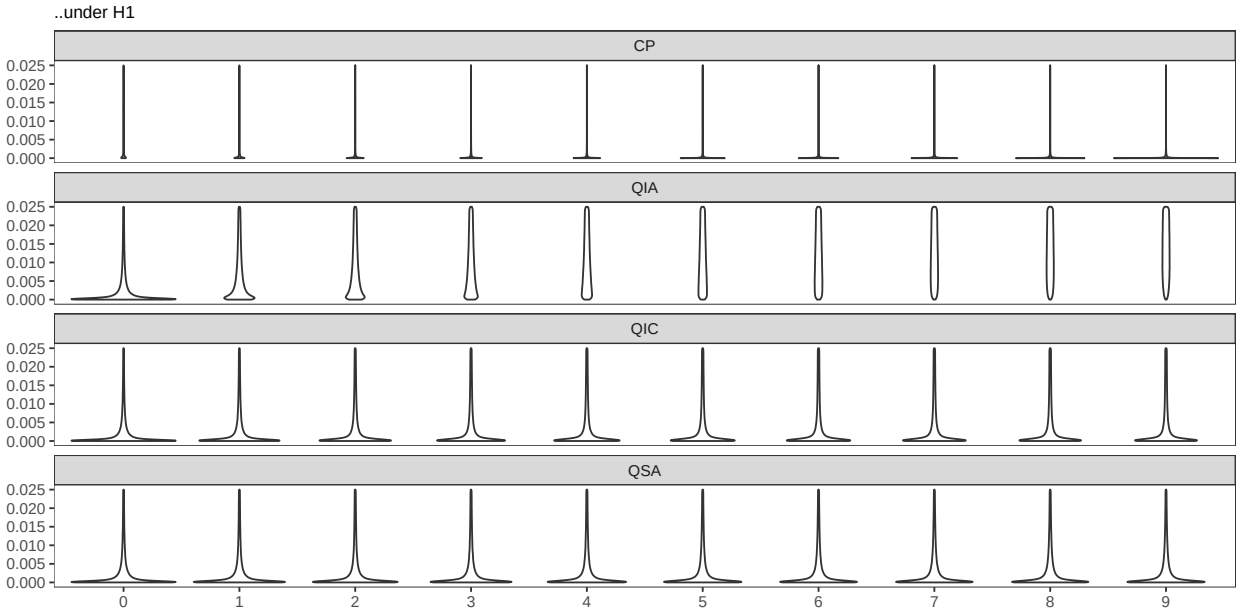

Figure **S.1**: Violin plots of the p-values of all significant original studies depending on the level of severity (on the x-axis) of the QRP.

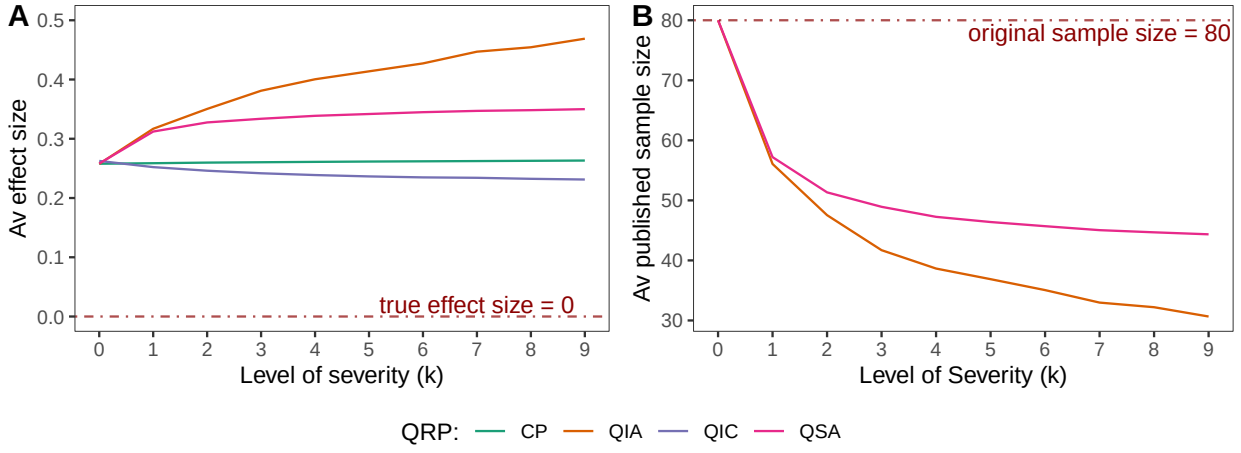

Figure **S.2**: Average effect size in the significant original study depending on the QRP and the level of severity k, under the null hypothesis (A); and the average published sample size of the significant original studies for QIA and QSA depending on the level of severity k, under the null hypothesis (B). For CP and QIC, the published sample size stays equal to the originally defined sample size of $n_o = 80$ and $n_o = 157$, respectively.

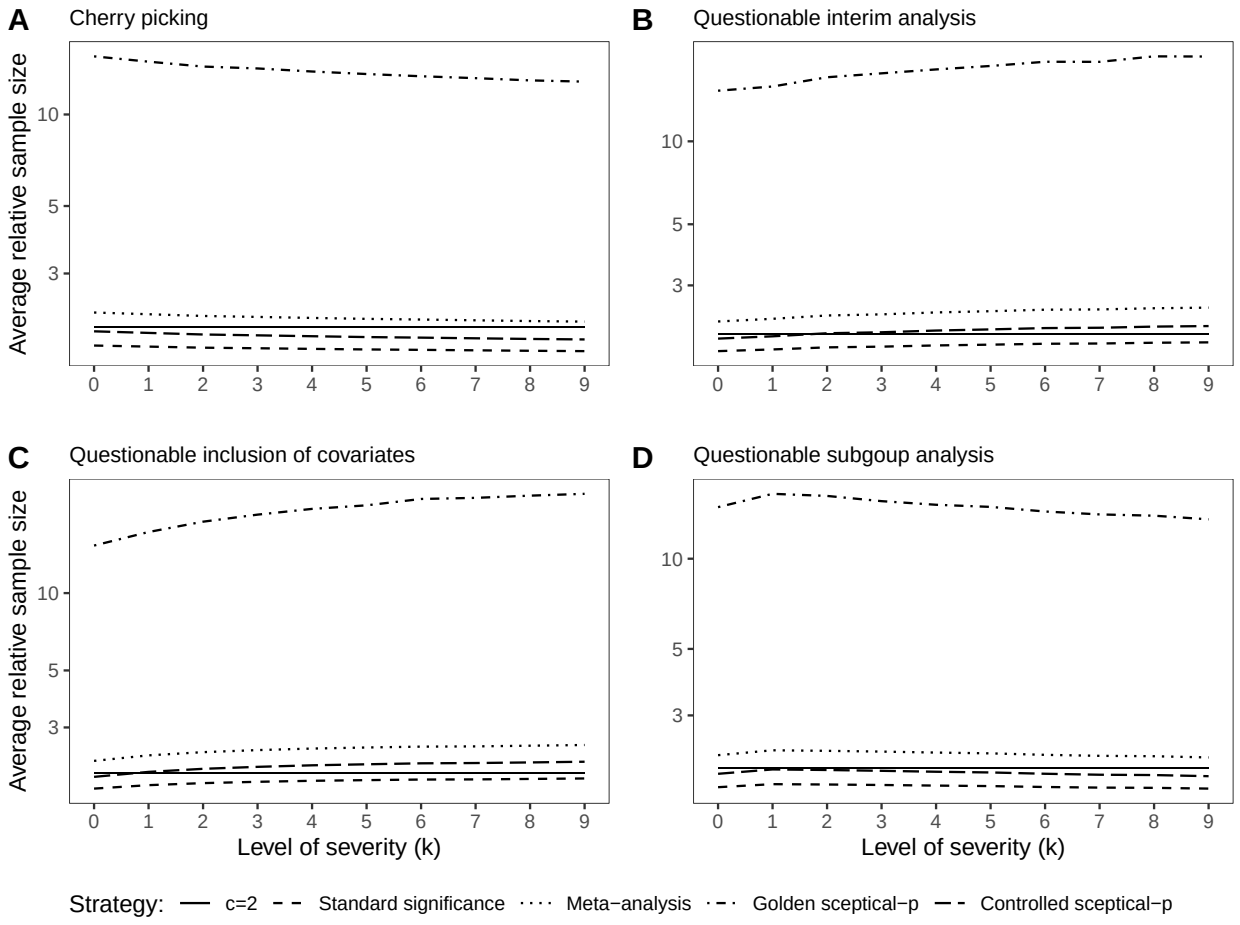

Figure **S.3**: The average relative sample size c depending on the stategy chosen to compute the sample size for all QRP and level of severity, under the null hypothesis.

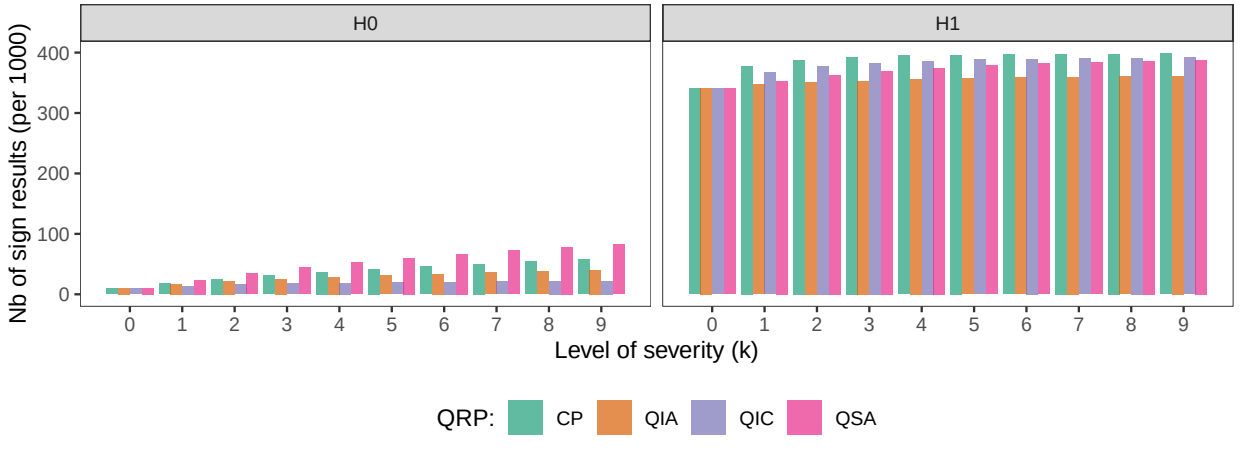

Figure **S.4**: The count of significant origincal results (per 1000), under the null and the alternative hypothesis, depending on the QP and the level of severity k.

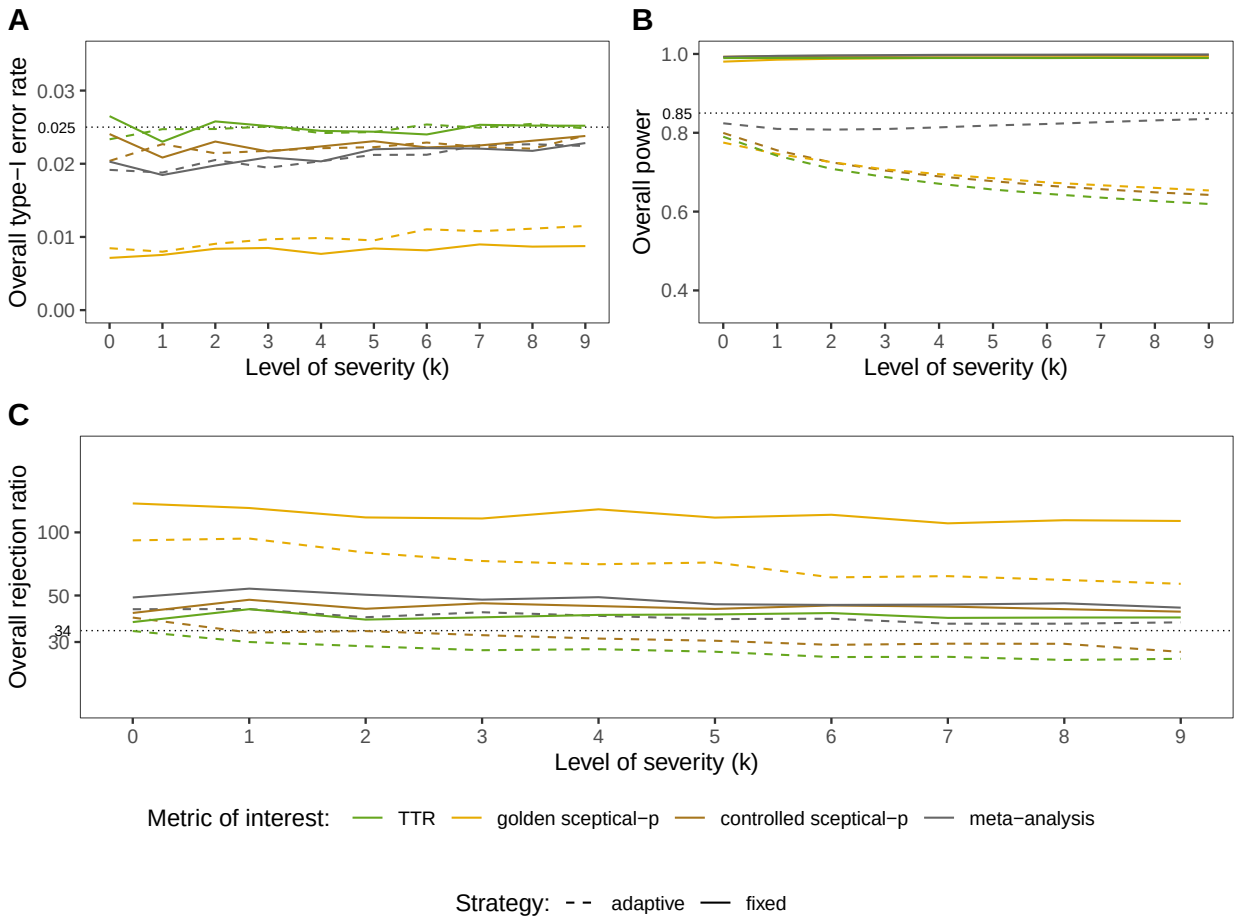

Figure **S.5**: For **scenario A**, the type-I error (A), the power (B), and pre-experimental rejection rates (C) are shown. The type-I error and the power are simply the proportion of successful replication among all the replications. The different levels of k, the metric to quantify replication success and the strategy to compute the replication sample size are considered. The target values expected from the simulation design are indicated with the dotted lines.

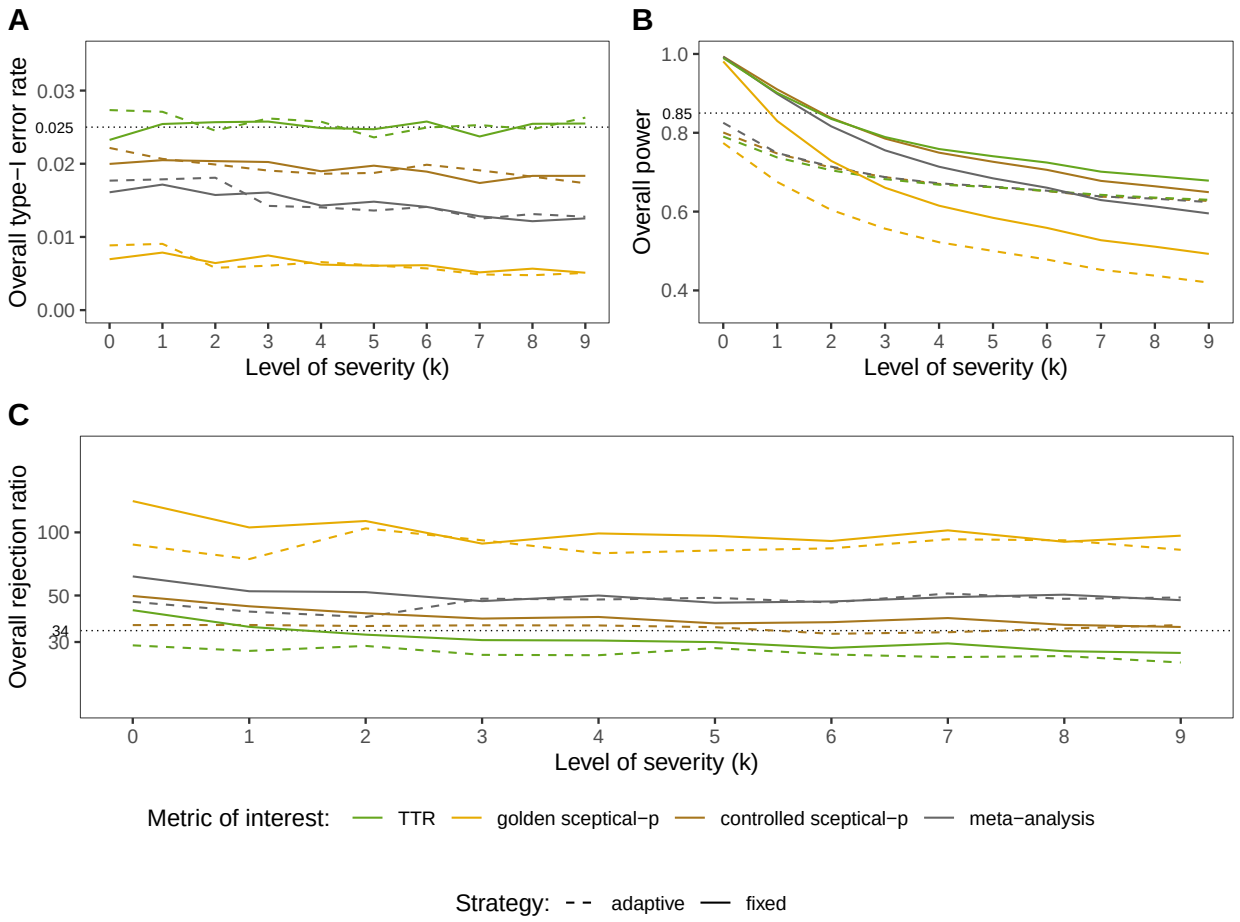

Figure **S.6**: For **scenario B**, the type-I error (A), the power (B), and pre-experimental rejection rates (C) are shown. The type-I error and the power are simply the proportion of successful replication among all the replications. The different levels of k, the metric to quantify replication success and the strategy to compute the replication sample size are considered. The target values expected from the simulation design are indicated with the dotted lines.

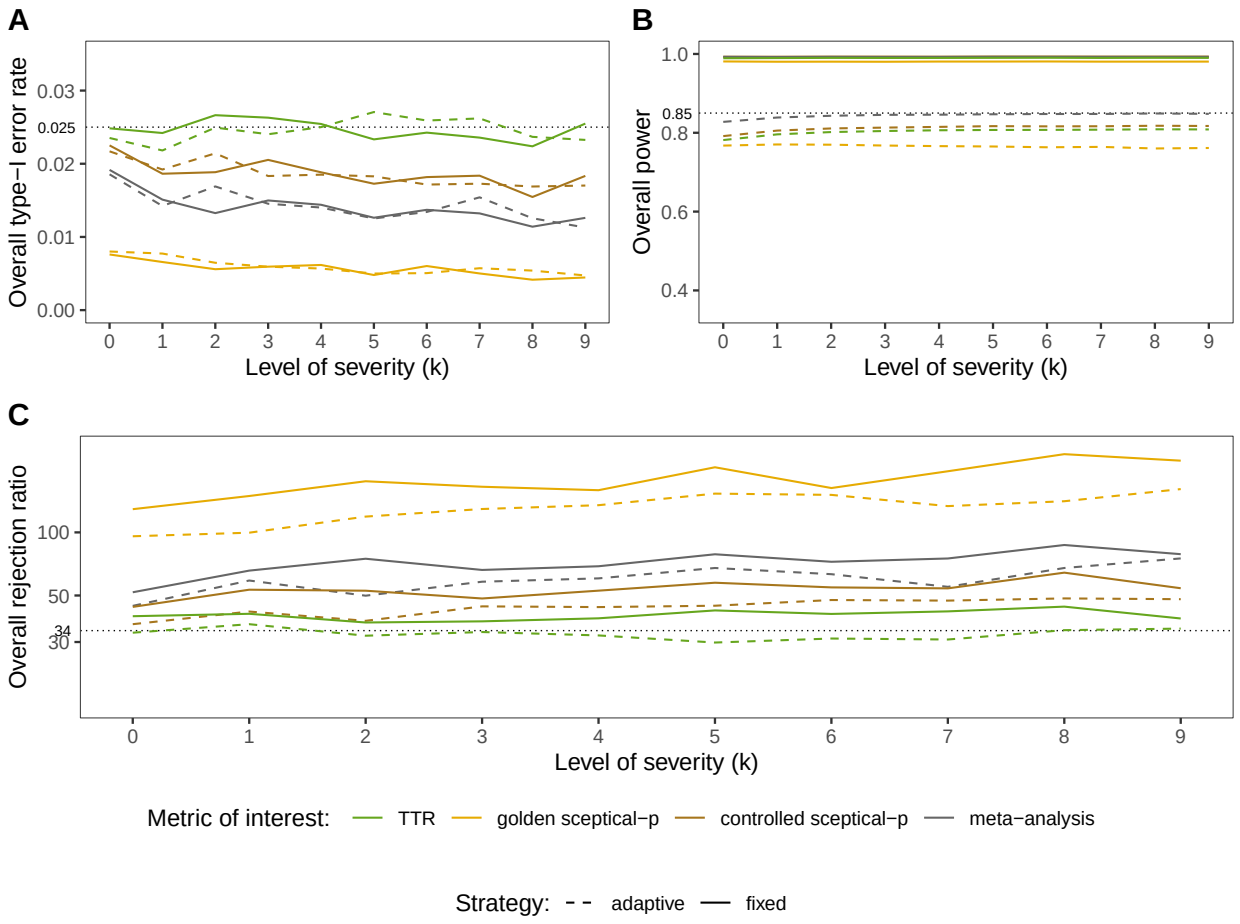

Figure **S.7**: For **scenario C**, the type-I error (A), the power (B), and pre-experimental rejection rates (C) are shown. The type-I error and the power are simply the proportion of successful replication among all the replications. The different levels of k, the metric to quantify replication success and the strategy to compute the replication sample size are considered. The target values expected from the simulation design are indicated with the dotted lines.

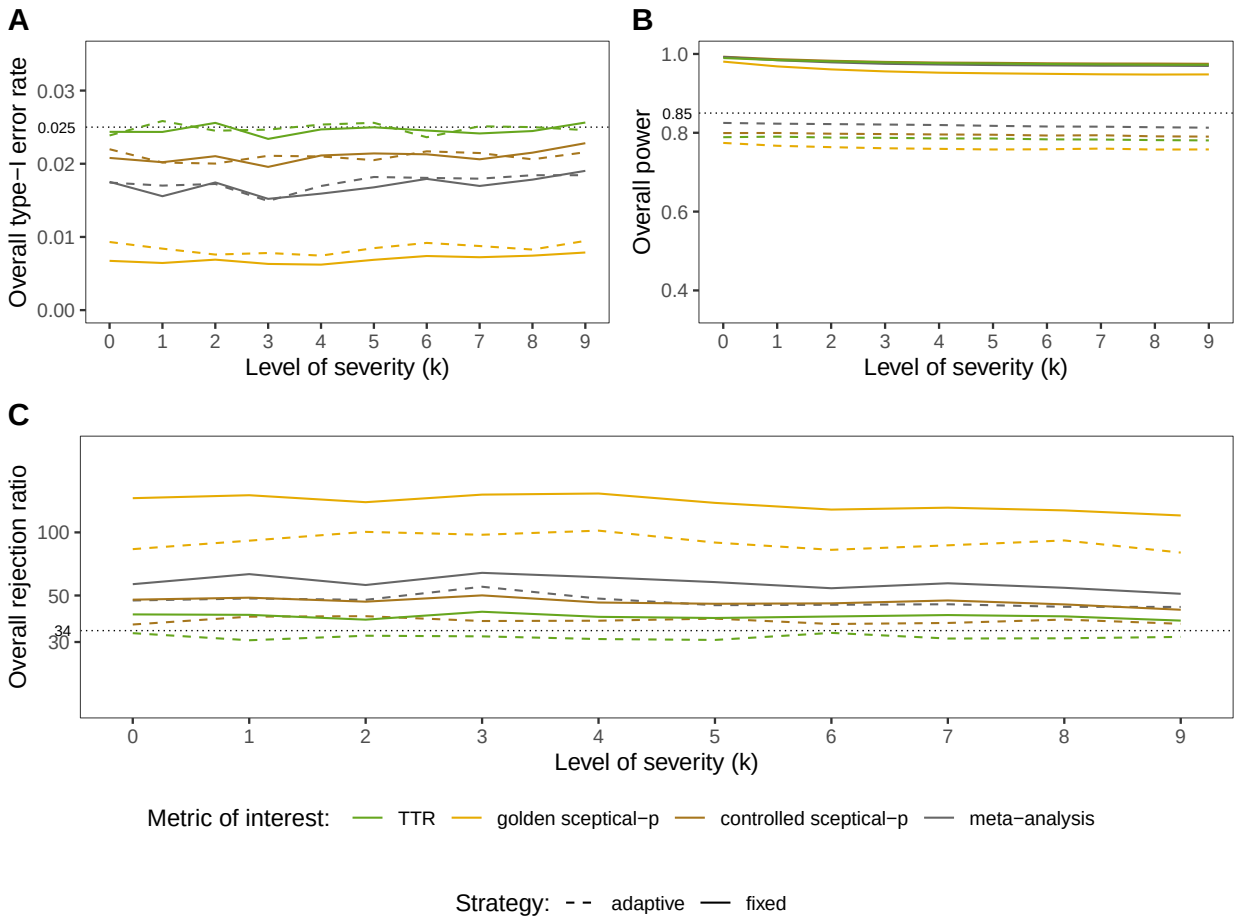

Figure **S.8**: For **scenario D**, the type-I error (A), the power (B), and pre-experimental rejection rates (C) are shown. The type-I error and the power are simply the proportion of successful replication among all the replications. The different levels of k, the metric to quantify replication success and the strategy to compute the replication sample size are considered. The target values expected from the simulation design are indicated with the dotted lines.