Sceptical Confidence Interval for Evidential Assessment of Replication Studies

---

Master Thesis in Biostatistics (STA495)

by

Xijin Chen
xijin.chen@uzh.ch

supervised by

Prof. Dr. Leonhard Held
Charlotte Micheloud

University of Zurich[UZH]

Zurich, June 2020

Abstract

The past few years have witnessed impressive efforts on the reproducibility of scientific research, which helps to adhere to good scientific results. However, there is compelling evidence that the majority of these discoveries will not stand the test of time and may fail to confirm the results of an original study. That is the so-called 'replication crisis'. There is no agreement on the standard for statistical assessment of replication success. In this thesis, the newly proposed sceptical confidence interval and sceptical p-value function can help to assess the replication success quantitatively. The comparison of the sceptical confidence interval with other alternative methods, such as the harmonic mean χ^2 confidence interval and meta-analysis, shows the properties of this approach. With the data from four collaborative and large-scale replication projects across different scientific disciplines, we evaluate the performance of these approaches under different scenarios.

Acknowledgments

The experience of working on this thesis is quite interesting to me. This research journey helps me to develop myself as a statistician and as a person. Many thanks to my supervisor Leonhard Held for his support and guidance, as well as this opportunity. I particularly want to thank Charlotte Micheloud, who has helped me patiently. I am also thankful to all my friends of the Master Program in Biostatistics, Seraphina Kissling, Uriah Daugaard, Lukas Kook, Mei-Yee Ng, Natalia Popova, Chiara Vanetta, and Samuel Pawel, in particular Katrin Petermann, who encouraged me quite a lot. I had the most amazing time with you. Besides, I would like to thank all the people in the Master Program in Biostatistics, especially Eva and Reinhard Furrer, Leonhard Held, and Torsten Hothorn. I am extremely grateful for the opportunity to do this master program and appreciate all of your work and help very much. My great gratitude also goes to my friends Yue Liu, Lina Liu, Cheng Zhong, and my parents for their continuous support.

Xijin Chen
June 2020

Contents

List of definitions

Chapter 1

Introduction

Reproducibility is a crucial principle for scientific research. The question about whether the research findings obtained from an initial study is more than just noise can often be addressed by examining whether the result can be replicated in another study. It is the subsequent evidence of the replication study that helps to support claimed research findings of the original study. Reproducibility also plays an important role in drug approval. For example, the standard 'two-trials rule' (Senn, 2008; Kay, 2014), based on two primary studies testing the same medical product, is required for drug regulation to make decisions for drug approval.

There is more than one way to conduct a replication study (Lykken, 1968; Sargent, 1981; Keppel, 1991; Anderson and Maxwell, 2016). Most types of replication try to distinguish between the direct replication and the conceptual replication (Nosek and Errington, 2017). The direct replication is defined as an attempt to recreate the conditions obtaining the previous findings, while the conceptual replication deliberately modifies important features of the study for generalization in a new way (Schmidt, 2009; Nosek and Lakens, 2014). The direct replication is more appropriate to evaluate research discoveries without violations due to procedural discrepancies as in the conceptual replication.

The past few years have witnessed great efforts on replicable research findings. There is more than one large-scale reproducibility project across multiple disciplines. For instance, Reproducibility Project: Psychology (RPP) was conducted by 270 researchers in 64 different institutions from 11 different countries in order to estimate the reproducibility of 100 studies published in three leading psychology journals in the year 2008 (Open Science Collaboration, 2015). There was only one replication for each study, and the procedure of the further replication is required to be as close to that of the original study with the aim of a direct replication. Only 36% to 47% of the original studies were successfully replicated in the RPP project according to the used criterion. Thus, many concluded that there is a 'replication crisis' in psychological science (Carey, 2015). This crisis is hardly unique to the field of psychology but emerged in several scientific disciplines, such as economics, social science, and philosophy (Camerer et al., 2016, 2018; Cova et al., 2019). In 2016, a poll conducted by the journal *Nature* reported that more than half (52%) of the surveyed scientists held the view that science was facing a replication crisis (Baker, 2016).

The fact that a particular replication may fail to confirm the discoveries of an original study is because false positive results of the original study are inevitable (Ioannidis, 2005). Several causes for the lack of reproducibility were pointed out in Benjamin et al. (2017), such as multiple testing, p-hacking, publication bias, and underpowered studies. The leading reason is convinced to be our over-reliance on the traditional Null Hypothesis Significance Testing, typically, a criterion founded on a p-value less than 0.05. Accordingly, the RPP project severely underestimates the reproducibility of psychological science because of the statistical standards of evidence (Gilbert et al., 2016).

A proper and reliable approach is required to decide whether to discard the results of an original study. Since it is impossible to know the truth with 100% in any research question,

there is no agreement on a single standard for the assessment of replication success (Stodden et al., 2014). In what follows, I list some common indicators of reproducibility:

1. Evaluation of replication studies against the null hypothesis

   It is a straightforward method with a threshold of significant results ($p < 0.05$) for the replication study, where we assume the significance of the original study holds. This indicator simply treats 0.05 as a bright-line criterion for replication success or failure, and it has been criticized a lot for misinterpretation (Goodman, 1992; Killeen, 2005; Simonsohn, 2015).

2. Comparison of the original and replication effect sizes

   It is a complementary method testing whether the replication effect is significantly different from that of the original result. Namely, it examines whether the original effect size is within the 95% confidence interval for the effect size estimate from the replication study (Open Science Collaboration, 2015). Obviously, evidence from the original study is not properly used.

3. Combination of the original and replication effect size estimates

   Meta-analysis is becoming increasingly popular for integrating research findings, as it can not only provide a more precise estimate, but also examine variability and heterogeneity between studies (DerSimonian and Laird, 1986; Haidich, 2010). However, meta-analysis is not appropriate if the goal is to support the claimed findings by another independent study (Held, 2020b).

4. Bayesian methods for quantitative assessment of the credibility of new research findings

   There are plenty of reasons to encourage the use of Bayesian methods (Matthews, 2001b). For instance, a result in favor of the null hypothesis is available. For the credibility of new research findings, Bayesian approaches are widely adopted and show a significant number of attractive properties (Verhagen and Wagenmakers, 2014; Matthews, 2018; Ly et al., 2019). Unfortunately, the Bayesian method is not friendly for researchers for the lack of convenience.

5. Sceptical p-value as an indirect measure for the replication success

   The sceptical p-value proposed by Held (2020b) helps to address the weaknesses mentioned above. It evaluates our scepticism about the initial findings via evidence from the replication study. This Bayes-non-Bayes compromise (Good, 1992) holds advantages of both Bayesian inference and frequentist inference, and frustrates those who would prefer quick answers to the replication success assessment. However, it is not able to illustrate the magnitude and direction of effect size. Besides, it may result in an unwanted claim of replication success.

In this thesis, the proposed sceptical confidence interval for the assessment of replication success is an extension of the sceptical p-value method. The ordinary confidence interval takes the uncertainty around the effect size into account. Likewise, the sceptical confidence interval is preferable to the sceptical p-value for the assessment of replication success, since the sceptical confidence interval carries more information about the direction and strength of the investigated effects. Moreover, conflicts between the original and the replication study are accessible.

The structure of this thesis proceeds as follows: Chapter 2 describes the development of the sceptical confidence interval. Additionally, alternatives such as the harmonic mean χ^2 confidence interval (Held, 2020a) and meta-analysis are presented. Chapter 3 demonstrates implementations of the sceptical confidence interval, and the comparison of different strategies is then illustrated. Finally, this thesis is closed with discussions in Chapter 4.

Chapter 2

Methods

In this chapter, notations used in this thesis, and the general framework are introduced. Then development of the sceptical confidence interval and other alternative replication success assessment methods are also illustrated.

2.1 Notation

General notations in this master thesis are summarized in Table 2.1. Both the original and the replication studies are required for the assessment of replication success. For general quantities, we suppose θ is the true and unknown effect size. Effect size estimates of the original and the replication study are expressed as $\hat{\theta}_o$ and $\hat{\theta}_r$. Corresponding standard errors are denoted by σ_o and σ_r, respectively. The variance ratio is defined as $c = \sigma_o^2/\sigma_r^2$, which can be equally expressed as the relative sample size n_r/n_o, since $\sigma_o^2 = \kappa^2/n_o$ and $\sigma_r^2 = \kappa^2/n_r$, where κ^2 is the unit variance from one observation. The concept of the sufficiently sceptical prior with mean value μ is the core issue for the sceptical confidence interval, and it will be introduced in Section 2.3.1. The two-sided significance level is denoted by α, and corresponding $z_{\alpha/2}$ is the $\alpha/2$-quantile of the standard normal distribution. Notations for different methods for the replication success assessment discussed in Section 2.3 and Section 2.4 are also listed below.

Table 2.1: Notations

	Original study	Replication study
Treatment effect estimate	$\hat{\theta}_o$	$\hat{\theta}_r$
Standard error	σ_o	σ_r
Test statistic	t_o	t_r
Ordinary p value (two-sided)	p_o	p_r
Relative sample size	$c = n_r/n_o$ (σ_o^2/σ_r^2)	
True and unknown effect size	θ	
Mean of sufficiently sceptical prior	μ	
Two-sided sceptical p-value and confidence interval	p_S, CI_S	
One-sided sceptical p-value and confidence interval	$\widetilde{p}_S$, $\widetilde{CI}_S$	
Harmonic mean χ^2 p-value and confidence interval	p_H, CI_H	

2.2 General framework

This section discusses the essential concepts required for the understanding of the sceptical confidence interval and related methods in this thesis.

2.2.1 Normal distribution and χ^2 distribution

The normal distribution and the χ^2 distribution are widely used probability distributions in inferential statistics, especially in hypothesis testing and in the construction of confidence intervals. The introduction of these two distributions is taken from Held and Sabanés Bové (2014).

Suppose x stands for the realization of a random variable X that follows a normal distribution with mean μ and variance σ^2, say $X \sim N(\mu, \sigma^2)$. The probability density function for x can be expressed as

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right\}.$$

The normal probability distribution is fundamental to statistical analysis. In this thesis, we assume that an effect size θ can be modeled as normally distributed after suitable transformations, see Section 2.6.

The χ^2 distribution with d degrees of freedom is the distribution of a sum of the squares of d independent standard normal random variables. If $X_i \sim N(\mu, \sigma^2)$, i=1,...,d, are independent, then $\sum_{i=1}^{n} x_i^2 \sim \chi^2(d)$. The corresponding probability density function of x is

$$f(x) = \frac{(\frac{1}{2})^{\frac{d}{2}}}{\Gamma(\frac{d}{2})} x^{\frac{d}{2}-1} \exp(-x/2),$$

where $\Gamma(x)$ is the gamma function.

The χ^2 distribution works in the case of multiple random variables, namely multiple studies.

2.2.2 Standard error and confidence interval

This section about the theorem of standard error and confidence interval is taken from Held and Sabanés Bové (2014).

Generally, an estimator $\hat{\theta}$ will not be equal to the true parameter θ. However, it will often be close to θ in a particular sense. The statistical variability of an estimator $\hat{\theta}$ can be quantified via the variance τ^2 or the standard deviation $\sqrt{\tau^2}$ of $\hat{\theta}$. Even though τ^2 is always regarded as unknown, we are able to estimate it with the consistent estimate of the standard deviation, say, the standard error. The standard error helps to quantify how much an estimator will vary under hypothetical repeat of the sampling procedures independently.

Based on the standard error, the concept of the confidence interval is available. A $\gamma \cdot 100\%$ confidence interval for θ is defined by confidence limits, $\hat{\theta}_l = h_l(X_{1:n})$ and $\hat{\theta}_u = h_u(X_{1:n})$, where $X_{1:n}$ is a random sample and $\gamma \in (0, 1)$ is fixed. That is

$$\Pr(\hat{\theta}_l \leq \theta \leq \hat{\theta}_u) = \gamma,$$

for all $\theta \in \Phi$. The relationship of these two statistics can be assumed as $\hat{\theta}_l \leq \hat{\theta}_u$.

For repeated random samples from a distribution with unknown parameter θ, a $\gamma \cdot 100\%$ confidence interval will cover θ in $\gamma \cdot 100\%$ of all cases.

2.2.3 Null Hypothesis Significance Testing (NHST) and the p-value

It is often of interest to quantify the evidence against a specified null hypothesis H_0 given the observed data. Such a null hypothesis can often be represented by a specific value θ_0 of the

unknown parameter, i.e., $H_0 : \theta = \theta_0$. For instance, in clinical studies, a common null hypothesis is that there is no difference in the effect of two treatments. As indicated in Pernet (2015), the statistical inference method, NHST, necessarily, is a combination of the significance test by Fisher (1925) and the test of acceptance based on critical rejection regions in Neyman and Pearson (1928).

The p-value is the primary statistical end product of NHST, and it is defined as the conditional probability of the observed or more extreme data given H_0. Namely, it serves to quantify the evidence against H_0. As indicated in Chapter 1, the p-value has been realized to be an unreliable indication of the magnitude of an effect for decades (Bracey, 1991; Keren and Lewis, 1993; Cohen, 1994). Recently, there is a growing realization that reported 'statistically significant' claims based on a p-value in scientific publications are routinely mistaken (Gelman and Loken, 2014). Shortcomings of NHST are likely contributing factors behind this replication crisis across multiple disciplines (Szucs and Ioannidis, 2017). To summarize, most of the scientific results are mistaken due to the convenient, yet ill-founded strategy on the basis of statistical significance (Ioannidis, 2005). For this reason, a group of 72 authors suggested to redefine statistical significance to avoid a high rate of false positives (Benjamin et al., 2017). Afterwards, this suggestion was argued by different opinions (Lakens et al., 2018; McShane et al., 2019).

2.2.4 Bayesian analysis with normal distributions

There is a growing awareness of Bayesian methods (Matthews, 2001b), which was said to have a number of intrinsic advantages over conventional methods (Lilford and Braunholtz, 1996; Bland and Altman, 1998; Spiegelhalter et al., 1999). The discussion of Bayesian analysis in this section is originated from Spiegelhalter et al. (2004).

Bayesian analysis, the so-called 'prior-to-posterior' analysis, can be simply regarded as a process of update from the prior to the posterior, in the light of the evidence from observed data. We denote the prior distribution of the unknown parameter θ as $p(\theta)$ and the observed evidence (i.e., the results of a clinical trial) as y. We are supposed to obtain the new probability of the posterior for different values of θ, taking account of the evidence y,

$$p(\theta|y) = \frac{p(y|\theta)}{p(y)} \times p(\theta).$$

Now $p(y)$ is just a normalizing factor to ensure that $\int p(\theta|y)d\theta = 1$, and its value is not what we are concerned about. The essence of Bayes theorem only considers the terms including θ, and hence it is often written as

$$p(\theta|y) \propto p(y|\theta) \times p(\theta),$$

which says that the posterior distribution is proportional to the product of the likelihood and the prior. Thus it can be expressed as

$$\text{Prior insight} + \text{Data likelihood} \rightarrow \text{Posterior insight}, \tag{2.1}$$

where the plus sign can be regarded as the combination of evidence.

2.2.5 Sceptical priors

Bayesian analysis is driven by the prior distribution, which expresses the personal opinions or the results of previous studies. The procedure of producing a posterior distribution is fixed, and the outcome of Bayesian analysis is highly dependent on the choice of the prior distribution. As summarized by Spiegelhalter et al. (2004), there are five broad approaches in general: Elicitation of subjective opinion; Summary of past evidence; Default priors; 'Robust Priors'; Estimation

of priors using hierarchical models. The sceptical prior is a widely-used default prior, with no consideration of all external information, personal opinions and suggestions from experts.

The distribution of a sceptical prior in Figure 2.1 is defined by specifying the best estimate as 0. Thus, the probability that an effect is more extreme than, or at least as extreme as the alternative hypothesis effect size estimate θ_A is small, i.e., 5% (Fayers et al., 1997). In other words, sceptical priors reflect the scepticism about large treatment effects. Only fairly solid evidence will be recommended by a careful sceptic (Kass and Greenhouse, 1989; Spieglehalter et al., 1994).

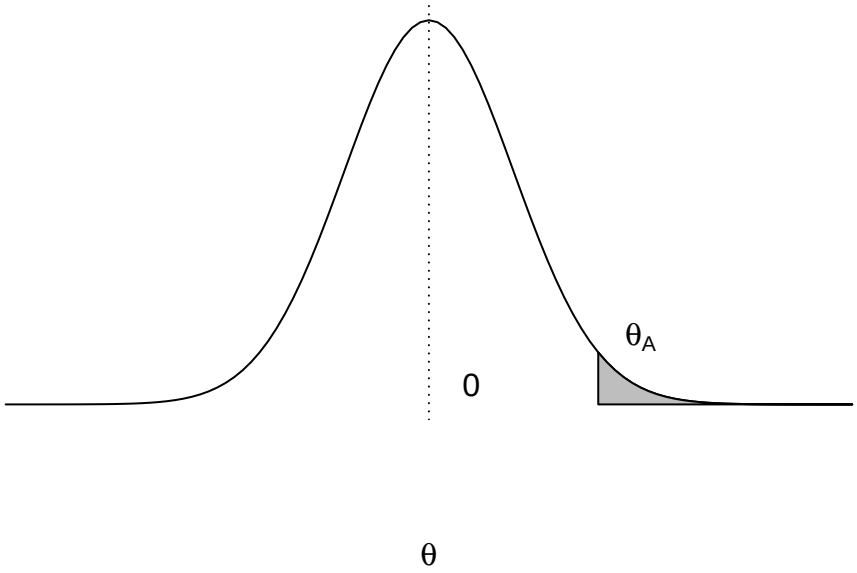

Figure 2.1: Sceptical prior with alternative hypothesis of effect esitmate θ_A. The sceptics' probability that the true difference is greater than θ_A is the shaded area.

Sceptical priors have been widely-used in a great number of case studies (DerSimonian and Laird, 1986; Fletcher et al., 1993; Freedman et al., 1994; Parmar et al., 1994; Heitjan, 1997; Higgins and Spiegelhalter, 2002).

2.2.6 Prior-predictive distribution

The prior-predictive distribution measuring the conflict between priors and subsequent data was firstly suggested by Box (1980). A predictive distribution for future observations can be derived based on a prior and the assumption of normal distributions (Spiegelhalter et al., 2004).

Suppose we have a normal distribution for some future data $Y_n \sim N\left[\theta, \sigma^2/n\right]$, where θ is the mean, σ is the standard deviation, and n is the sample size. A prior distribution in a similar form can be obtained $\theta \sim N\left[\mu, \sigma^2/n_0\right]$. The standard deviation σ is the same as the likelihood, and the prior is based on an 'implicit' sample size n_0. The advantage brought by this expression will be obvious for the prior-to-posterior analysis. In order to make predictions concerning future values of Y_n, taking into account our uncertainty about its mean θ. We can consider $Y_n = (Y_n - \theta) + \theta$, and thus Y_n can be regarded as the sum of two independent measures: $Y_n - \theta \sim N\left[0, \sigma^2/n\right]$ and $\theta \sim \left[\mu, \sigma^2/n_0\right]$. The sum of two independent normally distributed quantities is also normal, with the sum of the means and the variances, and hence Y_n will have a predictive distribution,

$$Y_n \sim N\left[\mu, \sigma^2\left(\frac{1}{n} + \frac{1}{n_0}\right)\right].$$

Given observed y_n, the predictive probability of observing a random variable Y_n less than that observed value y_n should be

$$P(Y_n < y_n) = \Phi\left(\frac{y_n - \mu}{\sigma\sqrt{\frac{1}{n_0} + \frac{1}{n}}}\right),$$

and hence the corresponding p-value can be obtained via the tail area associated with a standardized test statistic contrasting the prior and the likelihood,

$$z_n = \frac{y_n - \mu}{\sigma\sqrt{\frac{1}{n_0} + \frac{1}{n}}}, \tag{2.2}$$

showing that the Box's statistic in Equation 2.2 explicitly works to quantify the conflict between prior and data (Box, 1980).

2.3 Sceptical confidence interval

Definition 2.1 (Sceptical confidence interval) *The sceptical confidence interval reflects our initial scepticism about the treatment effect in the form of a range of sufficiently sceptical prior mean values μ. The sceptical confidence interval could be a one-sided version, a two-sided version, or under an unusual case with a set of disjoint intervals.*

This section is the core of the thesis, focusing on the development of the sceptical confidence interval. Firstly, the sceptical p-value (Held, 2020b) and the p-value function are introduced. Besides, the new standards of the sceptical p-value function and the sceptical confidence interval are proposed. Moreover, corresponding properties and limitations are then followed.

2.3.1 Sceptical p-value

The sceptical p-value (Held, 2020b) discussed in this section is the basis of the method of the sceptical confidence interval. In the context of null hypothesis testing, the ordinary p-value measures the statistical significance of a study, while the sceptical p-value quantifies the credibility of existing evidence from an original study by a replication study. When compared with the ordinary p-value of the replication study, the sceptical p-value is successfully extended to take the effect size estimates and sample sizes of both the original and replication studies into account. This new standard, with well-established theory, combines two approaches for analysis of credibility (Matthews, 2001b, 2018) and assessment of prior-data conflict (Box, 1980) simultaneously.

Analysis of credibility

Credibility is a core issue in the assessment of research findings. The desire of quantitative measurement of the credibility motivates the use of Bayesian inference (Matthews, 2001a).

The intrinsic shortcomings of statistical significance testing are well documented (Cristea and Ioannidis, 2018). As mentioned in Section 2.2.4, the inferential inadequacies of statistical significance testing, and p-values, in particular, has promoted the use of a Bayesian approach. Thus we can move beyond the p-value dichotomy and extract better insight from existing evidence.

For the sake of credibility, Bayes procedures can be implemented in a reversed way as introduced in Milner and Good (1951), and it has been widely implemented since then (Good, 1983; Greenland, 2006, 2011; Colquhoun, 2017, 2019). Instead of deriving a posterior based on

a pre-defined prior as mentioned in Section 2.2.4, it starts from a hypothetical posterior and answers the question of what prior is required to result in such a posterior. In the language of Bayesian inference,

$$\text{Prior insight} \leftarrow \text{Data likelihood} + \text{Posterior insight}, \tag{2.3}$$

which is the reversed version of Equation 2.1.

If the 95% credible interval of the calculated posterior distribution excludes 0, say no effect, the initial findings could be claimed as credible at the 95% level. That is how we come to the sceptical prior, answering the question of how suspicious we should be such that a positive result of a study will not be concluded as convincing.

Critical prior interval

Based on this reversed-Bayesian approach, Matthews (2001b) defines a critical prior interval for the assessment of credibility. The claimed findings can be regarded as credible at the specified level (i.e., 95%) if the initial evidence expressed by a conventional 95% confidence interval for the parameter of interest exists outside of the critical prior interval (CPI).

As shown in Matthews (2001b), the CPI with lower and upper bounds $(-S, +S)$ can be calculated from the data via

$$S = \frac{U - L}{4\sqrt{(UL)}}, \tag{2.4}$$

where S is the so-called scepticism limit, U and L are the lower and upper limit of the conventional two-sided confidence interval for the initial effect size. Confidence limits U and L depend on the specified confidence level (1-α/2)%. Thus, the scepticism limit itself depends on the specified significance level α.

Sufficiently sceptical prior

Based on Equation 2.4, variance of the sufficiently sceptical prior defined by Held (2019) is,

$$\tau^2 = \frac{S^2}{z^2_{\alpha/2}} = \frac{U - L}{16 z^2_{\alpha/2} UL} = \frac{z^2_{\alpha/2}\sigma^4}{\hat{\theta}^2 - z^2_{\alpha/2}\sigma^2} = \frac{z^2_{\alpha/2}\sigma^2}{t^2 - z^2_{\alpha/2}} = \frac{\sigma^2}{t^2/z^2_{\alpha/2} - 1},$$

where S is the sceptical limit as in Equation 2.4 and the variance σ^2 can be assumed to be known.

Thus the corresponding variance of the sufficiently sceptical prior of the original study is

$$\tau^2 = \frac{\sigma_o^2}{t_o^2/z^2_{\alpha/2} - 1}, \tag{2.5}$$

where $t_o = \hat{\theta}_o/\sigma_o$ is the test statistic and $t_o^2 \geq z^2_{\alpha/2}$ holds if the significance of the original study at level α holds.

Figure 2.2 based on fixed values of $z_{\alpha/2}$ and σ_o^2 shows the relationship between τ^2 and t_o^2 in Equation 2.5 intuitively. We are able to obtain the sufficiently sceptical prior τ^2 only if the original study result is significant at level α, say $t_o^2 \geq z^2_{\alpha/2}$. The sufficiently sceptical prior variance τ^2 can be either smaller or larger than σ_o^2, depending on the value of t_o^2. For a boundary value at $t_o^2 = z^2_{\alpha/2}$, say corresponding two-sided p-value $p_o = \alpha$, the sufficiently sceptical prior variance will be extremely large. If the squared test statistic t_o^2 is relatively large, i.e., $t_o^2 \geq 2z^2_{\alpha/2}$ (a substantially small p_o), then the sufficiently sceptical prior variance will be relatively small, i.e., $\tau^2 \leq \sigma_o^2$.

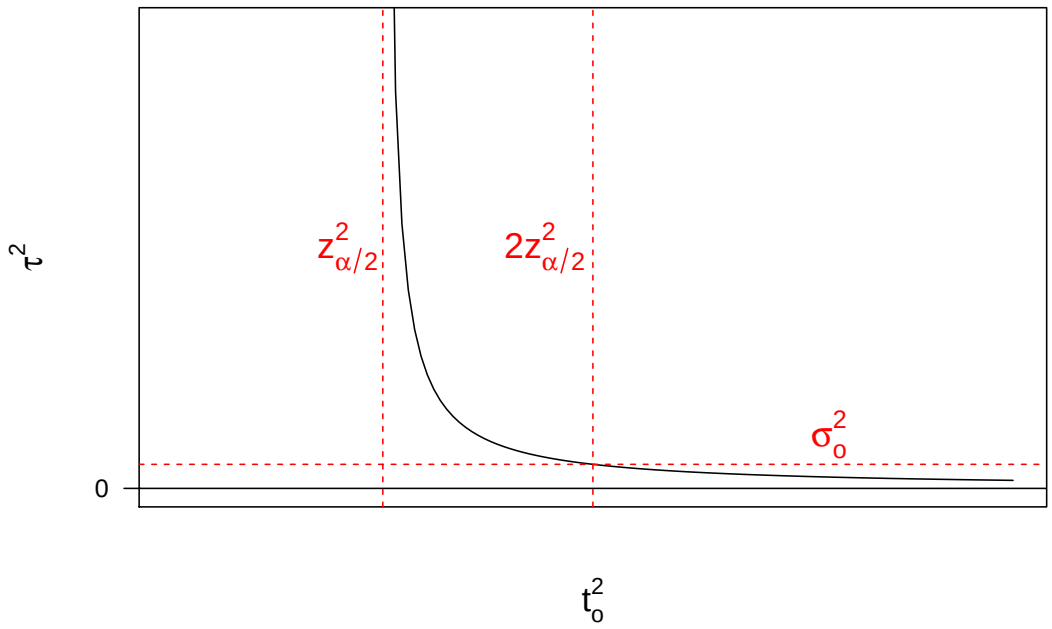

Figure 2.2: Variance of the sufficiently sceptical prior τ^2 as a function of the squared test statistic t_o^2, the two-sided significacne level is specified as α.

Figure 2.2 explains intuitively why the variance τ^2 is a nice interpretation of our scepticism about the previous findings expressed by t_o^2. Accordingly, the approach of analysis of credibility, based on a sufficiently sceptical prior, represents the argument from a sceptic who argues the claimed research findings of an original study would not be 'significant' anymore.

Assessment of prior-data conflict

The sufficiently sceptical prior derived from the original study works as an indicator of credibility. Since our goal is to assess replication success, we would like to check the conflict between the sufficiently sceptical prior and evidence from the replication study. The replication study result could be deemed credible if the prior-data conflict exists.

In Evans et al. (2006), a prior-data conflict exists whenever the data place little or no support to those values of θ where the prior provides its support. We could expect adequate supports of the prior and the posterior to be quite different if there is a significant prior-data conflict. Assessment of the existence of conflicts between prior and data is not the simple comparison between the replication effect size estimate $\hat{\theta}_r$ and the sufficiently sceptical prior we derive from the original study. Box's statistic (Box, 1980) helps to take the variance σ_r^2 of the replication study into account via the approach of prior-predictive distribution of $\hat{\theta}_r$ as discussed in Section 2.2.6. In a word, we are able to combine information of a sufficiently sceptical prior coming from the original study and the evidence from the replication study simultaneously.

The prior-predictive distribution (see Section 2.2.6) is assumed to be a normal distribution with mean μ and variance $\tau^2 + \sigma_r^2$. It is an established χ^2-distribution, with the test statistic,

$$t_{Box} = \frac{\hat{\theta}_r - \mu}{\sqrt{\tau^2 + \sigma_r^2}}.$$

Since the mean of a sceptical prior is $\mu = 0$ as mentioned in Section 2.2.5, the test statistic should be $t_{Box} = \hat{\theta}_r / \sqrt{\tau^2 + \sigma_r^2}$.

Given the significance level of α corresponding to the requirement for replication success, a small value of p_{Box} (i.e., $p_{Box} \leq \alpha$) or a large value of t_{Box} (i.e., $t_{Box}^2 \geq z_{\alpha/2}^2$) indicates the

existence of a conflict between the replication study and the sufficiently sceptical prior. Thus, the replication study is claimed as successful at level α. To summarize, by challenging a claim of a significant result of the original study, the replication success could be evaluated by the existing insight of the replication study.

In what follows, an example from Held (2020b) describes the procedure of the assessment of replication success, with the combination of analysis of credibility and prior-data conflict evaluation.

Table 2.2: *Example One* (Held, 2020b)

	Original study	Replication study
Effect estimate	$\hat{\theta}_o = 0.57$	$\hat{\theta}_r = 0.33$
Standard error	$\sigma_o = 0.1633$	$\sigma_r = 0.1652$
Ordinary p-values	$p_o = 0.0005$	$p_r = 0.0460$
Relative sample size	$c = n_r/n_o = \sigma_o^2/\sigma_r^2 = 0.98$	

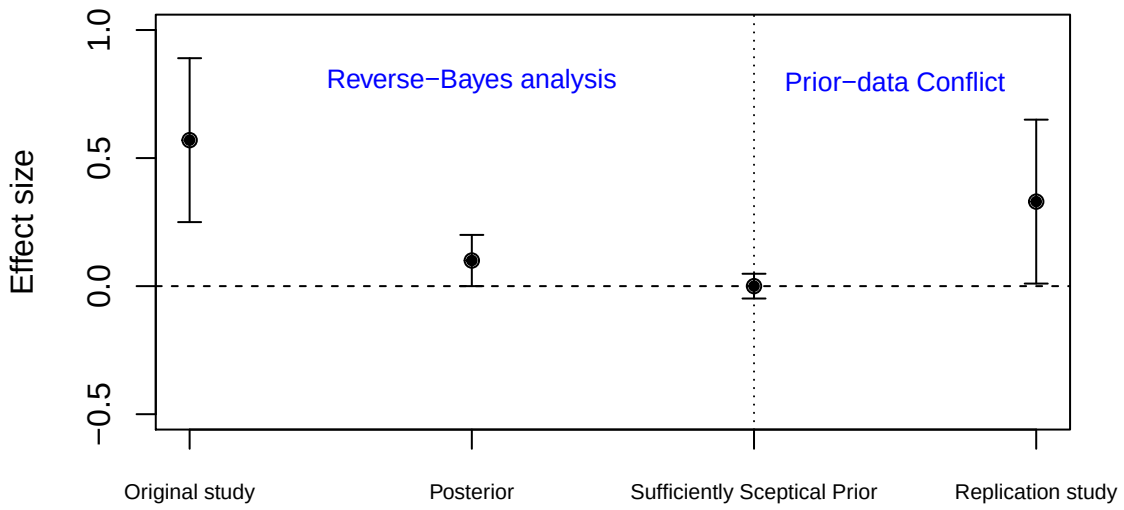

Figure 2.3: Process of the assessment of replication success for *Example One*

Figure 2.3 summarized the whole process of the replication success assessment for *Example One*. Firstly there is an effect estimate of the original study, and its 95% confidence interval is also illustrated. The reversed-Bayesian approach helps to turn the evidence from the original study into a sufficiently sceptical prior, based on a supposed posterior. Assessment of the replication study is then conducted by comparing the sufficiently sceptical prior and evidence from the replication study via measuring prior-data conflict.

For the application of this whole process in Figure 2.3, we have to specify the confidence level (1-$\alpha/2$)% at first, such that we can come to the variance of the sufficiently sceptical prior τ^2, and then the distribution of the test statistic t_{Box} as well as corresponding p_{Box}. Different significance levels α will result in different p_{Box} values. This gives rise to interpretation difficulties of the actual value of p_{Box} and motivates the development of the sceptical p-value p_S (Held, 2020b).

Sceptical p-value

The fact that the Box's tail probability p_{Box} depends on a specified significance level α is unwanted. Besides, the procedure of assessment described in Figure 2.3 is not achievable if the

original study result is not significant at level α, since the sufficiently sceptical prior would not exist, and the computation of p_{Box} is not achievable under a non-significant significant study (see Equation 2.5 and Figure 2.2). The sceptical p-value developed in Held (2020b) addressed these problems.

Suppose the test statistics of both original and replication study are $t_o = \hat{\theta}_o/\sigma_o$ and $t_r = \hat{\theta}_r/\sigma_r$, the corresponding relative sample size is $c = \sigma_o^2/\sigma_r^2$. The prior-predictive variance of $\hat{\theta}_r$ for the evaluation of prior-data conflict is:

$$\tau^2 + \sigma_r^2 = \sigma_r^2 \left(\frac{c}{t_o^2/z_{\alpha/2}^2 - 1} + 1 \right),$$

from which we can judge replication success by comparing the squared Box's test statistic and $z_{\alpha/2}^2$,

$$t_{Box}^2 = \hat{\theta}_r^2/(\tau^2 + \sigma_r^2) \geq z_{\alpha/2}^2.$$

Finally, the replication success is claimed by

$$(t_o^2/z_{\alpha/2}^2 - 1)(t_r^2/z_{\alpha/2}^2 - 1) \geq c. \tag{2.6}$$

The relationship $z_{\alpha/2}^2 < t_o^2$ always holds due to the significant result of the original study. Thus, $z_{\alpha/2}^2 < t_r^2$ should also hold to obtain a claim of replication success.

Therefore, we could come to the conclusion of this quantitative measure,

$$z_S^2 = \begin{cases} t_H^2/2 & c = 1 \text{ and} \\ \frac{1}{c-1}\left\{\sqrt{t_A^2[t_A^2 + (c-1)t_H^2] - t_A^2}\right\} & c \neq 1, \end{cases} \tag{2.7}$$

where $t_A^2 = (t_o^2 + t_r^2)/2$ is the arithmetic mean and $t_H^2 = 2/(1/t_o^2 + 1/t_r^2)$ is the harmonic mean of squared test statistics t_o^2 and t_r^2.

The sceptical p-value is then defined as $p_S = 1 - \Phi(z_S)$, and the requirement $z_S^2 \geq z_{\alpha/2}^2$ for a claim of replication success can be expressed as $p_S \leq \alpha$.

The sceptical p-value quantifies the conflict between the sufficiently sceptical prior and replication data, giving sight into evidence from both studies. Similar to the ordinary p-value, the sceptical p-value p_S can be evaluated under any specified significance level and has a nice interpretation. For instance, the calculated $p_S = 0.083$ in *Example One*, indicates a replication failure at level $\alpha = 0.05$ as $p_S > 0.05$. Therefore, this method developed within the framework of Bayesian inference has the nice property of frequentist inference, representing a Bayes-non-Bayes compromise (Good, 1992).

As mentioned in Held (2020a), the nominal level $\alpha = 0.05$, is too stringent to be the threshold for replication success. The calibrated level $\alpha = 0.13$ is more appropriate considering Type-I error control. More details about the choice of an appropriate level will be discussed in Section 2.5. We evaluate *Example One* as a replication success at the calibrated significance level $\alpha = 0.13$ as $p_S < 0.13$.

2.3.2 Sceptical p-value function and sceptical confidence interval

In this section, the concept of p-value function is firstly introduced, based on which, the sceptical p-value function and the sceptical confidence interval will be developed.

p-value function

Conventionally, the null hypothesis statistical test (NHST) is only based on ordinary p-value as introduced in Section 2.2.3. Due to inferential inadequacies of the p-value, a more widely

accepted method to show a magnitude of expected effect size is the confidence interval (Goodman, 1992; Killeen, 2005; Simonsohn, 2015). Thus, uncertainty of the inference is taken into account. The confidence interval around the point estimate has been advocated as an alternative, or a supplement, to NHST by many authors (Bolles and Messick, 1958; Gardner and Altman, 1986; Bailar and Mosteller, 1988; Cohen, 1992; Gonzalez, 1994; Cohen, 1994; Hunter, 1997).

A calculated confidence interval depends on an arbitrary choice of the confidence level. For instance, the most widely used one is the conventional 95% confidence interval. However, as argued by Blaker and Spjøtvoll (2000), the conclusion of a statistical analysis derived from the confidence interval is always too crude to be a summary of the information of the data.

To take the uncertainty about confidence levels into account, Cox (1956) introduced the confidence distribution, where confidence intervals are calculated based on various levels, instead of the single conventional fixed 95% level. This integration helps to summarize the crucial information required for a meaningful interpretation of results while avoiding the intrinsic pitfall of the significant result. For a few decades, this integration complements the conventional methods of presenting the results of medical studies, especially in cases where different confidence levels are required for medical decision making.

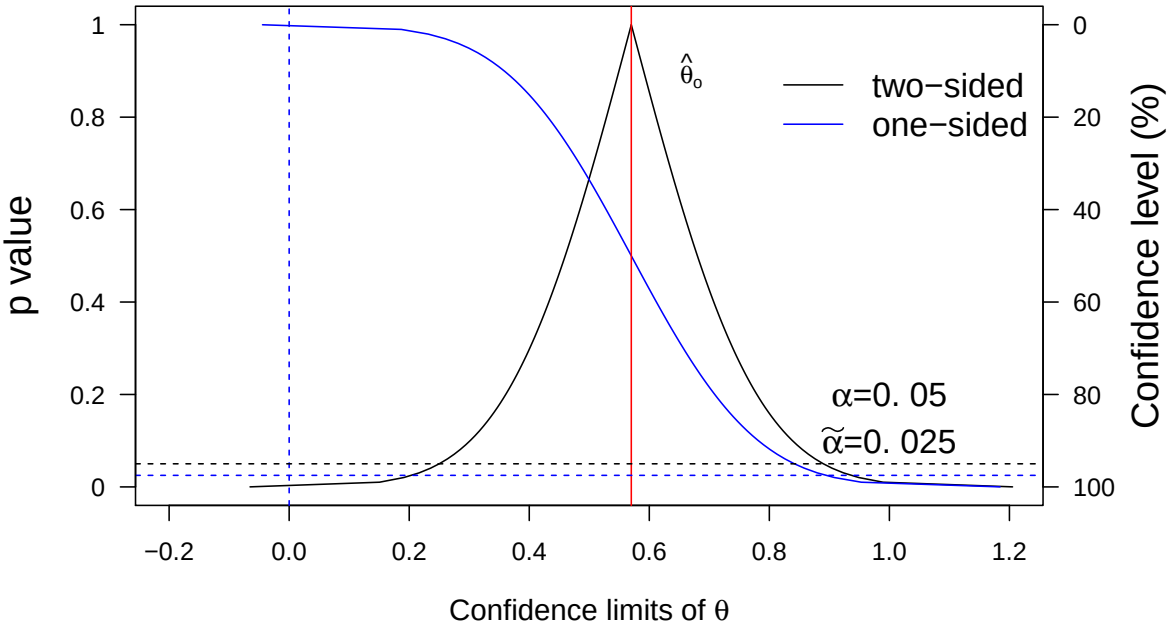

Figure 2.4: Two-sided and one-sided p-value functions for the original study of *Example One* (Table 2.2). The conventional two-sided and one-sided significance levels are plotted at the bottom.

Several terms have been adopted for this supplement method, which have quite different interpretations. Cox (1956) used the term 'confidence distribution', Birnbaum (1961) adopted 'confidence curves', while the name is specified as 'p-value function' in Stene and Miettinen (1987); Poole (1987b); Rothman (1998); Rothman et al. (2008). Other choices such as 'confidence curve', 'confidence interval function' and 'confidence distribution function' are also available (Poole, 1987a; Foster and Sullivan, 1987; Mau, 1988; Sullivan and Foster, 1990; Smith and Bates, 1992; Shakespeare et al., 2001). The choice of the term depends on the investigated question. In recent years, the 'p-value function' has attracted a surge of attention (Blaker, 2000; Bender et al., 2005; Berrar, 2017).

To obtain a graph of the p-value function, we can vary confidence levels from 0% to 100% and plot these values on the y-axis. The x-axis is constructed by the confidence limits of the effect size. In Figure 2.4, the cusp of the two-sided p-value function lies at $\hat{\theta}_o$ =0.57, which is

the point estimate of effect size of the original study in *Example One*. Intuitively, this may be regarded as a zero-percent confidence interval. The confidence interval we obtain at the 95% level is the conventional confidence interval for the original study, which is also demonstrated in Figure 2.3. The flexibility brought by this p-value function is that we can find a confidence interval at any confidence level by taking the points where the p-value function crosses through the corresponding horizontal line. The one-sided version of the p-value function, which is usually used for clinical relevance (Shakespeare et al., 2001) is also available in Figure 2.4. Corresponding one-sided confidence interval has the same lower limit at 97.5% as the conventional 95% two-sided confidence interval. At the confidence level 50%, the point estimate of the effect size $\hat{\theta}_o$ is observable.

The sceptical p-value discussed in Section 2.3.1 can be easily extended to a sceptical p-value function analogously, such that we give sights to various mean values μ of the sufficiently sceptical prior simultaneously. In the end, the sceptical confidence interval at any level will be easily obtained from the sceptical p-value function.

Sceptical p-value function and sceptical confidence interval (two-sided)

Definition 2.2 (Two-sided sceptical p-value function) *The (two-sided) sceptical p-value function for the two-sided sceptical p-values p_S corresponds to the null hypothesis $H_0 : \theta = \mu$. The x-axis represents confidence limits of mean values μ of the sufficiently sceptical prior. The y-axis is a range of two-sided sceptical p-values from 0 to 1, and corresponding confidence levels are from 100% to 0%.*

The sceptical p-value function helps to assess replication success at various levels and prevents investigators from misinterpreting their findings. In order to obtain different sceptical p-values corresponding to various μ values, test statistics of the original study and replication study in Equation 2.6 should be extended to $t_o = (\hat{\theta}_o - \mu)/\sigma_o$ and $t_r = (\hat{\theta}_r - \mu)/\sigma_r$, respectively.

The sceptical p-value function for *Example One* is displayed in Figure 2.5, where there are two cusps corresponding to two effect size estimates from the two investigated studies. While the (ordinary) two-sided p-value function in Figure 2.4 has only one cusp corresponding to the only effect size estimate $\hat{\theta}_o$ of the original study. We could obtain the largest $p_S = 1$ at $\mu = \hat{\theta}_o$ and $\mu = \hat{\theta}_r$, which correspond to substantially non-significant results of both studies, say $t_o = 0$ and $t_r = 0$. Replication failures then could be claimed due to substantially large p_S values. At $\mu = 0$, we obtain the sceptical p-value $p_S = 0.083$, as calculated in Section 2.3.1.

In *Example One*, effect size estimate of the original study is larger, $\hat{\theta}_o > \hat{\theta}_r$. The upper limits of sceptical confidence intervals are always obtained at $\mu > \hat{\theta}_o$, and the lower limits are always at $\mu < \hat{\theta}_r$. A list of μ values between $\hat{\theta}_o$ and $\hat{\theta}_r$, say $\hat{\theta}_r < \mu < \hat{\theta}_r$, indicates the so-called 'direction conflict' of these two studies. Unwanted direction conflict between the original and the replication study may result in a replication paradox, see Section 2.3.5.

Definition 2.3 (Direction conflict) *The direction conflict exists if the direction of the original study was failed to be replicated. Namely, the test statistics are not in the same direction. In the setting of a single replication, $t_o t_r < 0$ indicates the existence of a direction conflict. In the setting of multiple replications, the direction conflict can be defined as $t_{min} t_{max} < 0$, where $t_{min} = min\{t_1, ..., t_n\}$, $t_{max} = max\{t_1, ..., t_n\}$ and n is the number of studies.*

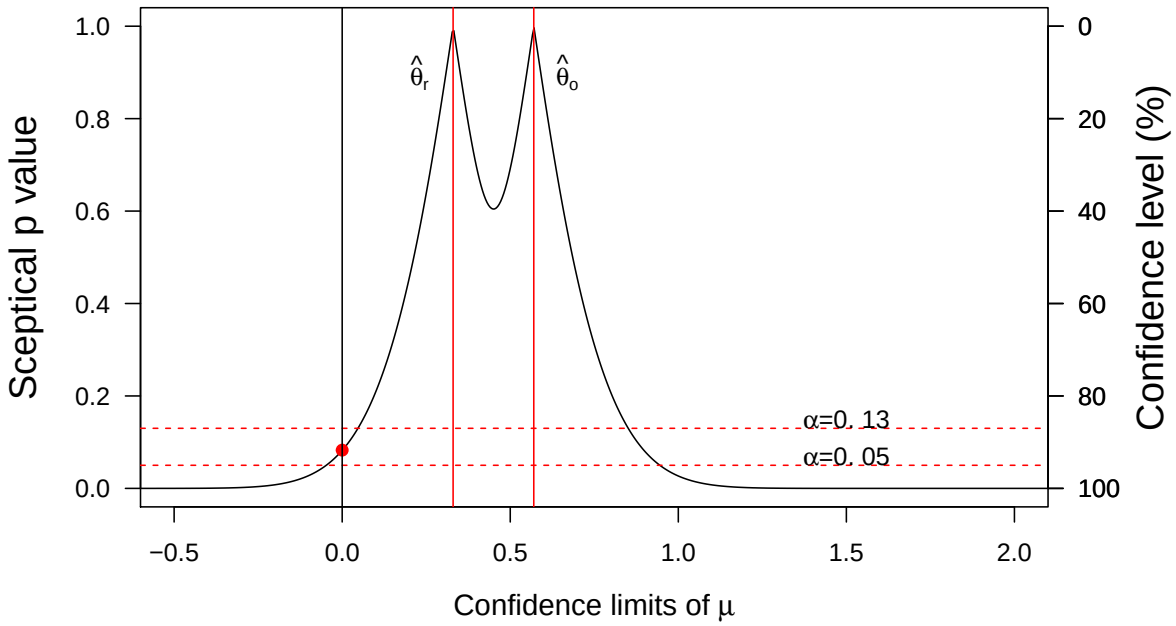

Figure 2.5: Sceptical p-value function of *Example One* (Table 2.2)

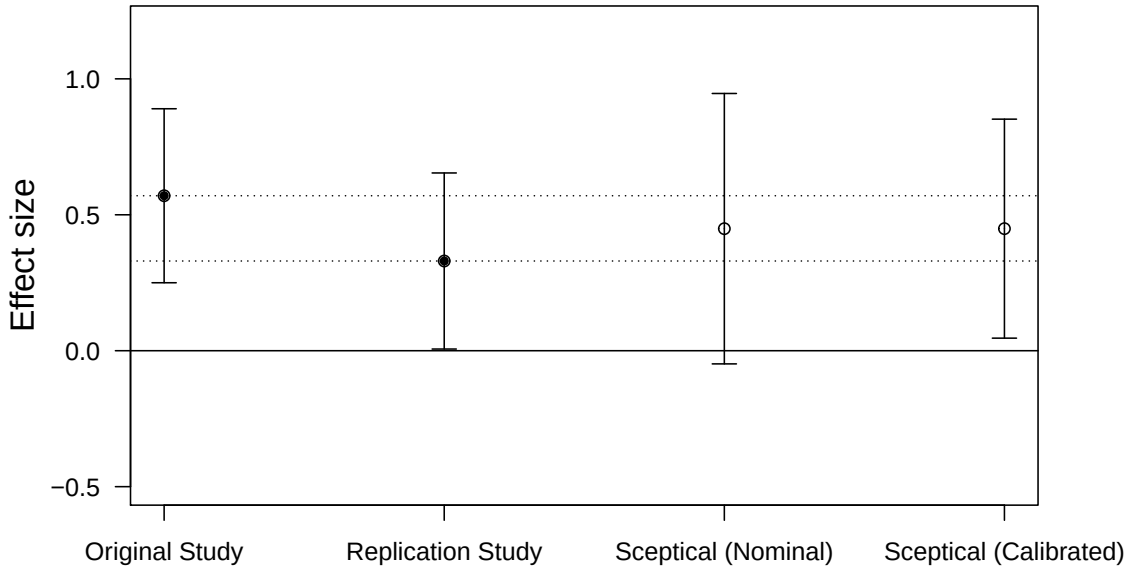

Figure 2.6: Conventional and sceptical confidence intervals of *Example One* (Table 2.2)

Definition 2.4 (Sceptical confidence interval in the two-sided version) *The sceptical confidence interval is the primary product of the sceptical p-value function. The 'two-sided' version indicates a two-sided test of the original study. A sceptical confidence interval at any confidence level* $(1-\alpha/2)\%$ *can be obtained by taking the confidence limits on the* x*-axis where the two-sided sceptical p-value function crosses through the corresponding horizontal line.*

In Figure 2.6, the original study and the replication study of *Example One* are both statistically significant at $\alpha = 0.05$. Sceptical confidence intervals at the nominal and calibrated level

are demonstrated. A claim of replication failure is declared at the nominal level as the nominal sceptical confidence interval covers $\mu = 0$. The calibrated sceptical confidence interval at the less stringent level $\alpha = 0.13$ is more appropriate and will result in a claim of replication success as $\mu = 0$ is excluded.

The sceptical p-value function is a simultaneous function of multiple parameters of both studies, from which unusual disjoint intervals may be obtainable. For example, if we set the significance level at $\alpha = 0.8$, we are supposed to observe two disjoint confidence intervals in a set, instead of one interval.

2.3.3 Sceptical confidence set

Definition 2.5 (Sceptical confidence set) *The bimodal (two-sided) sceptical p-value function can result in a sceptical confidence set, which is defined as a set of separable intervals at the same confidence level* $(1 - \alpha/2)\%$. *The sceptical confidence set is just an unusual case of the sceptical confidence interval (two-sided version) at the corresponding confidence level, with a separation between two disjoint intervals.*

The unusual sceptical confidence set could happen when the significance level α is large, which is not very common. It could also happen when there is a sizable difference between the original and replication study, see *Example Two* in Table 2.3.

Table 2.3: *Example Two*

	Original study	Replication study
Effect estimate	$\hat{\theta}_o = 0.90$	$\hat{\theta}_r = 0.10$
Standard error	$\sigma_o = 0.1796$	$\sigma_r = 0.1796$
Relative sample size	$c = n_r/n_o = \sigma_o^2/\sigma_r^2 = 1.00$	

The standardized difference, namely, the difference in means in units of standard deviation, can be utilized to quantify the difference between the original and replication study. Standardization allows for the comparison between variables measured on different scales (Austin, 2009; Cohen, 2013).

Definition 2.6 (Standardized between-study conflict) *The standardized between-study conflict measuring the discrepancy between the original study and the replication study is defined as,*

$$d = \frac{\hat{\theta}_o - \hat{\theta}_r}{\sqrt{\sigma_o^2 + \sigma_r^2}}, \tag{2.8}$$

based on which $d = (\hat{\theta}_o - \hat{\theta}_r)/\sqrt{2}\sigma$ *if the equal sample size assumption holds* ($\sigma = \sigma_o = \sigma_r$).

The standardized between-study conflict is usually larger than 0 as the original study effect size estimates are reliably larger than that of the replication studies (Open Science Collaboration, 2015). The absolute value of the standardized between-study conflict $|d|$ is more often used in this thesis with no consideration of the relationship between $\hat{\theta}_o$ and $\hat{\theta}_r$.

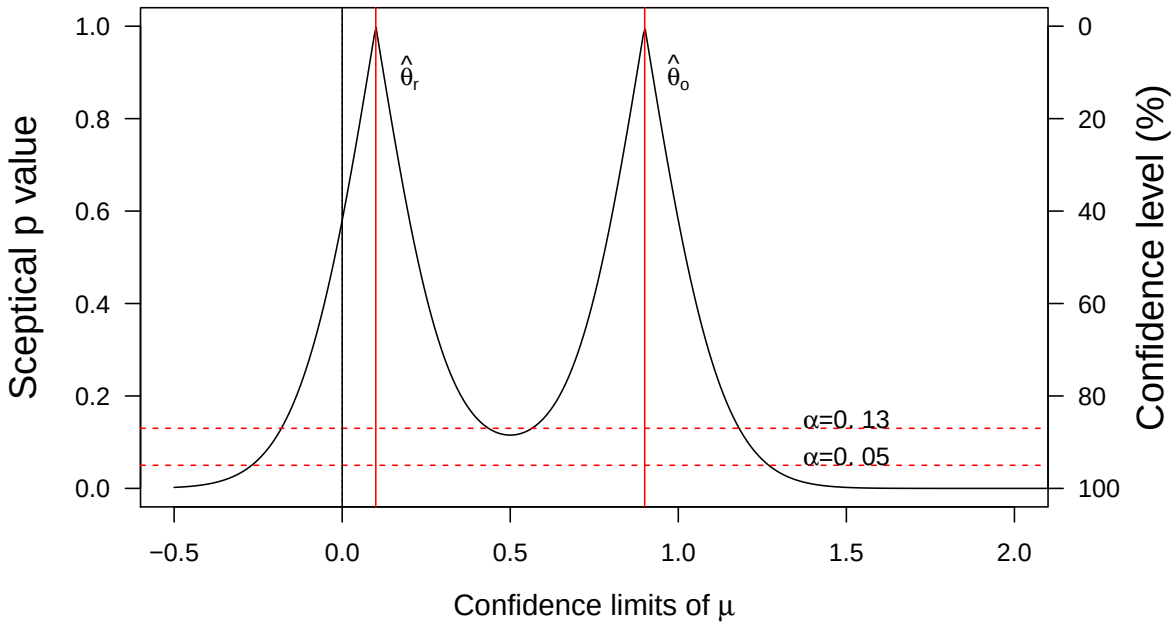

Figure 2.7: Sceptical p-value function of *Example Two* (Table 2.3)

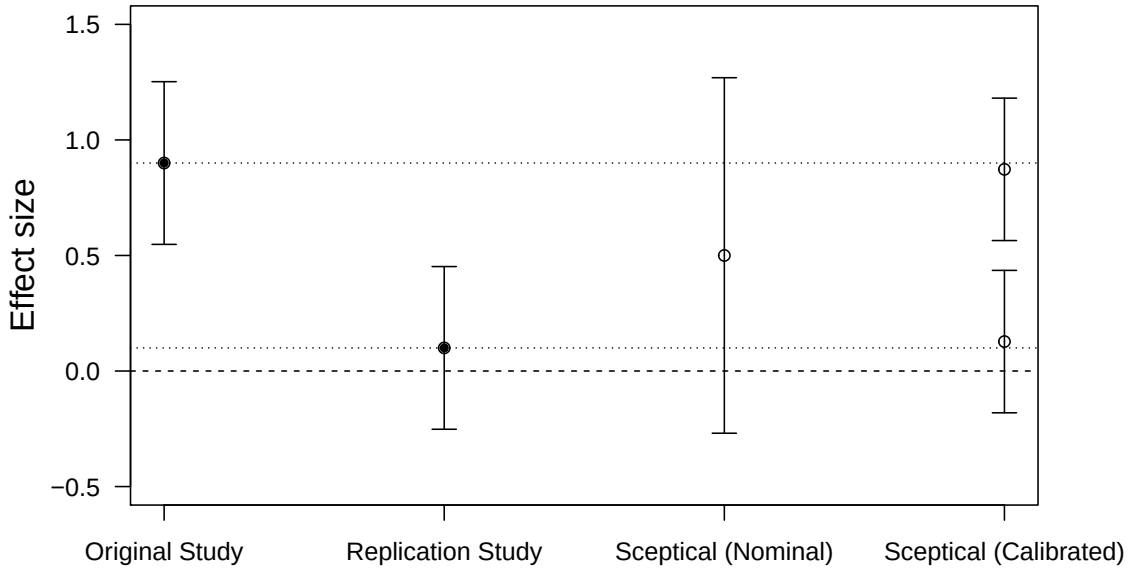

Figure 2.8: Conventional and sceptical confidence intervals of *Example Two* (Table 2.3)

Figure 2.7 and 2.8 illustrate the sceptical p-value function and sceptical confidence intervals for *Example Two* in Table 2.3. Due to the large standardized between-study conflict d=3.15, we obtain a sceptical confidence set at the calibrated level $\alpha = 0.13$, instead of a sceptical confidence interval. The direction conflict mentioned in Section 2.3.2 exists if $\hat{\theta}_r < \mu < \hat{\theta}_o$. There is a part of these μ values, for which the corresponding sceptical p-values ($p_S < \alpha$), will not be included in the sceptical confidence set, leading to a separation between two intervals. This gives rise to the question of what degree of standardized between-study conflict is required for the sceptical confidence set. The question will be discussed in Section 2.3.4 computationally and analytically.

On the one hand, this unusual phenomenon helps to explain the evidence from two studies

respectively, since one interval is around $\hat{\theta}_o$ and another one is around $\hat{\theta}_r$. The discrepancy between studies is obvious if we obtain a sceptical confidence set. On the other hand, such an unusual phenomenon may lead to the problem of the replication paradox, where an undesired claim of replication success would be made even these two studies have opposite directions. The problem of replication paradox and its solutions will be discussed in Section 2.3.5.

2.3.4 Properties of the sceptical confidence interval

To explore properties of the sceptical confidence interval, or the unusual sceptical confidence set due to large standardized between-study conflicts d, empirical and analytic results are illustrated in this section. Additionally, I compare the method of sceptical confidence interval with the method of sceptical p-value to show the advantages of the previous one.

Empirical results for the two-sided sceptical p-value methods

The two-sided sceptical p-value methods include the two-sided sceptical p-value and the sceptical confidence interval in two-sided versions, both of which are based the two-sided sceptical p-value. To obtain empirical results, we assume that effect size estimates $\hat{\theta}_o$ and standard errors σ_o for original studies are known and fixed at $\hat{\theta}_o = 0.6$ and $\sigma_o = 0.2$. With the assumption of an equal sample size ($c = 1$), standard errors of replication studies are also known as $\sigma_r = \sigma_o = 0.2$. The only variable is the effect size estimate $\hat{\theta}_r$ for replication studies. The calibrated sceptical confidence intervals and the sceptical p-values p_S, are plotted against the standardized between-study conflicts d in Figure 2.9. Note that the bottom plot for sceptical p-value p_S is different from the sceptical p-value function in Figure 2.5 and 2.7, since all of these sceptical p-values p_S are obtained at $\mu = 0$ in the bottom plot. Namely, the intrinsic difference between the method of sceptical confidence interval and the method of sceptical p-value is whether the test statistics $t_o = (\hat{\theta}_o - \mu)/\sigma_o$ and $t_r = (\hat{\theta}_r - \mu)/\sigma_r$ are obtained at $\mu = 0$. The sceptical p-value reflects point estimates for μ, while the sceptical confidence interval reflects interval estimates for μ.

There is a perfect symmetry in both plots in Figure 2.9. For calibrated sceptical confidence interval, either interval or set centers around mean value of effect size estimates $(\hat{\theta}_o + \hat{\theta}_r)/2$. For these sceptical confidence sets, each interval in the same set centers around effect size estimates of these two studies, $\hat{\theta}_o$ and $\hat{\theta}_r$, respectively.

The degree of standardized between-study conflict d required for an unusual sceptical confidence set can be obtained based on the derivation of Equation 2.6 with the assumption of $c = 1$, see Appendix A.1. The threshold in the form of the standardized between-study conflict d, depends only on the specified significance level α. To be specific, we are expected to obtain a sceptical confidence set, if $|d| > 2z_{\alpha/2}$.

Definition 2.7 (Threshold for the sceptical confidence set) *The threshold in the form of the standardized between-study conflict d for the sceptical confidence set should be*

$$d = \pm 2z_{\alpha/2}, \tag{2.9}$$

where α is the specified two-sided significance level, and the equal sample size assumption should hold (see Appendix A.1).

In Figure 2.9, the boundary of the direction conflict defined in Section 2.3.2 can be obtained at $\hat{\theta}_r = 0$. In terms of sceptical p-values in the bottom plot, we could observe the largest $p_S = 1$ at this boundary, which can also be obtained from Equation 2.6. It is reasonable to extend this conclusion to non-zero μ value cases, we are likely to obtain the largest $p_S = 1$ values if $t_r = (\hat{\theta}_r - \mu)/\sigma_r = 0$, equally $\mu = \hat{\theta}_r$. This confirms what we have observed in the sceptical p-value function in Figure 2.5 and 2.7.

The comparison of the top and the bottom plot in Figure 2.9 explains why the sceptical confidence interval is preferable to the sceptical p-value for the assessment of replication success. Information about the magnitude of effect size, the conflict between the original and the replication study, as well as the direction conflict defined in Section 2.3.2 are not obtainable from only a sceptical p-value. For example, we can easily judge whether there is a direction conflict by a calculated sceptical confidence interval, which is impossible with a sceptical p-value. Similarly, we can tell whether a claim of replication success is due to the problem of replication paradox mentioned in Section 2.3.3 based on the sceptical confidence interval, but it is unfeasible based on the sceptical p-value. The replication paradox will be discussed in Section 2.3.5.

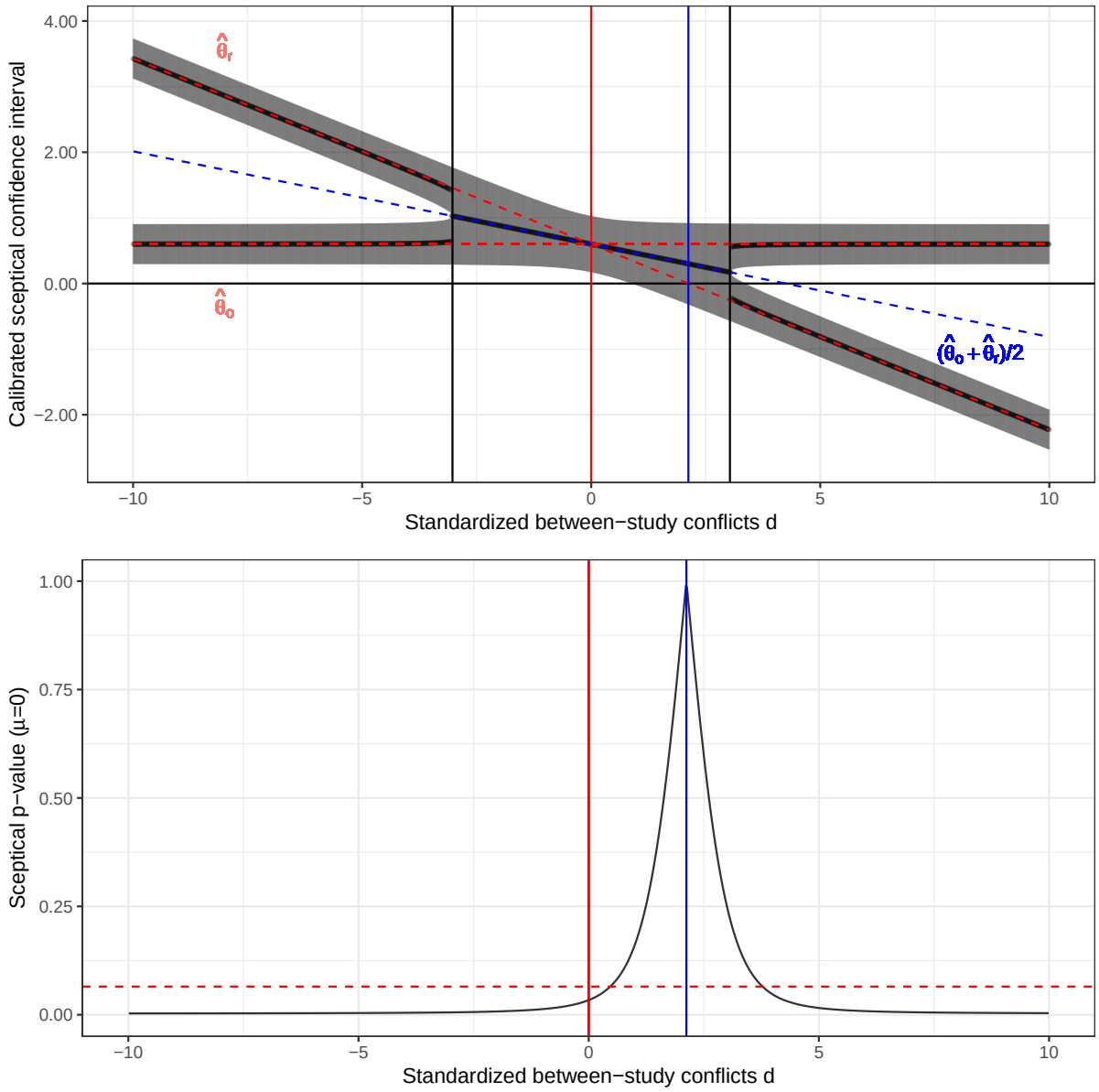

Figure 2.9: Empirical results of calibrated sceptical confidence intervals (CI_S) and sceptical p-values (p_S) with respect to different standardized between-study conflicts. The calibrated level $\alpha = 0.13$ is plotted for sceptical p-values.

Analytical results for the two-sided sceptical p-value methods

Under the equal sample size assumption ($c = 1$) and non-zero μ values, Equation 2.6 for replication success assessment can be expressed as

$$(t_o^2/z_{\alpha/2}^2 - 1)(t_r^2/z_{\alpha/2}^2 - 1) \geq 1, \tag{2.10}$$

where $t_o = (\hat{\theta}_o - \mu)/\sigma_o$, $t_r = (\hat{\theta}_o - \mu)/\sigma_o$ and $\sigma_o = \sigma_r$.

This criterion based on test statistics t_o and t_r is illustrated in Figure 2.10. The red regions indicate direction conflicts as t_o and t_r are in different directions. A claim of replication success can be declared if the values of t_o and t_r are on or beyond these four black curves, corresponding to Equation 2.10.

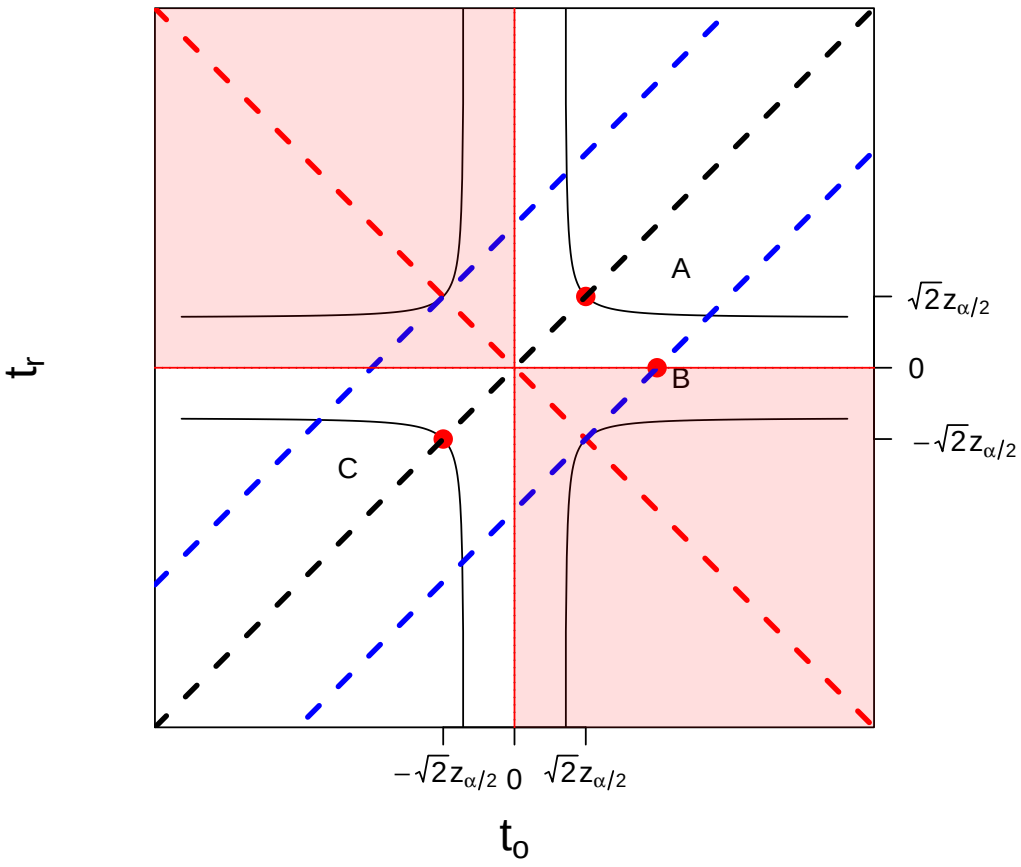

Figure 2.10: Replication success assessment under equal sample size assumption

Assume we have three pairs of studies for the assessment of replication success, study pair A, study pair B, and study pair C, each with a pair of original study and replication study (Table 2.4). Study pair A and study pair C have the same results of both original and replication studies, but in the opposite directions. They are just on the curves and have a narrow squeak regarding the assessment of replication success. While the study pair B, with a significant original study $t_o = 2\sqrt{2}z_{\alpha/2}$ and a non-significant replication study $t_r = 0$, can not be claimed as a replication success.

Table 2.4: *Three pairs of studies*

	t_o	t_r
study pair A	$\sqrt{2}z_{\alpha/2}$	$\sqrt{2}z_{\alpha/2}$
study pair B	$2\sqrt{2}z_{\alpha/2}$	0
study pair C	$-\sqrt{2}z_{\alpha/2}$	$-\sqrt{2}z_{\alpha/2}$

To obtain sceptical confidence intervals in Figure 2.10, we want the solutions for μ in Equation 2.10, namely, the intersection of these four curves and the line $t_o - t_r = (\hat{\theta}_o - \hat{\theta}_r)/\sigma$, where $\sigma = \sigma_o = \sigma_r$. The location of this line, which describes the relationship between the test statistics of the original and replication studies, depends on $\hat{\theta}_o$, $\hat{\theta}_r$ and σ simultaneously. For instance, the black dashed line crossed point A and point C could be an example of such a line. The intersections, point A and point C, correspond to two μ values and these μ values are limits of the sceptical confidence interval under the specified $\hat{\theta}_o$, $\hat{\theta}_r$ and σ. Intuitively, this line corresponds to $t_o = t_r$ is the narrowest sceptical confidence interval. The narrowest width of the sceptical confidence interval is $2\sqrt{2}z_{\alpha/2} \cdot \sigma$, which can be derived with the distance from point A to point C, see Appendix A.2. This kind of perfect replication is not realistic and achievable as there should be some differences between the original and replication study, thus the calculated sceptical confidence interval is always wider than $2\sqrt{2}z_{\alpha/2} \cdot \sigma$.

The blue dashed line crossed point B corresponds to $t_o - t_r = (\hat{\theta}_o - \hat{\theta}_r)/\sigma = 2\sqrt{2}z_{\alpha/2}$. If $t_o - t_r \geq 2\sqrt{2}z_{\alpha/2}$, namely, the line of $t_o - t_r = (\hat{\theta}_o - \hat{\theta}_r)/\sigma$ is above or under these two blue dashed lines, there will be four intersections, corresponding to four confidence limits of a secptical confidence set. This line can be expressed in the form of the standardized between-study conflict d defined in Equation 2.8. Namely, $|t_o - t_r| = \sqrt{2}(\hat{\theta}_o - \hat{\theta}_r)/\sqrt{\sigma_o^2 + \sigma_r^2} = 2\sqrt{2}z_{\alpha/2}$. Accordingly, the threshold for unusual sceptical confidence set $|t_o - t_r| = \sqrt{2}z_{\alpha/2}$ can be expressed as $|d| = 2z_{\alpha/2}$, which confirms the results in empirical results, see Figure 2.9.

The red dashed line $t_o + t_r = 0$ can also be expressed as $\mu = (\hat{\theta}_o + \hat{\theta}_r)/2$, which is the center line for sceptical confidence intervals and sceptical confidence sets as shown in empirical results in Figure 2.9.

2.3.5 Replication paradox

Definition 2.8 (Replication paradox) *The replication paradox describes a situation where the result of the replication study goes in the opposite direction of the original study, but the method still declares a replication success.*

The problem of replication paradox is not unique for the method of sceptical confidence interval but also emerged in the replication Bayes factor proposed in Ly et al. (2019).

Example Three in Table 2.5 illustrates an extreme example for such a paradox, where the effect size estimates of the original and replication study are the same but in opposite directions.

Table 2.5: *Example Three*

	Original study	Replication study
Effect estimate	$\hat{\theta}_o = 0.8$	$\hat{\theta}_r = -0.8$
Standard error	$\sigma_o = 0.2685$	$\sigma_r = 0.2685$
Relative sample size	$c = n_r/n_o = \sigma_o^2/\sigma_r^2 = 1.00$	

Figure 2.11 is the sceptical p-value function for *Example Three* in Table 2.5, the sceptical p-value at $\mu = 0$ is quite small, thus resulting in a claim of replication success at both nominal and

calibrated levels ($p_S < 0.05$). The same results can be observed in Figure 2.12, where sceptical confidence sets at both nominal and calibrated levels fail to cover $\mu = 0$. The claims at both levels are unrealistic as the replication success should not be declared with the existence of a direction conflict between two studies.

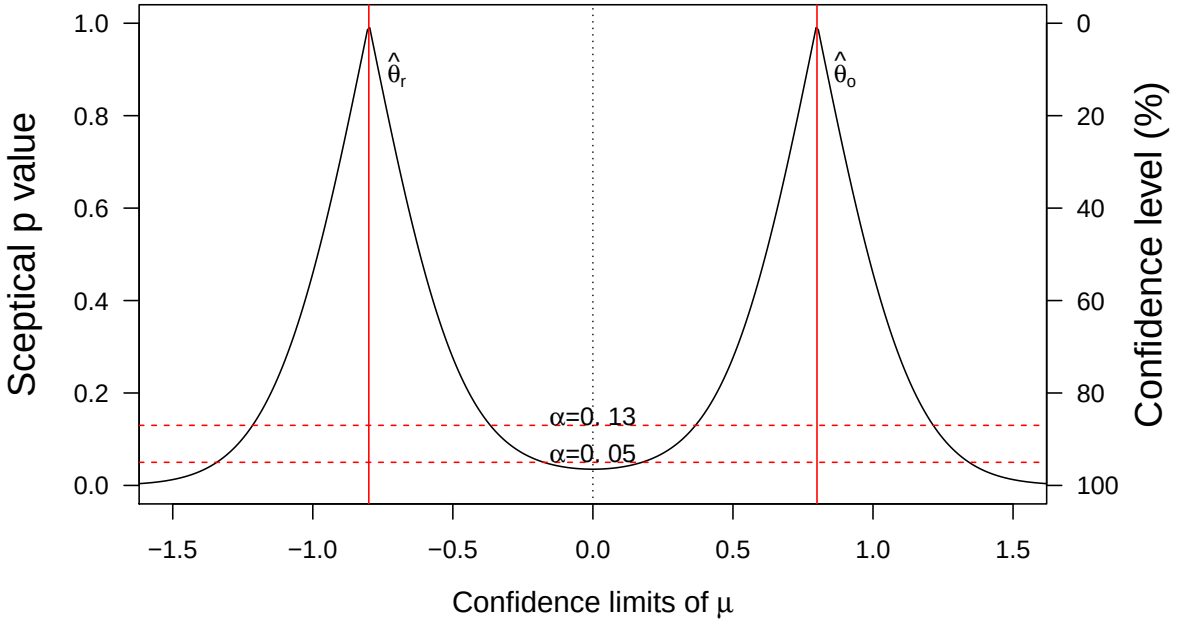

Figure 2.11: Sceptical p-value function of *Example Three* (Table 2.5)

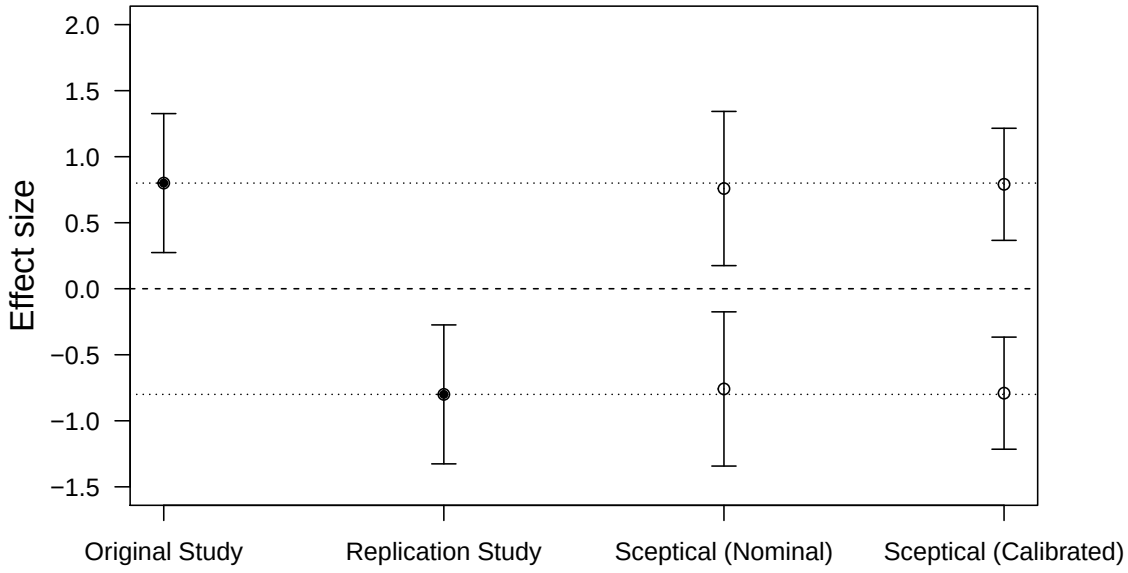

Figure 2.12: Conventional and sceptical confidence intervals of *Example Three* (Table 2.5)

There are two necessary but not sufficient requirements for the replication paradox. Firstly, the unusual sceptical confidence set due to a sizable standardized between-study conflict, say $|d| > 2z_{\alpha/2}$ is required (see Equation 2.9). Secondly, the direction conflict exists, namely $\mu = 0$ should be between $\hat{\theta}_o$ and $\hat{\theta}_r$.

In Figure 2.13, I reveal both requirements in the form of the standardized between-study

conflicts d. We assume that the original study has a positive result with the test statistic $t_o > 0$, thus the test statistic for the replication study $t_r > 0$ indicates no existence of the replication paradox, and we just focus on the right part where there may exist a direction conflict. The threshold for the sceptical confidence set $d = \pm 2z_{\alpha/2}$ defined in Equation 2.8, are presented by two vertical lines in red. We will not obtain sceptical confidence interval, but sceptical confidence set outside of the red vertical lines where $|d| > 2z_{\alpha/2}$. There is no direction conflict on the left side of the blue vertical lines, which represents $t_r > 0$.

The case on the left side is the same case as the empirical results in Figure 2.9. To be specific, there is always a direction conflict if we have observed sceptical confidence sets instead of sceptical confidence intervals. The boundary for the direction conflict (the blue vertical line) is smaller than the threshold for the sceptical confidence set (the red vertical line). Another case is on the right side, unusual sceptical confidence sets could happen when there is no direction conflict (the blue region). As shown in Appendix A.3, the question of which of these two cases we will observe depends on how significant the original study is. Considering a case with a positively significant original study $t_o < 2\sqrt{2}z_{\alpha/2}$, an unusual sceptical confidence set could only happen when there is a direction conflict, just as what we have observed in Figure 2.9. The sceptical confidence set could happen even though there is no direction conflict if $t_o > 2\sqrt{2}z_{\alpha/2}$.

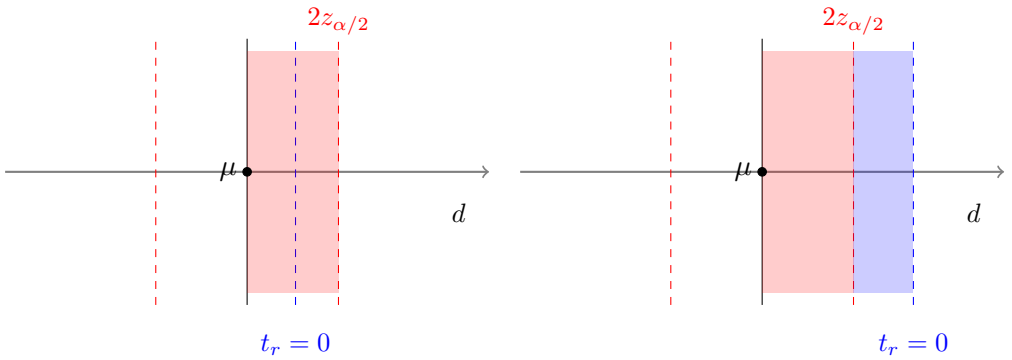

Figure 2.13: The relationship between two requirements for the replication paradox

In Section 2.4, two alternatives are introduced to avoid the replication paradox. One is originated from the one-sided sceptical p-value $\widetilde{p}_S$, equally the sceptical confidence interval in the one-sided version $\widetilde{CI}_S$. Another one is the harmonic mean χ^2 confidence interval based on the harmonic mean χ^2 test (Held, 2020a).

2.3.6 Influence of the relative sample size c

The discussion up until now is based on the assumption of an equal sample size. Generally speaking, replication studies do not have exactly the same sample size as the original studies in reality, they are expected to be larger than the original studies to ensure the statistical power (Maxwell et al., 2015). Cases with $c < 1$ are also possible when the resources for the replication study was limited or when the original sample size n_o was especially large.

I illustrate empirical results and analytical results in this section to show the influence of relative sample size c on the calculated sceptical confidence intervals.

Empirical results for the sceptical confidence intervals

To obtain the empirical results of calibrated sceptical confidence intervals CI_S, the original study ($\hat{\theta}_o = 0.6$, $\sigma_o = 0.1$) and the standardized between-study conflict ($d = 3$) are fixed. The calculated sceptical confidence intervals at $\alpha = 0.13$ are plotted against various relative sample sizes c in Figure 2.14.

In Figure 2.14 for the empirical results of sceptical confidence intervals or sets, perfect symmetry no longer exists. The dashed line in blue represented by $(\hat{\theta}_o+\hat{\theta}_r)/2$, is no longer the center line. Relative sample size c plays a more important role when it is relatively small. Width of different intervals in a sceptical confidence set now differs due to unequal sample sizes. On the left hand side where the relative sample sizes are relatively small, we obtain a wider interval around $\hat{\theta}_r$, since smaller replication study will result in less precise effect estimate. This reflects that the sceptical confidence set can transfer some useful information that the sceptical confidence interval cannot.

Smaller relative sample size c, say relatively smaller replication study than the original study, is more likely to result in an unusual sceptical confidence set. Since replication studies are expected to be larger than original studies (Open Science Collaboration, 2015), an unusual sceptical confidence set is less likely to happen (see replication projects results in Chapter 3). Additionally, we can also conclude that, with a fixed original study, the size of the replication study makes sense for the claim of replication success. Larger replication study helps to support the significant result of the original study, since larger replication study can ensure adequate statistical power. The problem of inadequate power of replication study can be settled by equivalence test and multiple replications as mentioned in Maxwell et al. (2015).

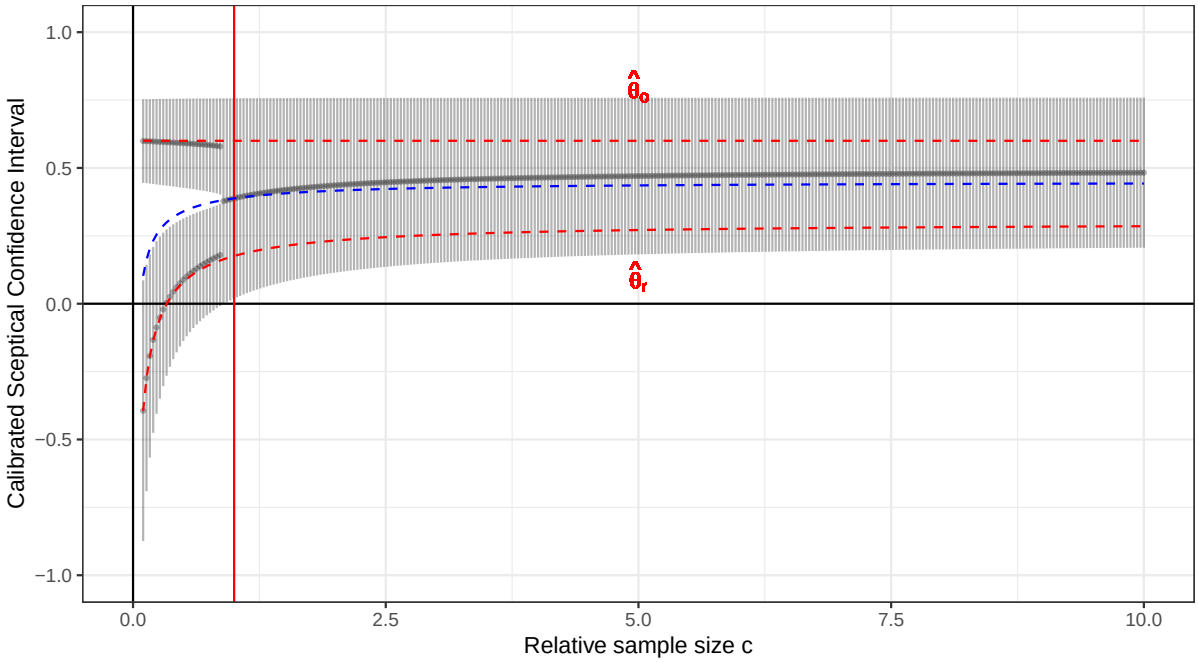

Figure 2.14: Empirical results of calibrated sceptical confidence intervals and calibrated sceptical confidence sets against the relative sample size c

Assume a case under a larger between-study conflict d, it is reasonable that an unusual sceptical confidence set could happen when the relative sample size is relatively larger compared with the empirical results in Figure 2.14. The influence of the relative sample size c on the sceptical confidence interval has practical value for the design of a replication study. It drives us to consider the required sample size for a replication to obtain reliable results for the assessment of replication success.

Analytical results for the sceptical confidence intervals

In Figure 2.15, the curves in black correspond to the assessment criteria under the equal sample size assumption in Equation 2.10. Profits brought by the equal sample size assumption no longer exist. Sceptical confidence intervals can be obtained in the same way as in Figure 2.10, where

we want the solutions for μ in

$$(t_o^2/z_{\alpha/2}^2 - 1)(t_r^2/z_{\alpha/2}^2 - 1) \geq c, \tag{2.11}$$

where $t_o = (\hat{\theta}_o - \mu)/\sigma_o$, $t_r = (\hat{\theta}_o - \mu)/\sigma_o$ and $c = \sigma_o^2/\sigma_r^2$.

The curves for the assessment criterion depend on the relative sample size c, which describes the relationship between different values of σ_o and σ_r. The intersections corresponding to solutions of μ are now determined by these curves and a line $\sqrt{c} \cdot t_o - t_r = (\hat{\theta}_o - \hat{\theta}_r) \cdot \sigma_r$, with a slope $\sqrt{c}$ instead of 1.

Suppose we observe a perfect replication regarding effect size estimates, where $\hat{\theta}_o = \hat{\theta}_r$, the line $\sqrt{c} \cdot t_o - t_r = 0$ will cross the point at $t_o = t_r = 0$. The calculated sceptical confidence interval, contracted by two intersections is not easy to obtain. The reason is that not only the line but also these four curves vary with different c values. The additional randomness brought by an additional unknown parameter c makes the analytical result not available.

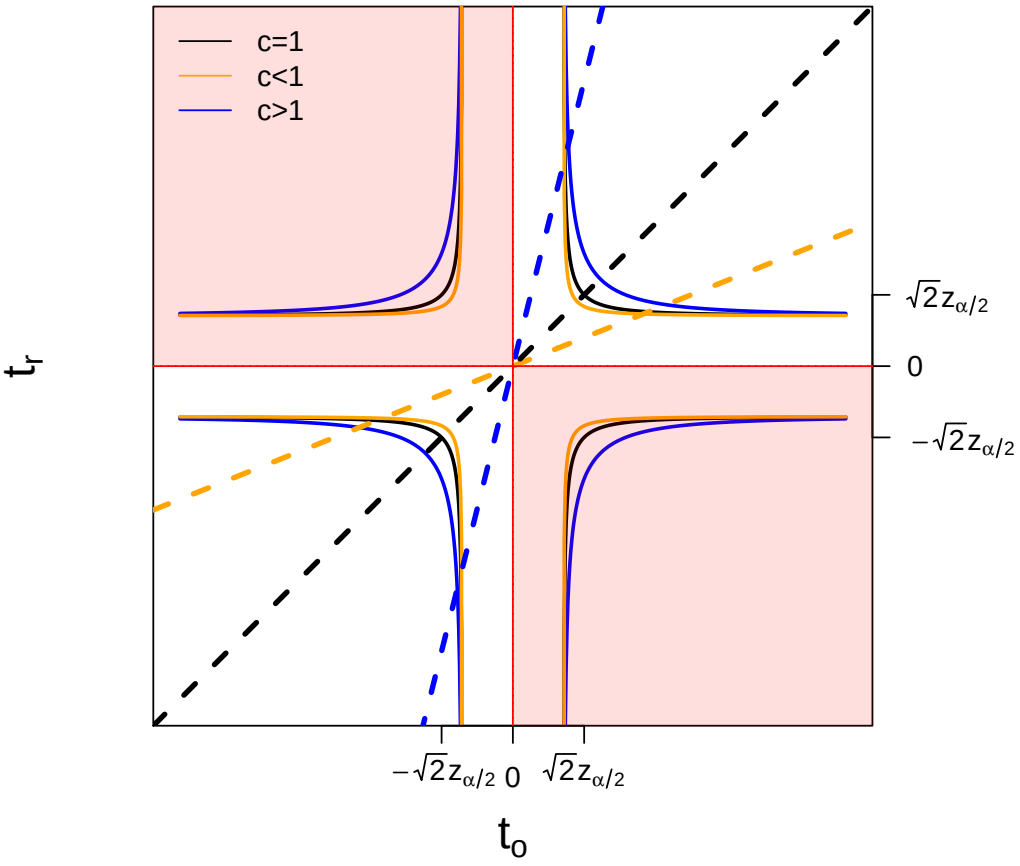

Figure 2.15: Replication success assessment under general sample size assumption

To summarize, we are expected to obtain an unusual sceptical confidence set not only when the conflict between the original and replication study is huge, but also when the replication study is not large enough.

2.4 Alternative assessment methods

To prevent the replication paradox mentioned in Section 2.3.5, I illustrate two alternatives in this section. The first one is based on the one-sided sceptical p-value, while the second on the is grounded on the harmonic mean χ^2 test (Held, 2020a). The same idea is that we adjust p-values

when there is a direction conflict, such that the undesired small p-values will not be observed. Additionally, I will present the widely-used evidence synthesis method, meta-analysis.

2.4.1 One-sided sceptical p-value method

In Section 2.3.4, original studies are fixed for the illustration of empirical results of calibrated sceptical confidence intervals and sceptical p-values. The 'no-conflict' direction of replication studies has already been determined. Thus the assumption of the direction of the original study allows us to avoid the replication paradox.

Sceptical p-value function and sceptical confidence interval (one-sided)

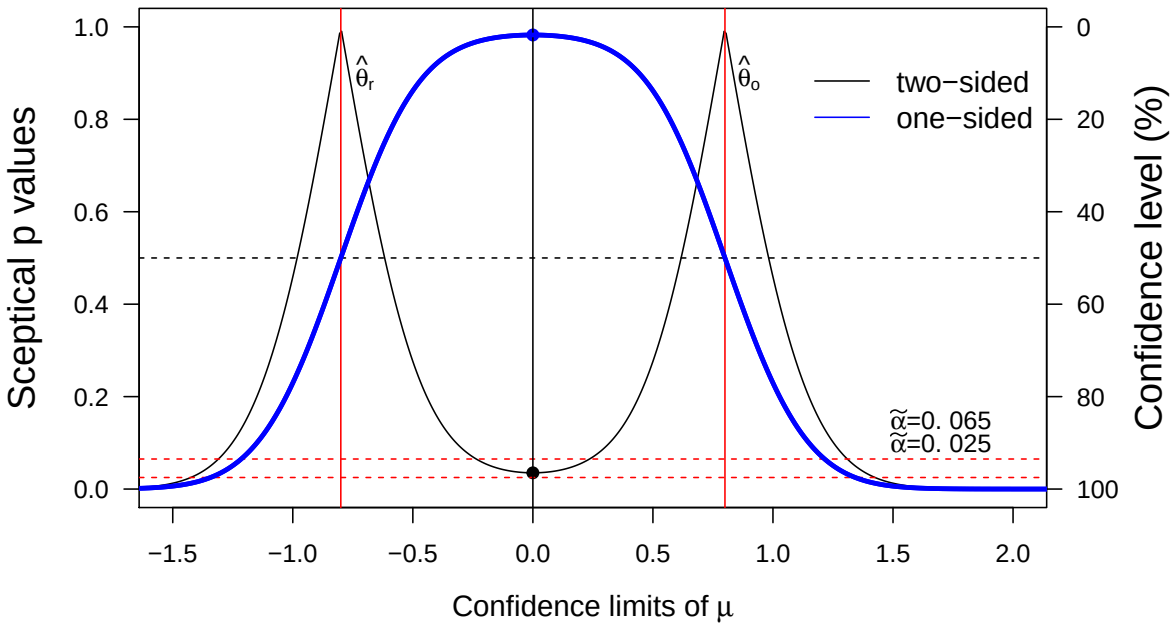

Figure 2.16: Sceptical p-value functions of *Example Three* (Table 2.5)

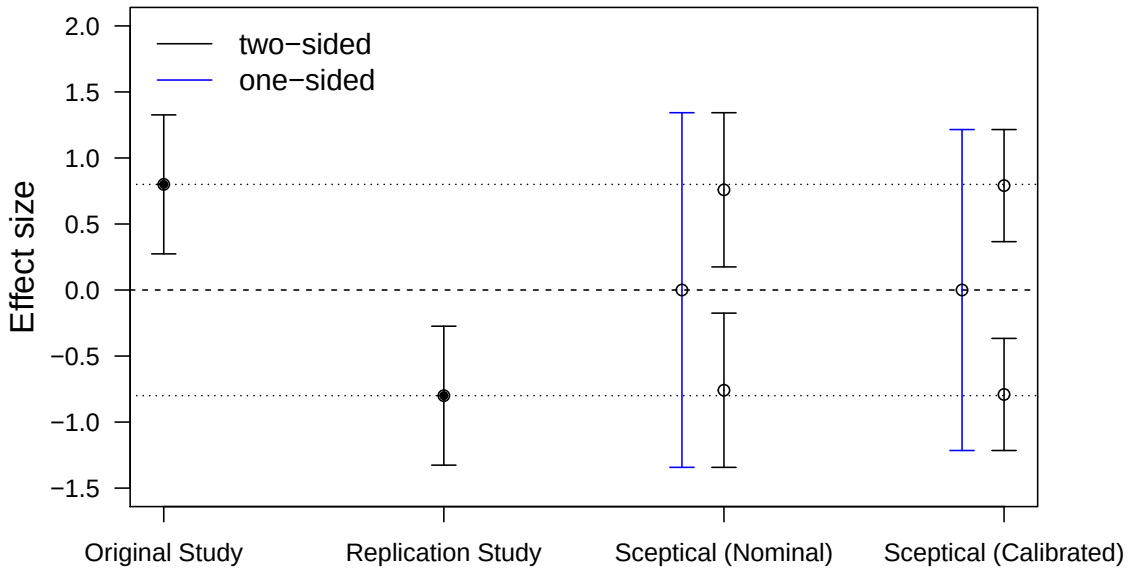

Figure 2.17: Conventional and sceptical confidence intervals of *Example Three* (Table 2.5)

Definition 2.9 (One-sided sceptical p-value function) *The one-sided sceptical p-value function for the one-sided sceptical p-values $\widetilde{p}_S$ corresponds to the one-sided test of the original study, i.e., the alternative hypothesis $H_1 : \theta > \mu$ to the null hypothesis $H_0 : \theta = \mu$.*

For generality, we always choose the one-sided significance level $\widetilde{\alpha} = \alpha/2$, such that the lower limit of the one-sided confidence interval, $\hat{\theta}_o - z_{\widetilde{\alpha}}\sigma_o$ at the confidence level $(1-\widetilde{\alpha})\%$, is the same as the counterpart of the two-sided confidence interval $\hat{\theta}_o - z_{\alpha/2}\sigma_o$ at the confidence level $(1-\alpha/2)\%$. Therefore, the variance τ^2 of the sufficiently sceptical prior in Equation 2.5 will depend on $z_{\widetilde{\alpha}}$, instead of $z_{\alpha/2}$.

Conventionally, one-sided p-values can be obtained by halving two-sided p-values. Similarly, values of $\widetilde{p}_S$ are obtainable by lowering p_S values to the half. Besides, we adapt $\widetilde{p}_S = p_S/2$ to $\widetilde{p}_S = 1 - p_S/2$, when there are direction conflicts defined in Section 2.3.2. The sceptical p-value function in the one- and two-sided versions for *Example Three* in Table 2.5 are illustrated in Figure 2.16. Unexpected small values of $\widetilde{p}_S$ when $\hat{\theta}_r < \mu < \hat{\theta}_o$ will no longer happen due to the adjustment.

When compared to the ordinary one-sided p-value function in Figure 2.4, one-sided sceptical p-value function is a two-tailed function with one cusp. Thus the corresponding sceptical confidence interval at a specified confidence level has both lower and upper limits, say a two-sided sceptical confidence interval.

Definition 2.10 (Sceptical confidence interval in the one-sided version) *The sceptical confidence interval in the one-sided version is a two-sided interval, where 'one-sided' corresponds to the one-sided test of the original study. A sceptical confidence interval in the one-sided version $\widetilde{CI}_S$ at the one-sided significance level $\widetilde{\alpha} = \alpha/2$ just ignores the possible separation in a sceptical confidence set at the two-sided significance level α. Same results could be obtained by sceptical confidence interval in the one-sided version $\widetilde{CI}_S$ at $\widetilde{\alpha}$ and sceptical confidence interval in the two-sided version CI_S at α if CI_S is not an unusual sceptical confidence set.*

In Figure 2.17, the sceptical confidence intervals $\widetilde{CI}_S$ at both nominal and calibrated levels ($\widetilde{\alpha}$=0.025 and $\widetilde{\alpha}$= 0.065) just ignore the separations of sceptical confidence set in two-sided version at $\alpha = 0.05$ and $\alpha = 0.13$. We take the upper limit of the larger interval and lower limit of the smaller interval to construct new limits.

Empirical results for the one-sided sceptical p-value methods

The one-sided sceptical p-value methods are based on the one-sided sceptical p-value. Empirical results for calibrated sceptical confidence intervals $\widetilde{CI}_S$ and $\widetilde{p}_S$ values are illustrated in Figure 2.18, where the fixed values $\hat{\theta}_o$, σ_o and σ_r are the same when compared with Figure 2.9. We will acquire sceptical confidence intervals, no matter how large the standardized between-study conflicts d are.

One-sided sceptical p-values $\widetilde{p}_S$ are shrunk from values in the two-sided version p_S, thus the value we obtained at $\hat{\theta}_r = 0$, equally $t_r = 0$ as $\mu = 0$, is halved to 0.5. However, for those with direction conflicts (right hand side of the blue line), one-sided sceptical p-values $\widetilde{p}_S$ are turned over to above 0.5. Therefore, this curve for one-sided sceptical p-values increase monotonically. Larger standardized between-study conflicts d will result in larger one-sided sceptical p-values. However, this is not a rule as the assumption of an equal sample size may not hold, see Section 3.1.1.

The one-sided sceptical p-value can not only avoid the replication paradox, but also transfer informative that the two-sided sceptical p-value cannot. To be specific, we know that there is a direction conflict once a one-sided sceptical p-value $\widetilde{p}_S > 0.5$ is obtained.

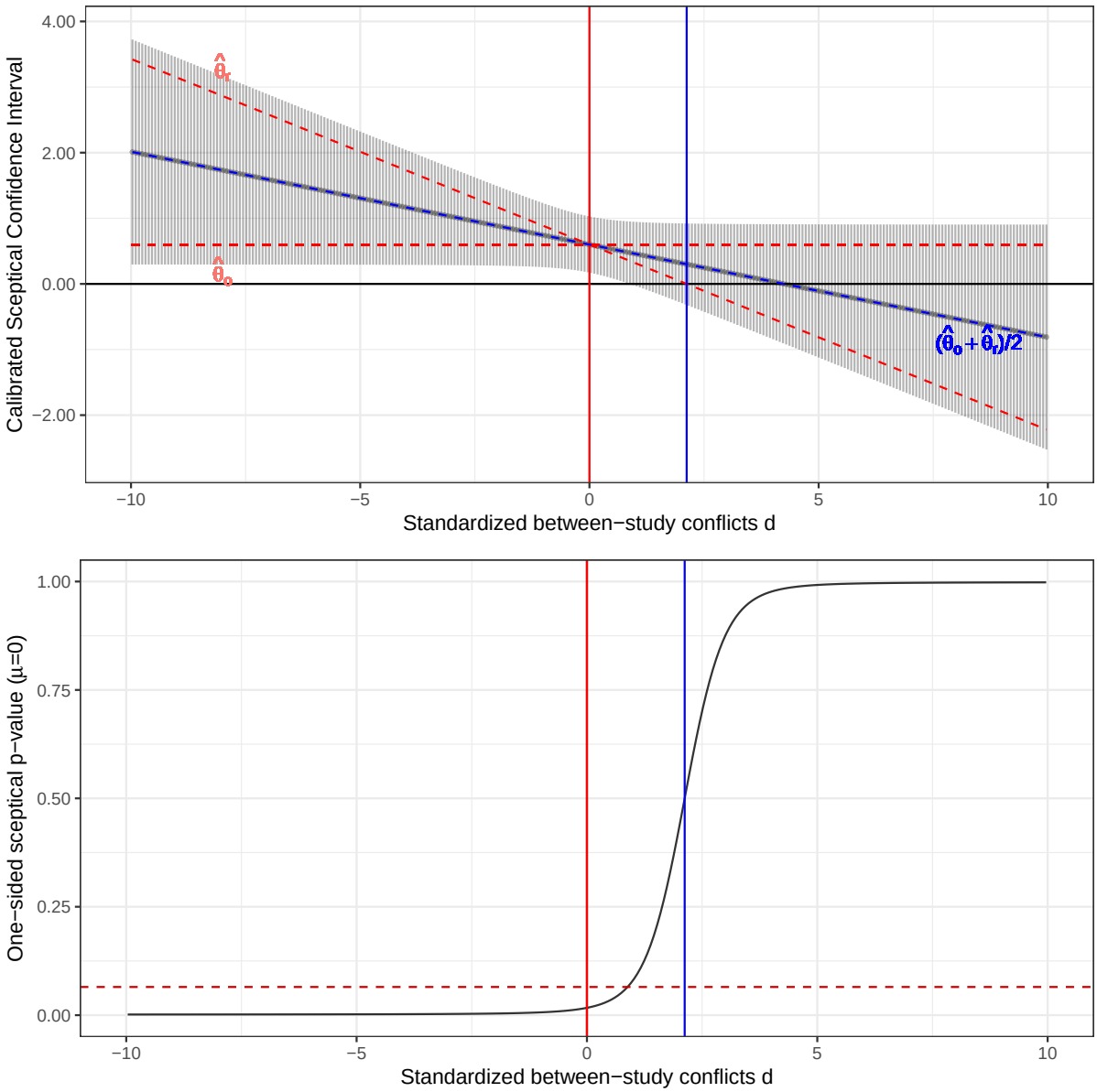

Figure 2.18: Empirical results of calibrated sceptical confidence intervals ($\widetilde{CI}_S$) and one-sided sceptical p-values ($\widetilde{p}_S$) with respect to different standardized between-study conflicts. The calibrated level $\widetilde{\alpha} = 0.065$ is plotted for sceptical p-values.

Limitations of the one-sided sceptical p-value methods

The one-sided sceptical $\widetilde{p}$-value p_S and corresponding sceptical confidence interval $\widetilde{CI}_S$ help to solve the problem of the replication paradox. However, several limitations make it not so convenient for implementation, and promote the development of alternative methods.

Firstly, the one-sided sceptical p-value and corresponding confidence interval are more appropriate for a one-sided formulated original study. However, we are more concerned about the solution for a situation that all effect estimates are positive and negative. Secondly, these two kinds of sceptical confidence intervals, CI_S and $\widetilde{CI}_S$ are both two-sided with at least an upper limit and a lower limit. Thus, terminology and explanation are ambiguous. Furthermore, even though a one-sided sceptical p-value $\widetilde{p}_S > 0.5$ carries the information about the direction conflict, the sceptical p-value function in the one-sided version is not as informative as the two-sided counterpart, since the between-study conflict is not readable anymore. Additionally, in terms of the one-sided sceptical p-value function, the range of one-sided sceptical p-values $\widetilde{p}_S$ obtained at

different μ values is from 0 to an arbitrary value below 1. For instance, the largest one-sided sceptical p-value is $\widetilde{p}_S = 0.98$ in Figure 2.16, slightly smaller than 1. We are expected to obtain even a smaller upper bound when the conflict between the original and replication study is not such huge. For this reason, the explanation for the one-sided sceptical p-value $\widetilde{p}_S$ is not as convenient as the two-sided sceptical p-value p_S, which is always from 0 to 1 as the conventional p-value, see Figure 2.4, 2.5 and 2.7. Last but not least, the well-established theorem based on the prior-data conflict in Section 2.2.6 is only applicable for the single replication. Thus the replication success assessment under multiple replications is not accessible.

Considering these shortcomings of the one-sided sceptical p-value function and corresponding sceptical confidence interval, I will introduce another method to solve the problem of replication paradox, which is originated from the approach for research synthesis for drug approval in Held (2020a).

2.4.2 Harmonic mean χ^2 p-value method

The newly proposed method for drug approval combines conventional one-sided p-values from all of the individual studies, and gives rise to a new one-sided harmonic mean χ^2 p-value for the overall treatment effect for drug regulations to make decisions (Held, 2020a).

As a by-product, this method of research synthesis can also be implemented for the assessment of replication success, based on the original and the replication study. It addresses the problem of replication paradox mentioned in Section 2.3.5, and avoids limitations of the one-sided sceptical p-value method in Section 2.4.1 at the same time.

Harmonic mean χ^2 test

To assess the overall evidence for the treatment effect, this approach is based on the harmonic mean $Z_H^2 = n/\sum_{i=1}^n 1/Z_i^2$ based on the squared Z-scores, where the number of studies $n = 2$ always holds in our case with a single replication study. The test statistics is

$$\chi^2 = nZ_H^2 = \frac{n^2}{\sum_{i=1}^n 1/Z_i^2}, \tag{2.12}$$

which is motivated by the sceptical p-value p_S and corresponding squared test statistic z_S^2 mentioned in Section 2.3.1, with the assumption of an equal sample size ($c = 1$) in Equation 2.7. The multiplicative factor n^2 ensures that the null distribution of χ^2 does not depend on n, the number of individual studies. Suppose the observed test statistic $\chi^2 = x^2$ and $x = \sqrt{x^2}$, the corresponding overall one-sided p-value can be assessed as

$$\widetilde{p}_H = \frac{Pr\{\chi^2(1) \geq x^2\}}{2^n} = \frac{1 - \Phi(x)}{2^{n-1}}. \tag{2.13}$$

The more appropriate two-sided harmonic mean χ^2 p-value $p_H = 2\widetilde{p}_H$ in a common scenario with a two-sided test of the original study $H_0 : \theta = 0$ is obtainable. The critical value, corresponding to the overall one-sided significance level α_H should be

$$c_H = \Phi^{-1}(1 - 2^{n-1}\alpha_H)^2. \tag{2.14}$$

To avoid the problem of replication paradox, the overall one-sided harmonic mean χ^2 p-value $\widetilde{p}_H$ is suggested to be adjusted. To be more precise, the overall $\widetilde{p}_H$ in Equation 2.13 cannot be larger than $1/2^n$, the probability obtaining n positive results in n studies under null hypothesis. Namely, we set an upper bound and report the inequality $\widetilde{p}_H > 1/2^n$, equally $p_H > 1/2^{n-1}$ if we focus on the two-sided harmonic mean χ^2 p-value.

For the assessment of replication success in our case, the harmonic mean χ^2 p-value works in the same way as the sceptical p-value p_S. To be specific, if $\widetilde{p}_H \leq \alpha_H$ or $p_H \leq 2\alpha_H$, a claim of success can be made at the one-sided significance level α_H.

Harmonic mean χ^2 p-value function and harmonic mean χ^2 confidence interval

The harmonic mean χ^2 p-value function and the harmonic mean χ^2 confidence interval are obtainable in the same way as the sceptical p-value method in Section 2.3.2. Specifically, the null hypothesis for the two-sided test of the original study is now extended to $H_0 : \theta = \mu$, where μ is the point estimate for the effect size.

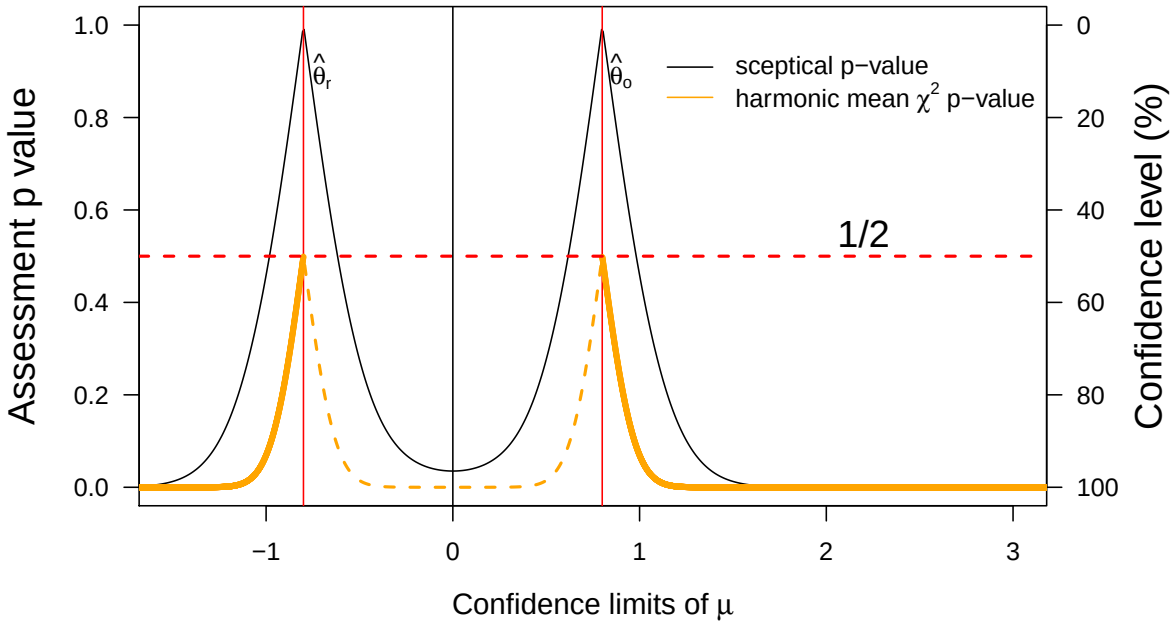

Figure 2.19: Harnomic mean χ^2 p-value function (two-sided) and sceptical p-value function (two-sided) of *Example Three* (Table 2.5)

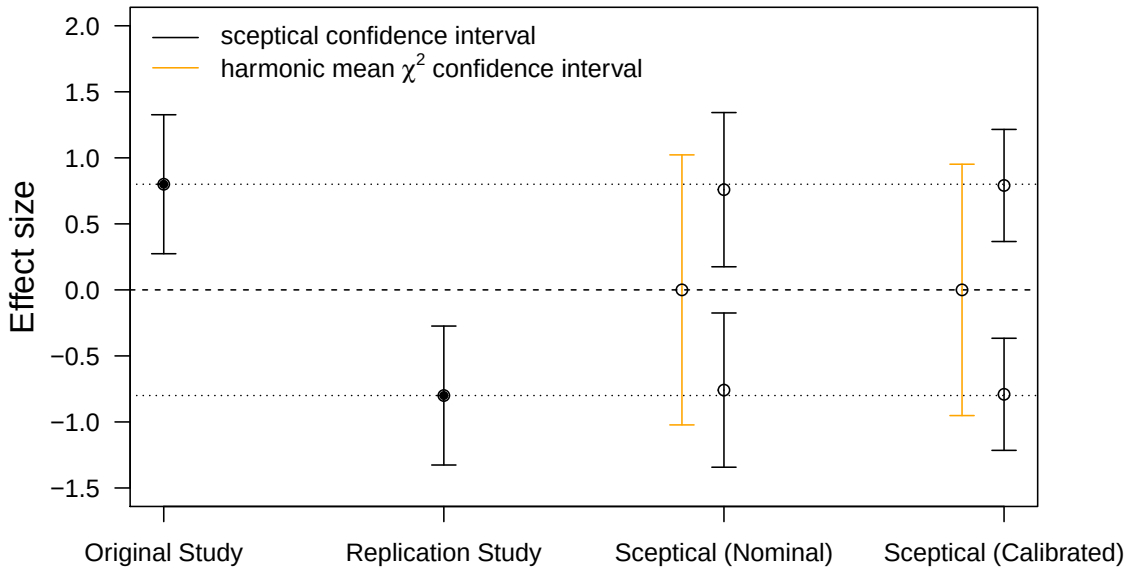

Figure 2.20: Conventional, sceptical and harnomic mean χ^2 confidence intervals of *Example Three* (Table 2.5)

Definition 2.11 (Harmonic mean χ^2 *p*-value function) *The harmonic mean χ^2 p-value function is a multimodal function with n cusps at effect size estimates of each study. An upper bound for two-sided harmonic p-values at 2^{n-1} could be obtained at the effect size estimates, $\mu = \hat{\theta}_{max}$ and $\mu = \hat{\theta}_{min}$ respectively, where $\hat{\theta}_{max} = max\{\hat{\theta}_1, ..., \hat{\theta}_n\}$, $\hat{\theta}_{min} = min\{\hat{\theta}_1, ..., \hat{\theta}_n\}$ and n is the number of investigated studies. An inequality instead of an exact value is accessible if there is a direction conflict, say $\hat{\theta}_{min} < \mu < \hat{\theta}_{max}$.*

Definition 2.12 (Harmonic mean χ^2 confidence interval) *The harmonic mean χ^2 confidence interval is a two-sided interval with an upper limit and a lower limit. Since there is an upper bound for the harmonic mean χ^2 p-value function, the harmonic mean χ^2 confidence interval is not obtainable if the specified significance level is large, i.e., $\alpha > 1/2^{n-1}$, where α is the two-sided significance level and n is the number of investigated studies.*

In Figure 2.19 and 2.20, the problem of replication paradox no longer exists. Assessment of replication success under direction conflicts can be reported as $p_H > 0.5$ as defined, and thus we will not observe unrealistically small p_H values.

The sceptical confidence interval CI_S and the harmonic mean χ^2 confidence interval CI_H at same levels are obtainable in Figure 2.20. The conclusion is that, the sceptical confidence interval is always wider than the harmonic mean χ^2 confidence interval under the same level, which means that the sceptical confidence interval is always a more stringent criterion for the assessment of replication success. Different choices of significance level α will be discussed in Section 2.5.

Multiple replications by the harmonic mean χ^2 test method

One of the merits of the harmonic mean χ^2 *p*-value (p_H) over sceptical *p*-values (p_S and $\widetilde{p}_S$) is that p_H is applicable in the setting with more than two studies, say multiple replications.

For the method of sceptical *p*-value, weight of each study was introduced to the result via the relative sample size *c*, equally the variance ratio. For the method of harmonic mean χ^2 test, weights ω_1,..., ω_n can also be introduced in Equation 2.12 (Held, 2020a), then the test statistic should be

$$\chi^2_\omega = \frac{\omega^2}{\sum_{i=1}^n \omega_i / Z_i^2}, \tag{2.15}$$

where $\omega = \sum_{i=1}^n \sqrt{\omega_i}$ is the overall weight and $\omega_i = 1/\sigma_i^2$ is the individual weight for the *i*th study. Thus the null distribution χ^2_ω does not depend on the weights ω_1,..., ω_n nor on *n*.

Table 2.6: *Example Four* (Fisher, 1999)

study number *i*	HR	log(HR)	SE	*p*-value (one-sided)
1	0.27	-1.31	0.41	0.00025
2	0.22	-1.51	0.85	0.0245
3	0.72	-0.33	0.29	0.128
4	0.57	-0.56	0.51	0.1305
5	0.53	-0.63	1.02	0.2575

To show how the method of harmonic mean χ^2 test works in the setting of multiple replications, *Example Four* originated from Fisher (1999) and discussed in Held (2020a) is summarized in Table 2.6. Considering 5 clinical trials on the effect of Carvediol, hazard ratios (HR) for

the treatment are listed. For simplicity, I still use $\hat{\theta}_i$ to denote the treatment effect estimates of individual studies, which corresponds to log hazard ratios log(HR). Standard errors of effect estimates SE are also available. This example was aimed at the evidence synthesis, but it is portable for our setting of replication success assessment under multiple replications, since the assessment of replication success or the integration of overall treatment effect does not depend on the order of individual studies in Equation 2.12 and 2.15.

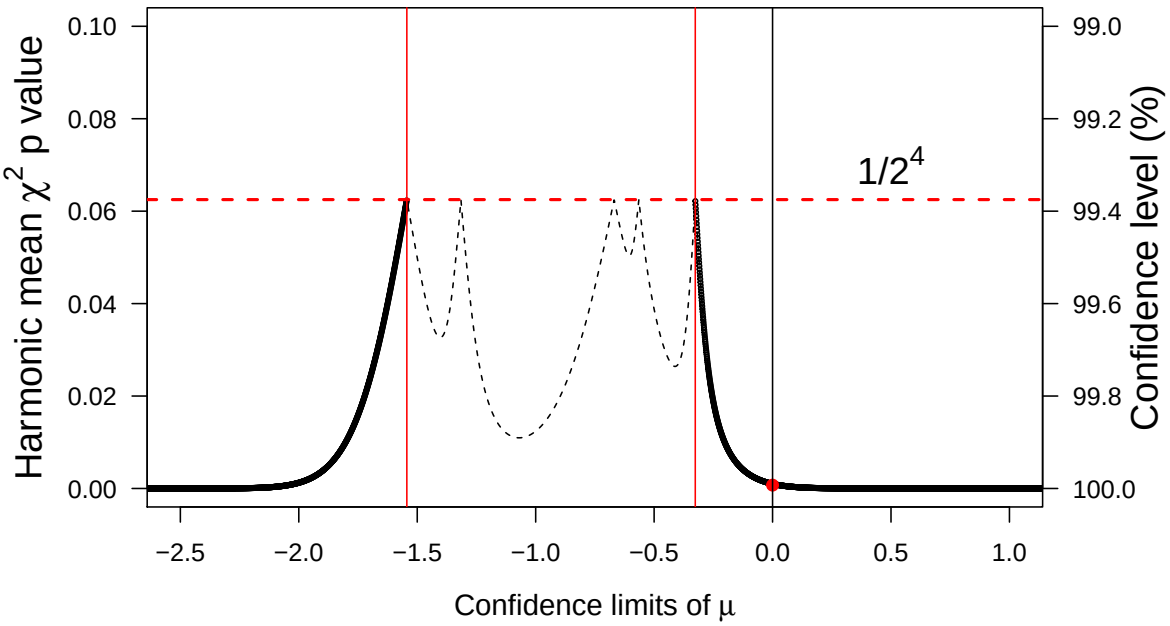

Figure 2.21: Unweighted harmonic mean χ^2 p-value function of *Example Four*. The horizontal line in red represents the upper bound depending on number of studies n.

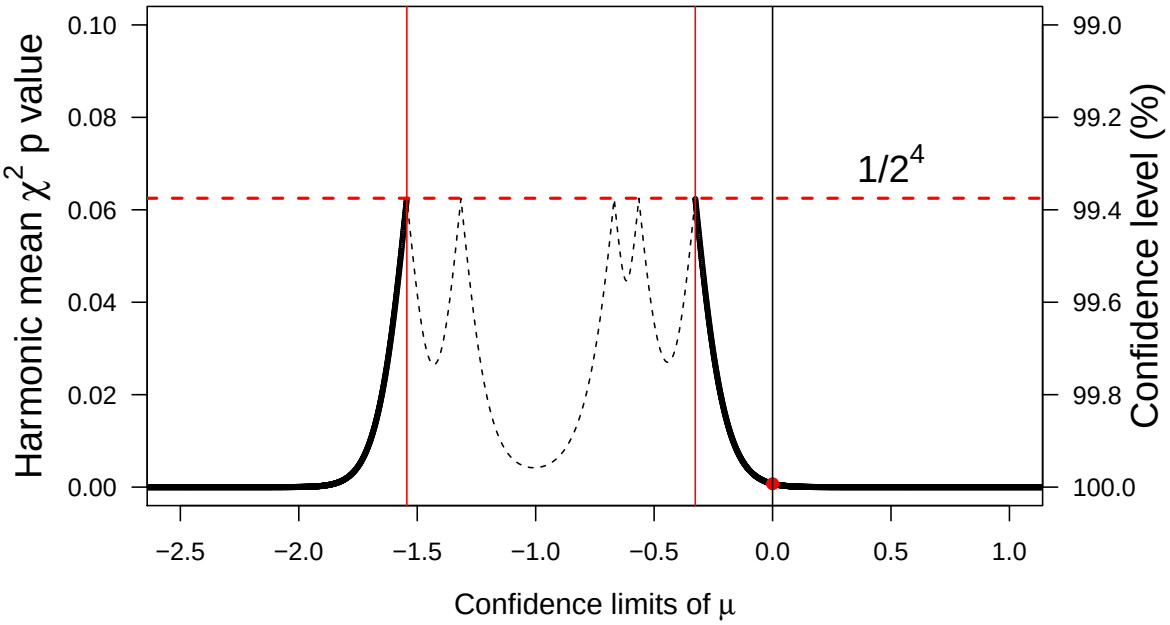

Figure 2.22: Weighted harmonic mean χ^2 p-value function of *Example Four*. The horizontal line in red represents the upper bound depending on number of studies n.

The unweighted and weighted harmonic mean χ^2 p-value function, corresponding to Equation 2.12 and 2.15 are illustrated in Figure 2.22 and 2.21. These two harmonic mean χ^2 p-value

functions against various μ values are always in the two-sided version. The weighted and unweighted ones are slightly different as they should be. In this thesis, I focus on what they have in common to show the properties of the harmonic mean χ^2 p-value function.

When compared to the weighted harmonic mean χ^2 p-value function for two investigated studies in Figure 2.19, there are 5 cusps in *Example Four*, corresponding to effect size estimates of 5 individual studies. We are likely to obtain the largest p_H values exactly at the most extreme treatment effect size estimates $\hat{\theta}_{\min}$ and $\hat{\theta}_{\max}$, where $\hat{\theta}_{\min} = \min\{\hat{\theta}_1, ..., \hat{\theta}_n\}$ and $\hat{\theta}_{\max} = \max\{\hat{\theta}_1, ..., \hat{\theta}_n\}$. The largest p_H value or the upper bound of the function is $1/2^{n-1}$, depending on the number of studies n. The direction conflict in this scenario, with multiple replication studies, will be observed if the largest test statistics and the smallest test statistics are in opposite directions, say $\hat{\theta}_{\min} < \mu < \hat{\theta}_{\max}$. We report the inequality $p_H > 1/2^{n-1}$ when there is a direction conflict, which corresponds to the middle part of the vertical lines in Figure 2.21 and 2.22.

Various harmonic mean χ^2 confidence intervals given different confidence levels are also obtainable. Unusual set of intervals are evitable because of the adjustment for p_H. The overall harmonic mean χ^2 confidence interval will include all of the individual effect size estimates. That is, the calculated upper limit of CI_H is larger than $\hat{\theta}_{\max}$, and the lower limit is smaller than $\hat{\theta}_{\min}$. For this reason, the direction conflict always leads to a claim of replication failure, since $\mu = 0$ will be included in the calculated harmonic mean χ^2 confidence interval.

The harmonic mean χ^2 p-value we obtained at $\mu = 0$ is $p_H = 0.0007$, which is much less informative than the harmonic mean χ^2 p-value function or the harmonic mean χ^2 confidence interval at a specified confidence level. The specification of an appropriate significance level should be based on the type-I error control under multiple studies, more details will be discussed in Section 2.5.

Discussion about the harmonic mean χ^2 test method

The harmonic mean χ^2 p-value function and the harmonic mean χ^2 confidence interval have several advantages over the sceptical p-value function and the sceptical confidence interval for the assessment of replication success.

Firstly, methods based on the harmonic mean χ^2 test solve the problem of the replication paradox in the two-sided sceptical p-value function and the corresponding confidence interval. Namely, undesired small p_H values and unwanted separation between intervals no longer exist. Secondly, they work for the general two-sided test for the original study, thus favorable over the methods based on one-sided sceptical p-value. Thirdly, the harmonic mean χ^2 p-value function carries adequate information. For example, the upper bound of the harmonic mean χ^2 p-value is fixed and known, effect size estimates of individual studies are obtainable at the cusps of this function. Additionally, the interpretation is quite convenient, no need to distinguish between the one-sided and two-sided version, since this method is based on the two-sided test of the original study. Finally, the range of p_H values is fixed, and the upper bound depends only on the number of studies n.

2.4.3 Meta-analysis

The natural method to integrate the data obtained from multiple studies is meta-analysis. The combination of the findings represents an attractive method to strengthen the evidence in evidence-based medicine (Glass, 1976; DerSimonian and Laird, 1986). It is of high interest to compare different methods for the combination of evidence from the original and replication studies to assess replication success. In this section, the meta-analysis is introduced and the comparison of different methods will be discussed in Chapter 3 by concrete examples.

There are two popular statistical models for meta-analysis (Borenstein et al., 2010), the fixed-effects model and the random-effects model. Under the fixed-effects model we assume that the true effect size for all studies is identical, and the only reason that the effect size varies between

studies is sampling error in estimating the effect size. By contrast, under the random-effects model we allow the true effect sizes to differ and the goal is not to estimate one true effect, but to estimate a distribution of these effects (Higgins et al., 2009).

To conduct a random-effects model, we have to specify the between-study-variance estimator τ^2 in advance. There is more than one choice for this estimator (Veroniki et al., 2016). Different estimators derive τ^2 using slightly different approaches, leading to somewhat different pooled effect size estimates and confidence intervals. In medical and psychological research, the most often used estimator is the *DerSimonian-Laird estimator*, which is also the default approach in many software and also in this thesis (DerSimonian and Laird, 1986).

The selection of these two kinds of meta-analysis model is critically important as it does not only affect the computations but also helps to define the goal of the analysis and the interpretation of the results. In Chapter 3, I will discuss more about the appropriate choice of model for meta-analysis with the aim of the assessment of replication success.

2.5 Comparison with the two-trials rule

All three confidence interval methods and corresponding p-value functions mentioned before work for the assessment of replication success. The differences are summarized in Table 2.7. In this section, I compare these criteria with the 'two-trials rule' (Kay, 2014), to show how these different criteria work for the drug approval.

The two-trials rule is the standard for drug approval, which requires adequate and well-controlled investigations to determine efficacy (Bobka, 1993). The guidance of the Food and Drug Administration (FDA) suggests that at least two primary studies testing the same medical product for drug regulation administrations to make decisions. Such studies are known as 'pivotal' efficacy trials (Food and Drug Administration, 1998). The reproducibility of specific results from the first study is confirmed by the second one, and the establishment of replication success is supported.

Table 2.7: The comparison of replication success assessment methods

	Critical value	Replication Paradox	Assessment
$p_S = 2\{1-\Phi(z_S)\}$	$z_S^2 = [\Phi^{-1}(1-\alpha_S)]^2$	No adjustment	$p_S \leq 2\alpha_S$
$\widetilde{p}_S = 1-\Phi(z_S)$	$z_S^2 = [\Phi^{-1}(1-\alpha_S)]^2$	$\widetilde{p}_S = 1-p_S/2$	$\widetilde{p}_S \leq \alpha_S$
$p_H = 2\{1-\Phi(2z_S)\}$	$z_H^2 = \left[\Phi^{-1}(1-2\alpha_H)\right]^2$	$p_H > 0.5$	$p_H \leq 2\alpha_H$

Note: $z_S^2 = 1/(1/t_o^2 + 1/t_r^2)$ and $z_H = 2z_S$. Significance levels α_S and α_H are in one-sided versions.

In Table 2.7, all measures are motivated by half of the harmonic mean of squared z-values, $z_S^2 = 1/(1/t_o^2 + 1/t_r^2)$. The test statistics t_o and t_r can be regarded as the same as the test statistic z_o and z_r in the method of harmonic mean χ^2 test in Section 2.4.2. The major difference between using a Z-score and a T-statistic is that the population standard deviation should be estimated when using a Z-score. Besides, a T-statistic is also used when the sample size is small. In what follows, I will always use t_o and t_r as the test statistics. The one-sided sceptical p-value $\widetilde{p}_S$ and the two-sided harmonic mean χ^2 p-value p_H get rid of the problem of replication paradox in a similar way, see Section 2.4. Unexpected small assessment p-values are redefined if these values are obtained when the direction of the original study cannot be substantiated, say the direction conflict.

Figure 2.23 with the test statistic of the original study t_o and the test statistic of the replication study t_r shows how these different criteria work for the drug approval. When it comes to the two-trials rule, the acceptance regions for drug approval are shown in red in the upper right and the bottom left corners, where two trials should have the results in the same direction (i.e., $\tilde{p} \leq 0.025$ in favor of the experiment treatment). Both original and replication studies are required to be significant at the one-sided level $\widetilde{\alpha} = 0.025$. For the one-sided sceptical p-value method ($\widetilde{p}_S$), the acceptance regions for drug approval are out of these two black curves, either the upper right corner or the bottom left one, depending on the assumption of direction of the original study. In terms of the harmonic mean χ^2 test approach, the

corresponding acceptance regions are out of the blue curves, which is quite similar to the method of the one-sided sceptical p-value.

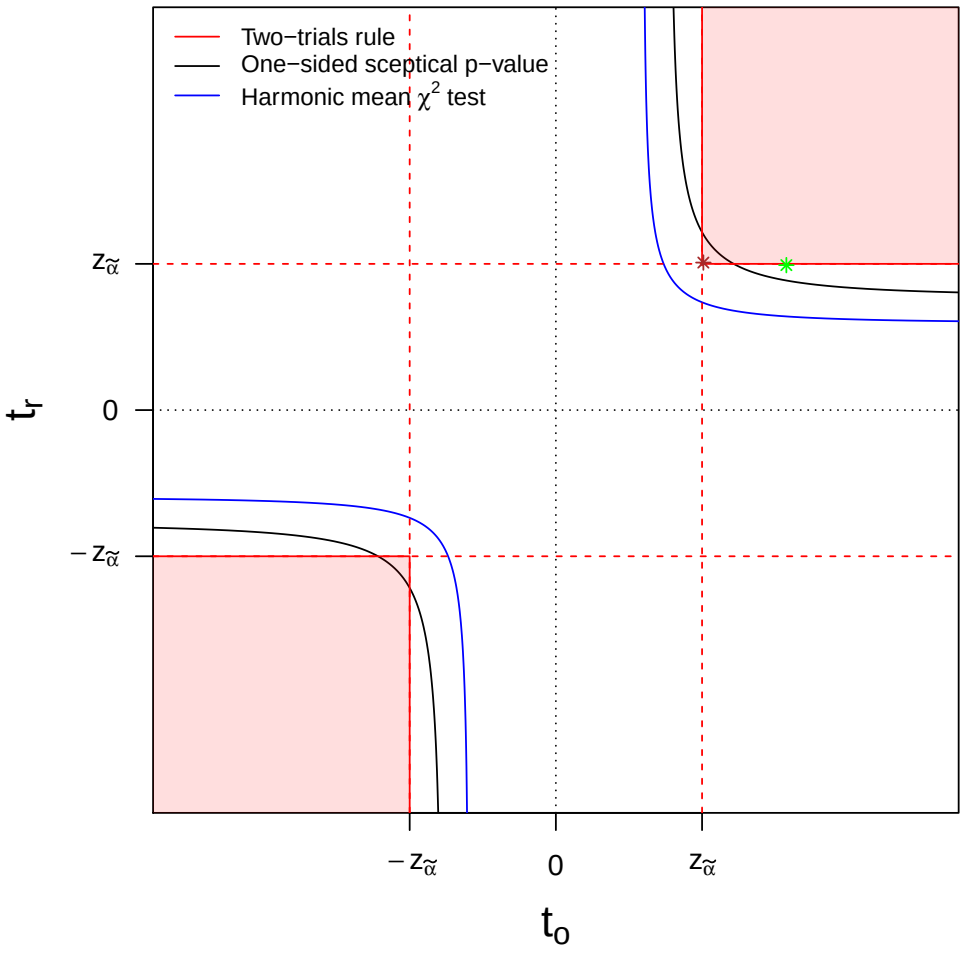

Figure 2.23: Comparison of different assessment methods for replication success, $\widetilde{\alpha}$ is the one-sided significance level.

In all of these three methods, Type-I error regarding the false-positive result is 0.025 (1 in 40). That is if the positive treatment effect is confirmed to be ineffective, then we are able to see two positive trials accidentally only 1 in 1600 cases, since the corresponding overall Type-I error for two trials is $0.025 \times 0.025 = 0.000625$ (1 in 1600) (Kay, 2014).

Table 2.8: Different bounds on the one-sided study specific p-values p_i

overall p-value	α_H	bound	critical value	study-specific p_i
$\widetilde{p}_H$	1/1600	sufficient	z_H^2=9.14	$p_i = 0.016$
		necessary	z_H^2=9.14	$p_i = 0.065$
		liberal	z_H^2=7.68	$p_i = 0.025$

It is straightforward to calculate the study specific p-value p_i, based on the bound of critical values as shown in Table 2.7, under the same type-I error control $\alpha_H = 1/1600$. Thus we are able to achieve an appropriate significance level, see Appendix A.4. Table 2.8 has illustrated different options based on different bounds for the critical values. It means that the requirement for study-specific p_i, namely, the significance level for individual studies, should be satisfied to make a claim of replication success. For

instance, the necessary criterion in Table 2.7 indicates the requirement $p_i \leq 0.065$, i=1, 2, for a claim of replication success based on n=2 studies.

There is more than one choice for the appropriate choice of the significance level for assessment, equally the calculated study-specific p-value p_i. In terms of the harmonic mean χ^2 test method, the sufficient bound and necessary bound are both applicable, while the liberal version gives us a different viewpoint. Since the calculated study-specific p_i of the sceptical p-value method requires the assumption of an equal sample size, more details see Appendix A.4. It is reasonably suggested to use the calculated study-specific p-value p_i of the harmonic mean χ^2 $\widetilde{p}_H$ for the sceptical p-value method. To be specific, the liberal version of p_i provides the nominal level at $\widetilde{\alpha} = 0.025$ for the sceptical p-value method, but the critical values is different from that of the harmonic mean χ^2 p-value method. The calibrated level $\widetilde{\alpha} = 0.065$ originated from the necessary bound for the harmonic mean χ^2 $\widetilde{p}_H$ is also applicable, see *Example One*, *Example Two* and *Example Three* in Section 2.3 and 2.4.

Two simple examples are illustrated in Table 2.9 to show how these criteria work differently in a same scenario, which is consistent with the criteria shown in Figure 2.23. Different criteria lead to different results for these two examples. The two-trial rule leads to non-approval or no replication success with $p_o = 0.0001$ and $p_r = 0.0026$ (the green point in Figure 2.23), but an approval or a replication success is supported if both trials have $p_o = p_r = 0.024$ (the brown point in Figure 2.23). With the implementation of the sceptical p-value method, we come to the opposite results in both cases. That is, non-significant results of the replication study may support the claimed result of the original research findings, while significant result of the replication study may not be under a non-significant result of the original study. When it comes to the approach of harmonic mean χ^2 test, it seems to be less stringent for the replication success assessment when compared with the other two methods, even though we use a low significance level (the liberal version). This explains why the harmonic mean χ^2 confidence interval CI_H is narrower than the sceptical confidence interval CI_S under the same significance level in the example in Figure 2.20.

Table 2.9: Comparison of different criteria for replication success

Method	t-statistics	$\widetilde{\alpha}$	$p_o = p_r = 0.024$	$p_o = 0.001$, $p_r = 0.026$
Two-trails rule	t_o, t_r	0.025	Success	No success
Sceptical p-value	$z_S^2 = 1/(1/t_o^2 + 1/t_r^2)$	0.065	No success	Success
Harmonic mean χ^2 test	$z_H^2 = 2/(1/t_o^2 + 1/t_r^2)$	0.025	Success	Success

Note: p_o, p_r are in one-sided versions, and the one-sided significance level is $\widetilde{\alpha}$.

To summarize, a non-significant result of a replication study, that fails to satisfy the two-trials rule, does not necessarily mean that the confirmed findings by an original study should be refused. Drug approval will be encouraged by means of alternative criteria. The widely-recognized replication crisis may not be a real crisis but the problem with the assessment measure used. The method of the harmonic mean χ^2 test and the method of the sceptical p-value provide us with a more reliable perspective to assess the replication success.

2.6 Data

In this master thesis, the dataset *RProjects* with 143 studies in total, is used to implement methods mentioned in this Chapter. The dataset from the R package *ReplicationSuccess* (Held, 2020b), is based on four large-scale replication projects, where there is only one replication study for each original study, say, single replication.

In all studies, effect size estimates are transformed to correlation coefficients r and Fisher z-transformed correlation coefficients $\hat{\theta} = tanh^{-1}(r)$. Correlation coefficient r has several virtues over alternative measures such as Cohen's d. Correlation coefficients are readily interpretable as they are bounded. Besides, the Fisher z-transformation makes the analysis more straightforward as their standard error can be expressed in the form of the sample size, $\text{se}(\theta) = 1/\sqrt{n-3}$. Thus, correlation coefficients has become the most common used metric for effect size in replication projects (Open Science Collaboration, 2015).

- **Reproducibility Project: Psychology (Open Science Collaboration, 2015)**
  As mentioned in Chapter 1, this project attempted to reproduce 100 studies from the field of psychology. Among these studies, 73 are selected to the dataset *RProjects*, based on which the standard errors of the Fisher z-transformed effect estimates are available (Johnson et al., 2017).

- **Experimental Economics Replication Project (Camerer et al., 2016)**

  There are 18 published experimental economics studies replicated in this project, which are taken from two high impact economic journals between 2011 and 2015.

- **Social Sciences Replication Project (Camerer et al., 2018)**

  In the project for social sciences, 21 published studies included are taken from the journals *Nature* and *Science* between 2010 and 2015.

- **Experimental Philosophy Replicability Project (Cova et al., 2019)**

  In this project, 40 replications of experimental philosophy studies were involved. The original studies were published in one of 35 specified journals between 2003 and 2015. In the package *RProjects*, 31 study pairs were selected based on which the effect estimates on correlation scale and the effective sample sizes for both the original and replication studies are available.

2.7 Software

All analyses were conducted in the R programming language (R Core Team, 2020), with base packages and some specific packages for analysis: Dataset and functions for the assessment methods in this thesis are from the package *ReplicationSuccess* (Held, 2020b); Figures are generated by the packages *plotrix* (Lemon J, 2006), *ggplot2* (Wickham, 2016), *ggpubr* (Kassambara, 2020), and *ggrepel* (Slowikowski, 2020); Meta-analysis are conducted by the *meta* package (Balduzzi et al., 2019); The nested tables were generated using the package *xtable* (Dahl et al., 2019). Formats are organized via the package *biostatUZH* (Haile et al., 2019). The functions for newly developed methods in this thesis are available in Appendix A.5.

Chapter 3

Results

In this chapter, the different methods mentioned in Chapter 2 are implemented on the four replication project datasets (see Section 2.6). The comparison of these methods then will be discussed. Confidence intervals for the assessment are calculated based on Fisher z-transformed correlation coefficients and corresponding standard errors.

In Section 2.5, different choices for the bounds for the one-sided study-specific p_i are illustrated. The best choice of the bound requires more discussion. For simplicity, the two-sided calibrated significance level $\alpha = 0.13$ (equally one-sided at $\widetilde{\alpha} = 0.065$) and the two-sided nominal level $\alpha = 0.05$ (equally one-sided at $\widetilde{\alpha} = 0.025$) will be implemented by these various methods in this chapter.

3.1 Replication projects

Different methods have different upsides and downsides as discussed in Chapter 2. In this section, I would like to implement these methods on datasets from the R package *ReplicationSuccess* (Held, 2020b) to show the availability of different instruments.

3.1.1 One-sided sceptical p-value

In this section, I focus on the one-sided sceptical p-values of these replication projects. The one-sided sceptical p-values are preferable to two-sided sceptical p-values, such that there is no problem of the replication paradox due to unrealistically small sceptical p-values.

In Figure 3.1, one-sided sceptical p-values are plotted against the absolute values of the standardized between-study conflicts $|d|$, defined in Equation 2.8, for all 143 studies in these four replication projects. All of these one-sided sceptical p-values are obtained at $\mu = 0$ and are adjusted to values larger than 0.5, when the direction conflict defined in Section 2.3.2 exists.

The claim of replication success will depend only on the specified significance level. We make claims of replication success on the nominal level if $\widetilde{p}_S < 0.025$. Similarly, claims of replication success can be achieved on the calibrated level if $\widetilde{p}_S < 0.065$. The vertical line in red indicates the threshold for unusual sceptical confidence set $|d| = 2z_{\alpha/2}$, equally $|d| = 2z_{\widetilde{\alpha}}$ where $\widetilde{\alpha} = \alpha/2$. Even though the threshold is originated from the method of sceptical confidence interval or set, it makes sense to roughly infer the sceptical p-value based on this threshold. To be specific, a large one-sided sceptical p-value $\widetilde{p}_S$ will be expected if the absolute standardized between-study conflict is relatively large $|d| > 2z_{\widetilde{\alpha}}$. There are some rare cases that $\widetilde{p}_S$ values are relatively small even the between-study conflicts $|d|$ values are large. For instance, in the replication project of *Psychology*, the study with the largest between-study conflict $|d|$ has a quite small one-sided sceptical p-value $\widetilde{p}_S = 0.14$. The reason for this irregular case will be more clearly illustrated with the method of sceptical confidence interval in Section 3.1.2.

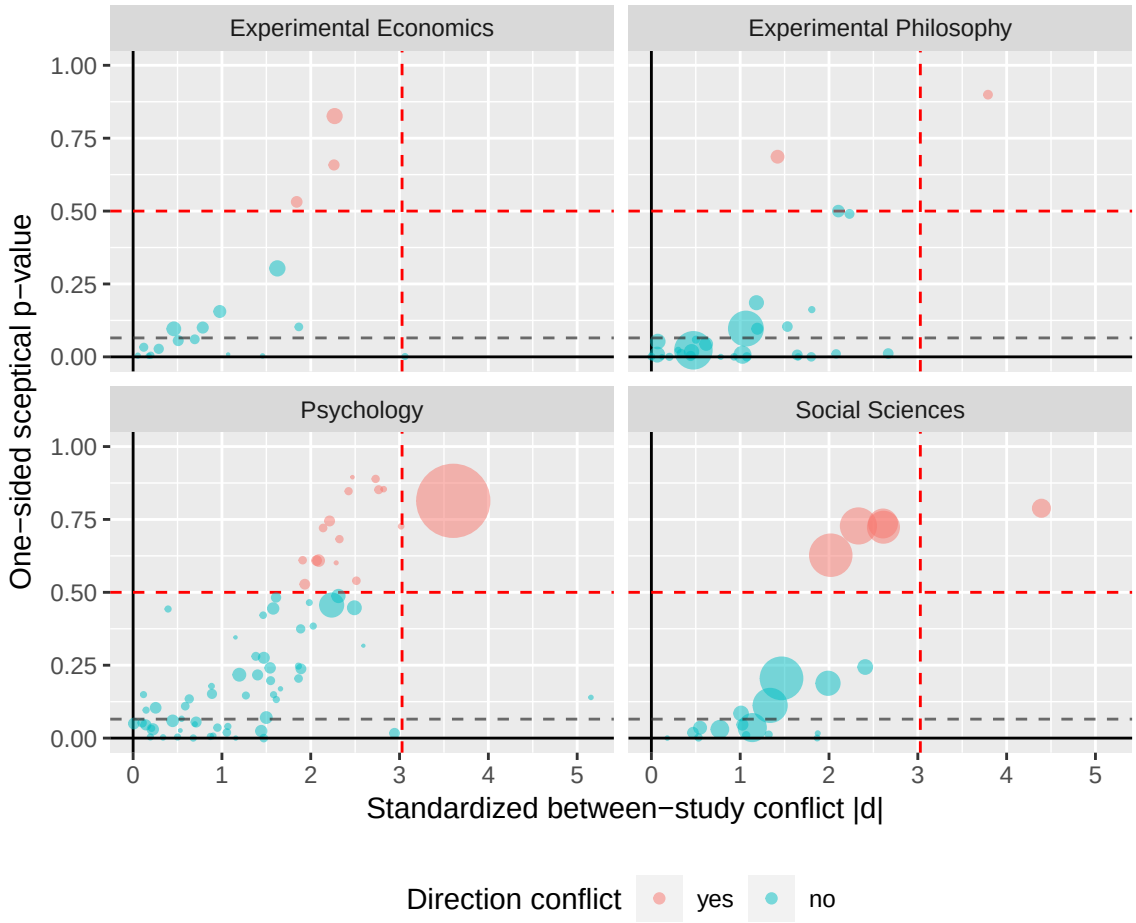

Figure 3.1: One-sided sceptical p-values for the replication projects. Relative sizes c are expressed by the size of circles. The horizontal line in black is the calibrated level $\widetilde{\alpha} = 0.065$. The vertical line in red represents the threshold for sceptical confidence set under the equal sample size assumption $|d| = 2z_{\widetilde{\alpha}}$.

3.1.2 Sceptical confidence interval

In this section, sceptical confidence intervals in the two-sided version (CI_S) for these replication success projects are demonstrated. Corresponding significance levels at the nominal one $\alpha = 0.05$ and the calibrated one $\alpha = 0.13$ show different results of the sceptical confidence intervals CI_S.

Even though the two-sided sceptical p-value p_S is not practically valuable, when compared with the one-sided sceptical p-value $\widetilde{p}_S$, due to the replication paradox mentioned in Section 2.3.5. The sceptical confidence interval CI_S in a two-sided version could provide us more information when compared with the counterpart in a one-sided version $\widetilde{CI}_S$. The possible separation reveals the information about distinguishable conflicts between the original and the replication studies. For instance, suppose there are substantially significant results of both original and replication study, we could achieve a quite small sceptical p-value (either p_S or $\widetilde{p}_S$), or a sceptical confidence interval ($\widetilde{CI}_S$) far away from $\mu = 0$. Yet the sceptical confidence interval CI_S may take the message about discrepancy between the original and the replication study if it comes to a set with two disjoint intervals.

In Figure 3.2, nominal sceptical confidence intervals of all studies from these four replication projects are plotted against the absolute standardized between-study conflicts $|d|$. Normally, the original studies are always more significant than the replication studies (Open Science Collaboration, 2015).

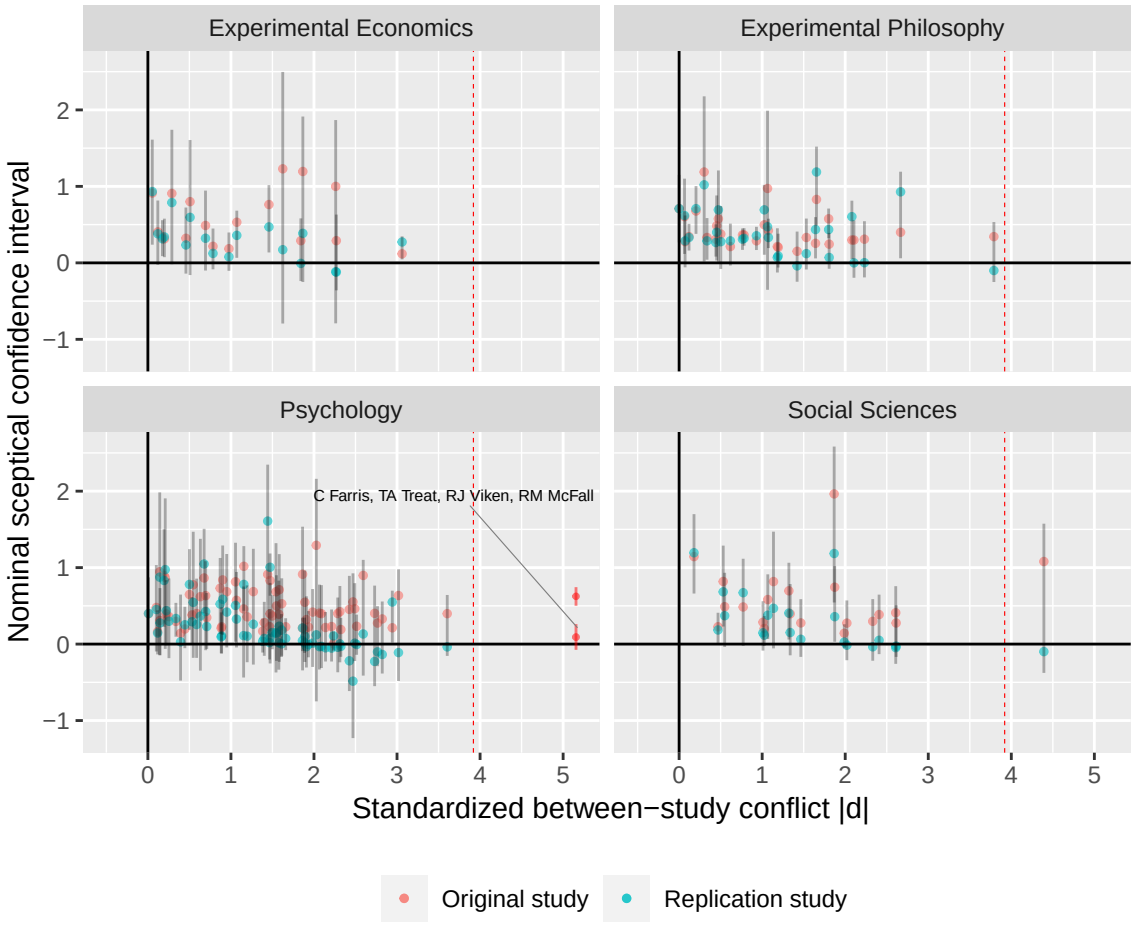

Figure 3.2: Nominal sceptical confidence intervals for the replication projects at $\alpha = 0.05$. The vertical line in red represent the threshold for sceptical confidence sets under the equal sample size assumption $|d| = 2z_{\alpha/2}$.

There is only one unusual case with the sceptical confidence set from the replication project of *Psychology*, that is *C Farris, TA Treat, RJ Viken, RM McFall*. This study is the one with the largest standardized between-study conflict $|d|$ within the replication project of *Psychology*. It is also the special case in Figure 3.1, which has a quite small one-sided sceptical p-value $\widetilde{p}_S$ even with a large $|d|$. With the method of the one-sided sceptical p-value, the conclusion is that this study cannot be concluded as a replication success as $\widetilde{p}_S = 0.14$ is larger than both the nominal and calibrated levels $\widetilde{\alpha} = 0.025$ and $\widetilde{\alpha} = 0.065$. The method of sceptical confidence interval leads to a set of intervals [-0.07, 0.26] and [0.50, 0.74]. The reason of a declare of no replication success is because $\mu = 0$ is included in one of these two intervals, say one of these two investigated studies is not significant enough. Additionally, we can conclude that there is a huge discrepancy between the original study and the replication study, which even results in an unusual sceptical confidence set.

The study with largest standardized between-study conflict $|d|$ from *Social Sciences* has a $|d|$ value larger the threshold (see Section 2.3.4), say $|d| > 2z_{\alpha/2}$. However, there is no unusual sceptical confidence set since the relative sample size c =4.47 is much larger than 1. As mentioned in Section 2.3.6, a pair of original and replication study with a relatively large relative sample size c is less likely to result in an unusual sceptical confidence set.

More unusual sceptical confidence sets are expected at the calibrated level $\alpha = 0.13$ as shown in Figure 3.3. The calibrated sceptical confidence intervals are always narrower than the corresponding nominal ones at $\alpha = 0.05$ in Figure 3.2. Additionally, there are more cases of sceptical confidence set at calibrated level as we expect. Across all these four replication project, the four studies with the largest standardized between-study conflict $|d|$ lead to a sceptical confidence set. All of these four studies has a standardized between-study conflict $|d|$ larger than the threshold for the unusual sceptical confidence set, say $|d| > 2z_{\alpha/2}$.

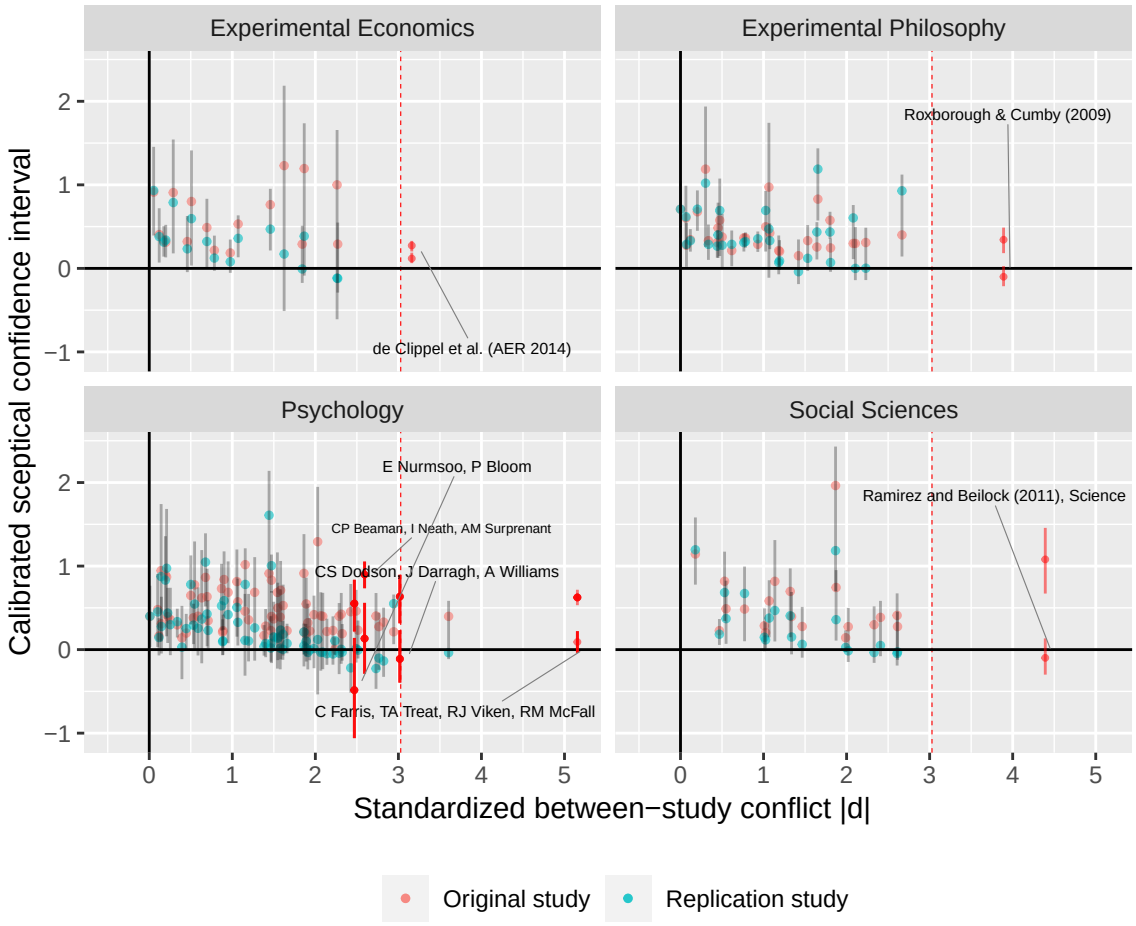

Figure 3.3: Calibrated sceptical confidence intervals for the replication projects at $\alpha = 0.13$. The vertical line in red represent the threshold for sceptical confidence sets under the equal sample size assumption, say $|d| = 2z_{\alpha/2}$.

However, there are other three cases in the replication project of *Psychology*: *'CP Beaman, I Neath, AM Surprenant'*, *'CS Dodson, J Darragh, A Williams'*, and *'E Nurmsoo, P Bloom'*, from which we observe unusual sceptical confidence sets even though under relatively smaller between-study conflicts ($|d| < 2z_{\alpha/2}$). It shows that the relative sample size c could affect the result of sceptical confidence interval. As shown in Table 3.1, all of these three unusual cases have comparatively small relative sample size, 0.13, 0.83, and 0.23 respectively.

Table 3.1 illustrates all of these unusual cases of sceptical confidence sets from these four projects in Figure 3.3. Names of the studies are consistent with the dataset in the R package *ReplicationSuccess* (Held, 2020b), some of which carry the information about the publication year. Most of the unusual cases of sceptical confidence sets occur when the between-study conflict is large enough, but this is not a rule as the relative sample sizes are different across different studies. If there is no distinguishable between-study conflicts $|d|$, the sceptical confidence set could also happen due to a small relative sample size c.

Table 3.1: Unusual cases with sceptical confidence sets

Study	Project	t_o	t_r	c	$\lvert d\rvert$
de Clippel et al. (AER 2014)	Experimental Economics	3.32	7.61	0.99	3.06
Roxborough & Cumby (2009)	Experimental Philosophy	3.68	-1.42	1.75	3.79
CP Beaman, I Neath, AM Surprenant	Psychology	8.89	0.48	0.13	2.59
CS Dodson, J Darragh, A Williams	Psychology	3.81	-0.61	0.83	3.02
C Farris, TA Treat, RJ Viken, RM McFall	Psychology	10.40	1.09	0.51	5.16
E Nurmsoo, P Bloom	Psychology	3.02	-1.28	0.23	2.47
Ramirez and Beilock (2011), Science	Social Sciences	4.45	-0.86	4.47	4.39

Note: t_o and t_r are test statistics of these two studies, d is the standardized between-study conflicts and the threshold is at d=2z$_{\alpha/2}$ = 3.03.

3.1.3 Harmonic mean χ^2 p-value

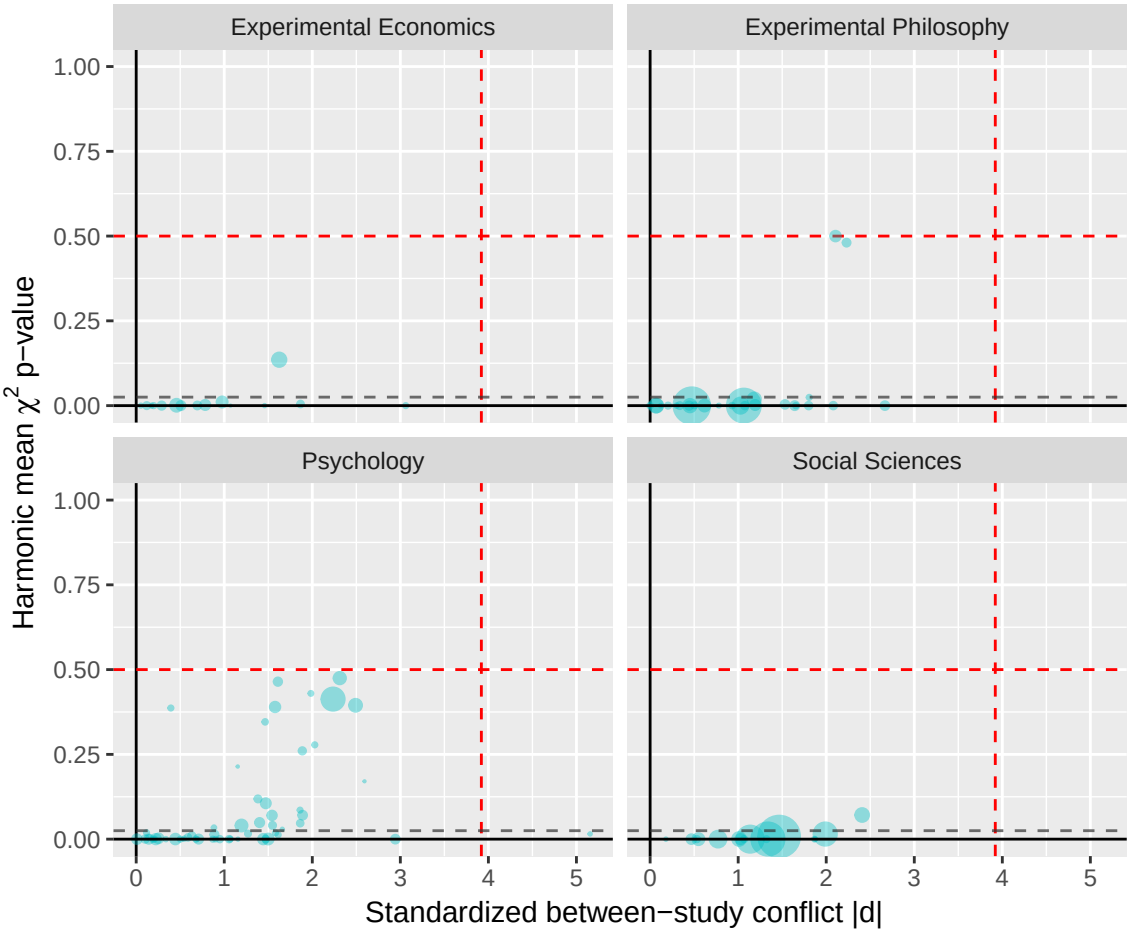

Figure 3.4: Harmonic mean χ^2 p-values for the replication projects. Relative sample sizes c are expressed by the size of circles. The horizontal line is the nominal at $\alpha = 0.05$. The vertical line in red represents the threshold for sceptical confidence set $|d| = 2z_{\alpha/2}$.

To avoid the replication paradox, we keep the harmonic mean χ^2 p-values only when there are no direction conflicts. This adjustment is achieved by setting an upper bound for these p-values. As shown in Figure 3.4, there are no harmonic mean χ^2 p-values larger than 0.5, since studies with direction conflicts have no exact harmonic mean χ^2 p-values. Instead, we could only obtain an inequality, see Section 2.4.2.

Similarly, borrowing the idea of the threshold $|d| = 2z_{\alpha/2}$ for the sceptical confidence set defined in Equation 2.8, we are able to make some initial inferences about the harmonic mean χ^2 p-values p_H. Normally, p_H values are reasonably inaccessible when there are large standardized between-study conflicts $|d|$, since large values of $|d|$ (i.e., $|d| > 2z_{\alpha/2}$) usually indicate the direction conflicts. The relationship between the threshold and the bound for direction conflict is discussed in Section 2.3.4 and Appendix A.3.

3.1.4 Harmonic mean χ^2 confidence interval

The harmonic mean χ^2 confidence interval in Figure 3.5 is favorable over the harmonic mean χ^2 p-value in Figure 3.4. Firstly, the magnitude of effect size is demonstrated, just like the reason why traditional confidence interval is always preferable to traditional p-value for the interpretation of results. Secondly, with the direction conflict, we cannot obtain exact values for harmonic mean χ^2 p-value, since we set a boundary to adjust for the replication paradox, see Figure 3.4. Whereas the harmonic mean χ^2 confidence intervals are still obtainable, once the significance level is not larger than the upper bound, say $\alpha \leq 2^{n-1}$.

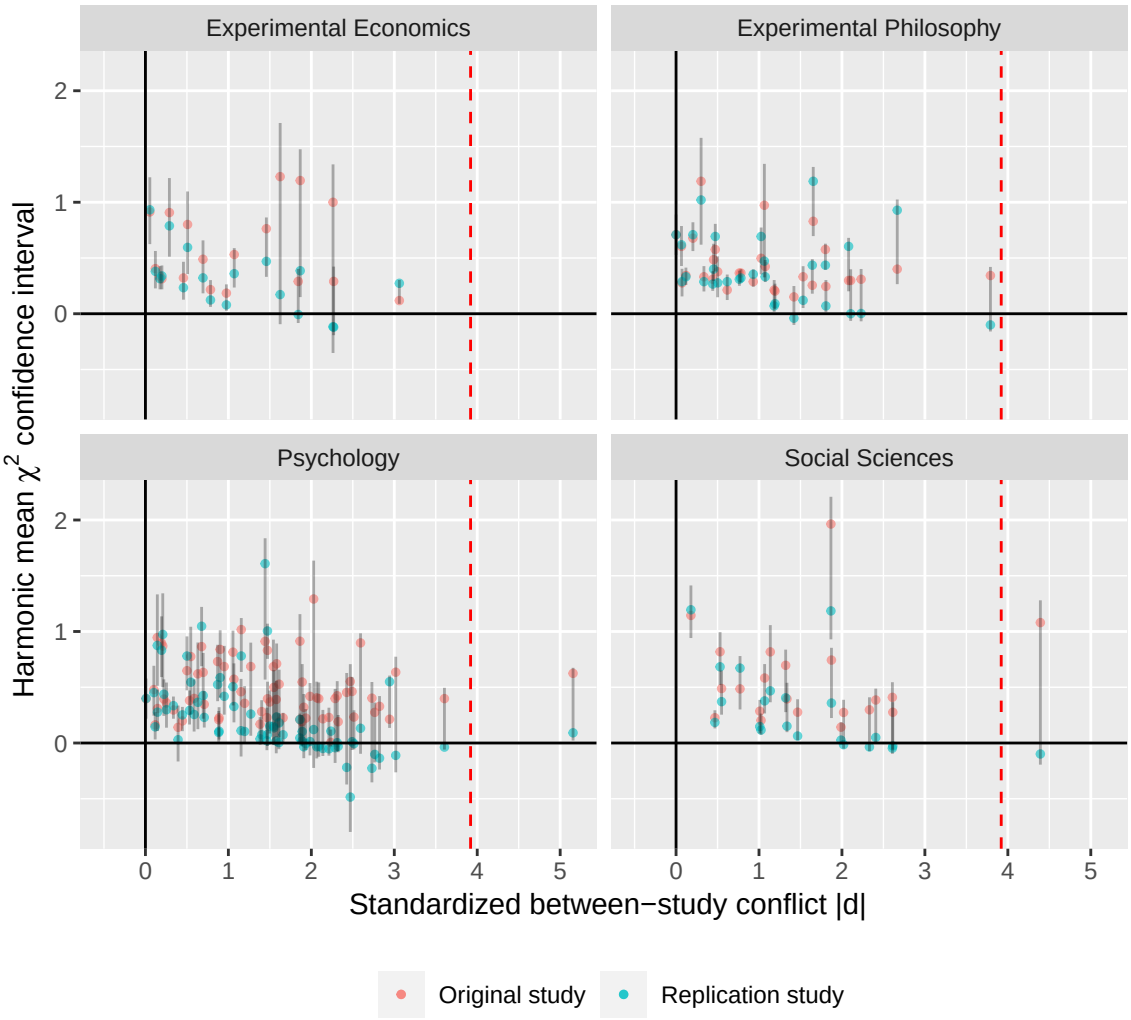

Figure 3.5: Harmonic mean χ^2 confidence interval for the replication projects. The vertical line in red represents $|d| = 2z_{\alpha/2}$, the threshold for the sceptical confidence sets under the equal sample size assumption.

When compared to the sceptical confidence intervals CI_S in Figure 3.2 and 3.3, there is no separation in harmonic mean χ^2 confidence intervals, which helps to avoid the replication paradox, but leads to the cost of information about the discrepancy between both studies at the same time.

3.2 Comparison of the different methods

In this section, results of the sceptical confidence interval and harmonic mean χ^2 confidence interval are compared with the meta-analysis, which is regarded as the gold standard submitted to the FDA. The question of which model of meta-analysis is more appropriate for the assessment of replication success has aroused as mentioned in Section 2.4.3. To answer this question, the three examples from Table 2.2, 2.3, and 2.5 are demonstrated at first. Furthermore, empirical results of different confidence intervals for the assessment are illustrated. Additionally, I will apply these different methods on some selected studies from these four replication projects mentioned in Section 2.6.

3.2.1 Three examples for the comparison

I illustrate these three examples from Chapter 2 in Figure 3.6. The ordinary confidence intervals for the original and replication studies are on the top. The confidence intervals for the replication success assessment are in the below, including calibrated sceptical confidence intervals in the two-sided version CI_S,

harmonic mean χ^2 confidence intervals at the nominal level $\alpha = 0.05$, and conventional 95% confidence intervals by meta-analysis with the fixed-effects model as well as the random-effects model.

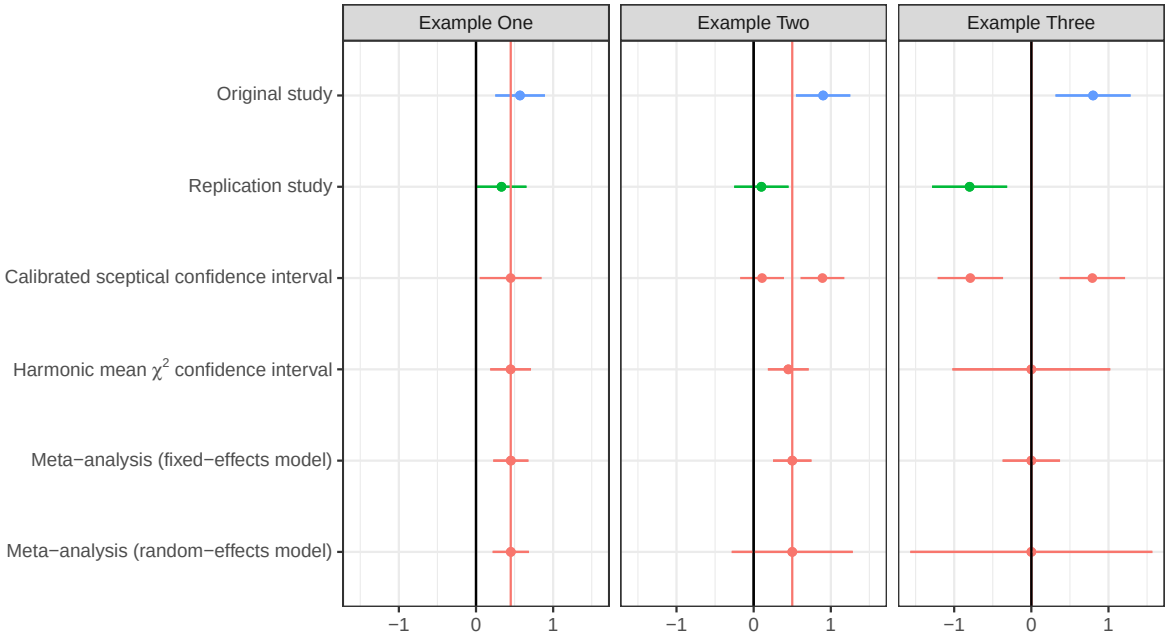

Figure 3.6: Different results of the replication success assessment for the *Three Examples* in Chapter 2

All of these confidence intervals for the replication success assessment center around the mean values of effect size estimates of the original and replication study approximately, since the assumption of an equal sample size holds (in *Example One*, $c \approx 1$). Additionally, the calibrated sceptical confidence sets in *Example One* and *Example Two* center around point estimates $\hat{\theta}_o$ and $\hat{\theta}_r$ of both two studies separately.

The calibrated sceptical confidence intervals are always wider than the harmonic mean χ^2 confidence intervals, even though the harmonic mean χ^2 confidence intervals are at the nominal level. Thus, the sceptical confidence interval is always a more stringent method in contrast to the harmonic mean χ^2 confidence interval for the replication success assessment. In terms of the meta-analysis results, the results from the *Example Two* and *Example Three* show that the assumption of homogeneity in the fixed-effects models is untenable. It is more appropriate to implement the random-effects model meta-analysis to take the heterogeneity across studies into account, especially when there is a distinguishable conflict between the original and replication study.

Due to the requirement of the direct replication mentioned in Chapter 1, the protocol of the replication study is required to be as closely as possible to that of the original study. Simons et al. (2014) points out that in the Reproducibility Project: Psychology (RPP), even though replication studies are reproduced in an extremely close way as in original studies, researcher might not be measuring exactly the same effect as studies from a variety of laboratories that all followed an identical in this project. This supports the use of the random-effects model meta-analysis for the replication success assessment.

3.2.2 Empirical results for the comparison

To explore the relationship among these different methods, I illustrate empirical results of corresponding assessment confidence intervals against varying standardized between-study conflicts d. The original study is fixed at $\hat{\theta}_o = 0.6$ as well as $\sigma_o = 0.1$, and the equal sample size assumption $c = 1$ holds to acquire empirical results.

For the aim of illustration, I show the results of sceptical confidence intervals CI_S and fixed-effects model meta-analysis firstly, then the results of harmonic mean χ^2 confidence intervals and random-effects model meta-analysis will be discussed.

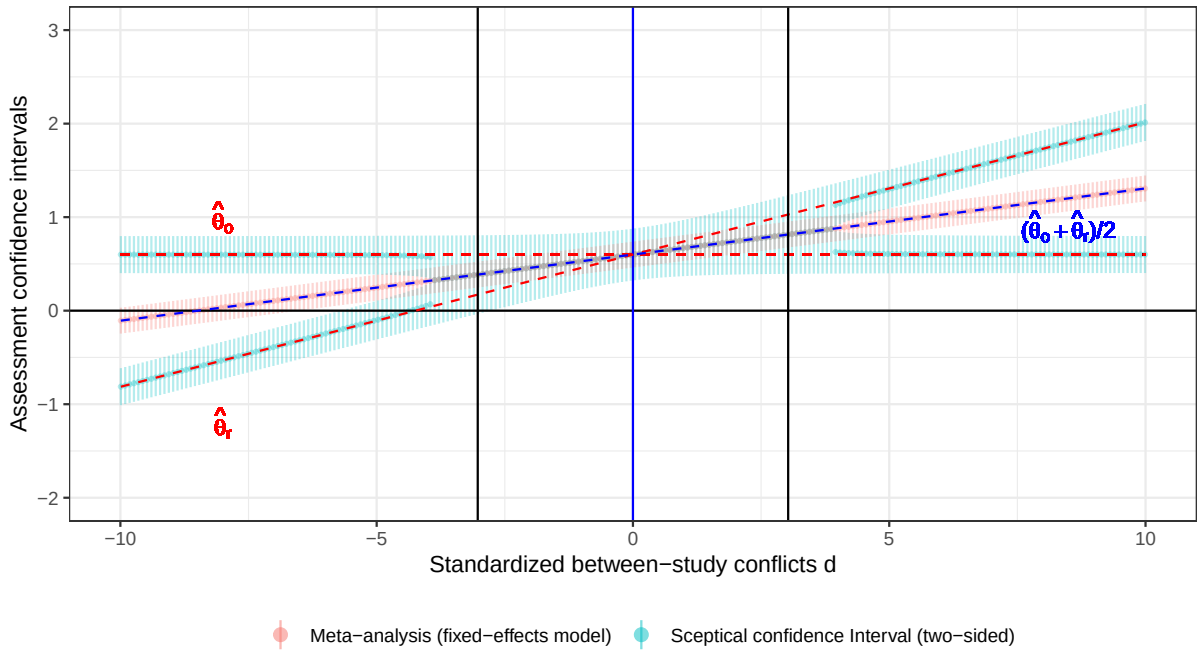

Figure 3.7: Empirical results of calibrated sceptical confidence intervals ($\alpha = 0.13$) and meta-analysis results with fixed-effects model ($\alpha = 0.05$)

In Figure 3.7, since the heterogeneity between the original and replication studies is not taken into account, different standardized between-study conflicts d have no effect on the width of the calculated confidence intervals of meta-analysis. Namely, the variance of the results of meta-analysis cannot reveal the between-study conflict. The overall effect estimates are just the average of the effect size estimates of both original and replication studies.

In terms of the method of sceptical confidence interval CI_S, the problem of replication paradox mentioned in Section 2.3.5 is unwanted. However, unusual sceptical confidence set could carry the information about possible discrepancies between these two studies. Specifically, suppose we have observed positive results of both original and replication studies, the sceptical confidence sets with two intervals centering around two effect size estimates could reflect how different the results from these two studies are, while the fixed-effects model meta-analysis could not.

Figure 3.8 shows the results of calibrated sceptical confidence intervals ($\widetilde{CI}_S$) and meta-analysis with the random-effects model. For the sceptical confidence interval in the one-sided version, we use the one-sided significance level at $\widetilde{\alpha} = 0.065$. While the results of meta-analysis are in the traditional two-sided significance level $\alpha = 0.05$. In both cases, confidence intervals are wider with larger absolute values of the standardized between-study conflicts $|d|$. That is we are less certain about the interval estimate of the overall assessment result if the original and replication studies differ quite a lot. Another interesting point is that meta-analysis results are more likely to be affected by the between-study conflicts. To be more precise, when the between-study conflict $|d|$ are small, we could obtain wider calibrated sceptical confidence interval. Reversely, meta-analysis with random-effects model will result in wider interval with larger standardized between-study conflict $|d|$. Thus meta-analysis is more likely to be affected by discrepancies between studies. In other words, with two distinguishable studies, we are less likely to make a claim of replication success by the meta-analysis.

The sceptical confidence intervals $\widetilde{CI}_S$ always have a lower limit slightly smaller than the smaller effect size estimate $\min\{\hat{\theta}_o, \hat{\theta}_r\}$, and an upper limit slightly larger than the larger effect size estimate $\max\{\hat{\theta}_o, \hat{\theta}_r\}$. Accordingly, when we observe two substantially positive results of both studies (on the right hand side of the Figure 3.8), the sceptical confidence interval will not cover $\mu = 0$, no matter how different these two studies are. However, it is not the same case regarding the random-effects model meta-analysis results. Large discrepancies will come to a claim of replication failure, as $\mu = 0$ will be covered in the calculated confidence intervals. This is the intrinsic difference between the method of sceptical confidence interval and the random-effects model meta-analysis. Namely, the random-effects model meta-analysis will treat large discrepancy between studies as the replication failure, while the

method of sceptical confidence interval will not, once both studies are significant enough.

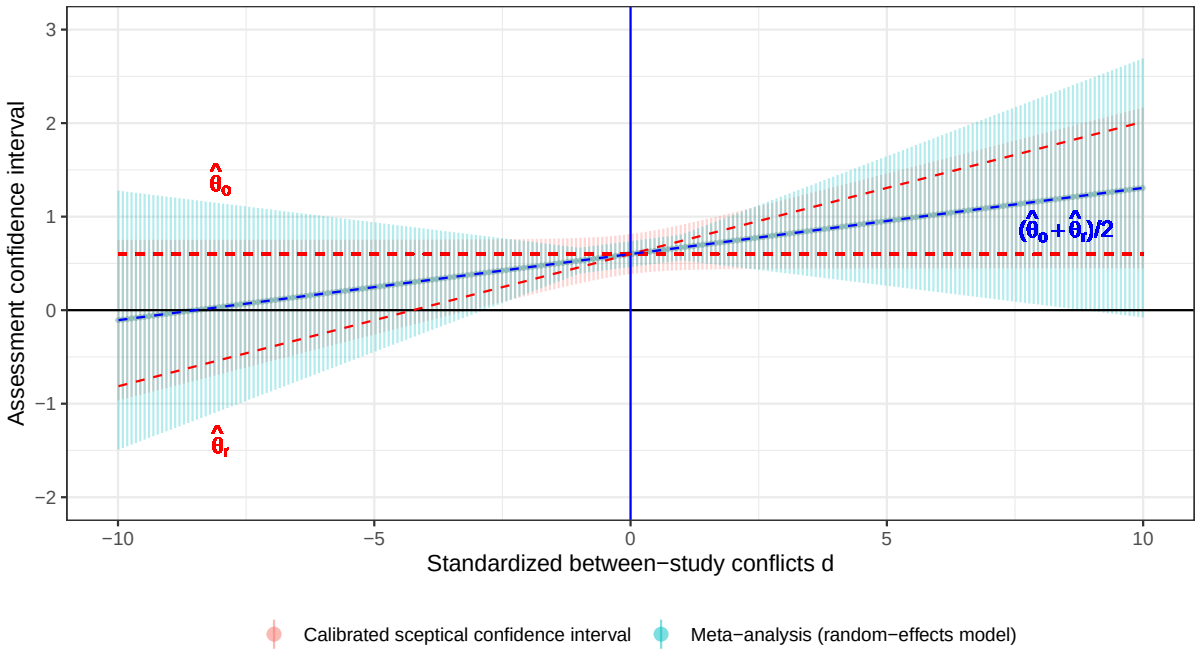

Figure 3.8: Empirical results of calibrated sceptical confidence intervals ($\tilde{\alpha} = 0.065$) and meta-analysis with random-effects model ($\alpha = 0.05$)

To summarize, the random-effects model meta-analysis is more appropriate than the fixed-effects model meta-analysis as the heterogeneity is taken into account in the former case, which is important especially when there is a great conflict between the original and replication study. The method of sceptical confidence intervals and random-effects model meta-analysis are similar in a particular sense. We are likely to come to a wider confidence intervals for the assessment if the standardized between-study conflict d is large. However, the response to various values of d is slightly different. The random-effects model meta-analysis is more likely to be affected by discrepancies between studies. And most of all, these two methods substantially differ in terms of the assessment for studies with large discrepancies. To be more specific, the random-effects model meta-analysis will not claim a replication success once there is a large discrepancy, no matter how significant the original and replication studies are. While the result of the sceptical confidence interval will be the opposite. Further research is required to explore the degree of between-study conflict which will come to different results of the claim of replication success under these two methods.

3.2.3 Replication studies with equal sample size for the comparison

The random-effects model meta-analysis treats studies as exchangeable, which means that the true effect of each study is considered to be a random measure from the population distribution. For the comparison of different methods, we assume that the assumption of exchangeability of the original study and replication study always holds.

In Figure 3.9, calibrated sceptical confidence intervals CI_S are listed for studies from these four replication projects. From the bottom to the top, the relative sample sizes c vary from small values to large values. The whole areas are divided into three categories, studies with small relative sample size $c < 1$ are on the bottom, studies with equal sample size $c = 1$ are in the middle (only in the case of *Experimental Philosophy* and *Psychology*), and studies with large relative sample size $c > 1$ are on the top. Unusual sceptical confidence sets are also illustrated separately.

To compare different assessment confidence intervals, I select the studies with an equal sample size $c = 1$, and list them in Table 3.2. There are 1 case from the project of *Experimental Philosophy* and 9 cases from the project of *Psychology*, 10 in total, satisfying the assumption of an equal sample size.

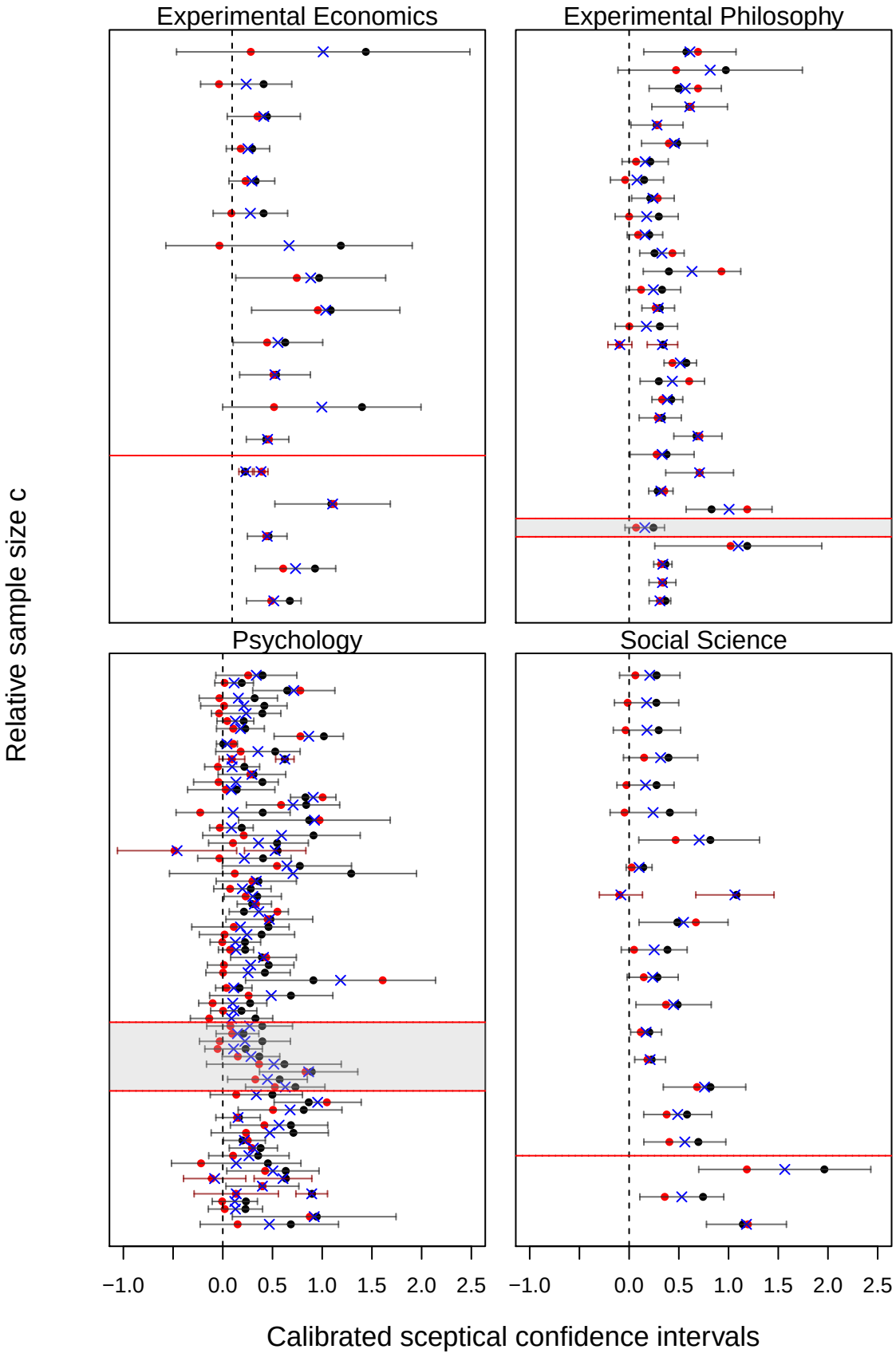

Figure 3.9: Calibrated sceptical confidence intervals for the replication projects. Original effect estimates $\hat{\theta}_o$ are in black and replication effect estimates $\hat{\theta}_r$ are in red. Calibrated sceptical confidence sets are expressed by red arrows. Crosses in blue stand for mean values of confidence limits. All studies are divided into three categories according to the relative sample size.

Table 3.2: Cases of studies under the assumption of an equal sample size

Study	Project	t_o	t_r	c	$\lvert d \rvert$
AL Morris, ML Still	Psychology	3.33	1.10	1.00	1.58
S Farrell	Psychology	3.53	2.02	1.00	1.07
S Pacton, P Perruchet	Psychology	3.90	3.63	1.00	0.19
NO Rule, N Ambady	Psychology	1.96	1.75	1.00	0.15
PW Eastwick, EJ Finkel	Psychology	0.70	0.15	1.00	0.39
C McKinstry, R Dale, MJ Spivey	Psychology	2.46	2.75	1.00	0.21
DR Addis, AT Wong, DL Schacter	Psychology	3.49	4.20	1.00	0.50
SE Palmer, T Ghose	Psychology	3.16	0.30	1.00	2.03
V Lobue, JS DeLoache	Psychology	3.46	1.18	1.00	1.61
Alicke et al.	Experimental Philosophy	3.58	1.03	1.00	1.81

Note: t_o and t_r are the test statistics for the original and the replication study.

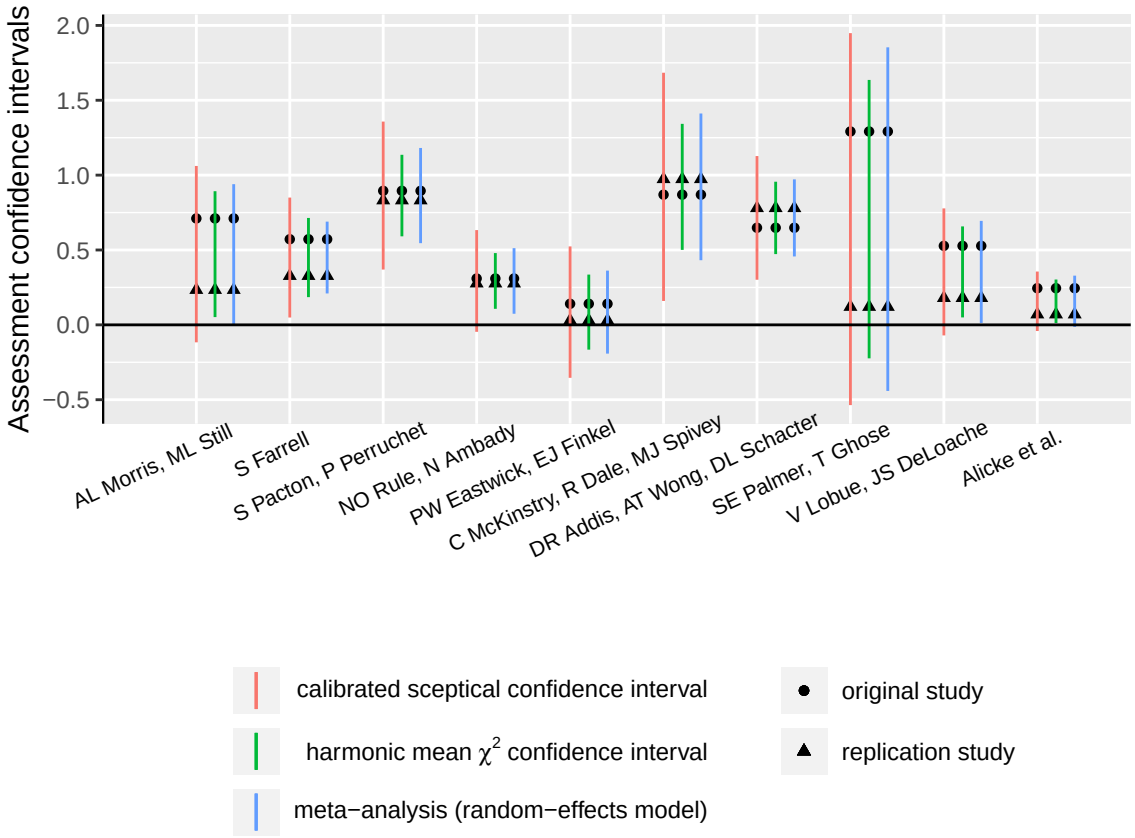

Figure 3.10: Assessment confidence intervals for the 10 selected studies

In Figure 3.10, calibrated sceptical confidence interval CI_S at $\alpha = 0.13$, harmonic mean χ^2 confidence intervals at $\alpha = 0.05$ and meta-analysis (random-effects model) results at $\alpha = 0.05$ are demonstrated for these 10 studies holding the equal sample size assumption, and the assumption of exchangeability is assumed to hold as well. Effect size estimates of both original and replication studies are expressed in the form of correlation coefficients r mentioned in Section 2.6.

Calculated confidence intervals for the assessment of replication success vary under same pairs of studies. Thus the resulted claim of replication success may be different using different criteria. For instance, for the study *'AL Morris, ML Still'*, the sceptical confidence interval will result in a declare of replication failure, while the harmonic mean χ^2 confidence interval and random-effects model meta-analysis will come to a claim of replication success. The most stringent calibrated sceptical confidence intervals are always the widest, which means that the between-study conflicts of these studies are not large enough, otherwise wider confidence intervals of meta-analysis results would be observed, see Section 3.2.2. Results from the random-effects model meta-analysis are mostly wider than harmonic mean χ^2 confidence interval results. There is only one special case *'S Farrell'*, where the relationship between meta-analysis results and sceptical confidence interval is the opposite. It is unclear which of these criteria is more appropriate for the assessment, and we are expected to come to different results under various criteria.

Chapter 4

Discussion

Reproducibility is a feature of science, which requires reliable assessment approaches for a better insight of evidence. This thesis provides a new perspective to evaluate the research findings.

Firstly, methods developed from the sceptical p-value (Held, 2020b), namely the method of sceptical p-value function, and the method of sceptical confidence interval, are theoretically well-established via the analysis of credibility and check of the prior-data conflicts. The sceptical p-value function helps to investigate the results of sceptical p-values systematically, from which we can observe sceptical confidence intervals at various confidence levels. The discussion about the unusual sceptical confidence set aroused concern about the direction conflict, the between-study conflict as well as the replication paradox.

Secondly, the harmonic mean χ^2 test (Held, 2020a) is an attractive substitution for the assessment of replication success. The primary product, the harmonic mean χ^2 p-value function, and the harmonic mean χ^2 confidence interval help to assess the replication success in the same way as the methods based on the sceptical p-value. That is, if the corresponding p-value is smaller than the significance level or the point estimate $\mu = 0$ is not included in the calculated confidence interval, a claim of replication success can be declared. The harmonic mean χ^2 p-value function and the harmonic mean χ^2 confidence interval can address the problem of the replication paradox, and have a nice interpretation of the result. What is more, the most compelling feature of this method is its application in the situation of multiple replications.

These two kinds of methods mentioned above evaluate the replication success analogously. However, the starting point and the explanation are quite different. Methods based on the sceptical p-value provide a list of mean values of the sufficiently sceptical prior, which connects evidence from the original and the replication study. Approaches originated from the harmonic mean χ^2 test offers us a plausible range of point estimates of effect size, based on which the test statistics are integrated, and then comes to a new statistic under the χ^2 distribution.

4.1 Sceptical p-value function & sceptical confidence interval

Based on the theoretically well-founded sceptical p-value (Held, 2020b), the result of the original study was transformed into a sufficiently sceptical prior with the mean value $\mu = 0$, based on which the evidence from the replication study is then examined.

The sceptical p-value function works as an integration for sceptical p-values under varying μ values. The sceptical confidence interval can then be obtained at any level from the sceptical p-value function. The sceptical p-value function for two-sided sceptical p-values is a bimodal function, which may give rise to an unusual sceptical confidence set. The one-sided version of the sceptical p-value function avoids unrealistically small sceptical p-values when there is a direction conflict, but has a few limitations for implementation.

In contrast to the sceptical p-value (both p_S and $\widetilde{p}_S$), the sceptical confidence interval demonstrates the range of plausible value of mean values μ of the sufficiently sceptical prior, which reveals the range of effect size estimates. Additionally, the direction and strength of the demonstrated effects are also accessible via a confidence interval. The sceptical confidence set under a large discrepancy between studies may result in the unwanted replication paradox. To substantiate the direction of an original study, the one-sided sceptical p-value is recommended. In that case, there would be no sceptical confidence set, thus no replication paradox.

4.2 Harmonic mean χ^2 p-value function & harmonic mean χ^2 confidence interval

Methods under the harmonic mean χ^2 test solve the problem of replication paradox, via an adjustment under the direction conflict. To be specific, we set an upper bound for the harmonic mean χ^2 p-values, according to the number of investigated studies. Thus an inequality, instead of an unrealistically small p-value will be reported.

The harmonic mean χ^2 p-value function assesses p_H values under different μ values and CI_H under varying significance levels. With the help of the adjustment, we acquire a function curve that opens upward, where multiple cusps at individual effect size estimates are also obtainable. The upper bound for the harmonic mean χ^2 p-value function is at $1/2^{n-1}$, where n stands for the number of studies.

The harmonic mean χ^2 confidence interval have only an upper and a lower limit, no separation or disjoint intervals. Another appealing property of the harmonic mean χ^2 confidence interval is that it can be easily extended to a setting with multiple replications based on the harmonic mean of squared statistics.

4.3 Conclusion

Reproducibility is exactly the beating heart of science, and it has come under scrutiny recently. In this master thesis, multiple alternatives with their upsides and downsides are introduced in detail. Researchers should choose an appropriate one under a particular circumstance. Researches about quantitative measures of the replication success will never stop.

In contrast to the standard two-trials rule, methods based on the sceptical p-value may come to different results for drug approval under two primary studies. Both the sceptical confidence interval and the meta-analysis with a random-effects model take the discrepancy between two studies into account. The second approach is more responsive to the between-study conflicts. The fundamental difference between these two methods is the theoretical groundwork behind them. What is more, the random-effects model meta-analysis will declare a replication failure if there is a substantial difference between studies. While the sceptical confidence interval will come to the opposite conclusion once both studies are significant enough. This underlying difference requires further research about the appropriate choice of the approach under different scenarios.

Maxwell et al. (2015) indicates that designing replication with adequate power encounters several primary difficulties. We can consider the equivalence test and multiple replications for the sake of power. To assess the replication success of an original study with adequate power, multiple replications help to get rid of the problem of inadequate sample size and single replication (Klein et al., 2014; Ebersole et al., 2016; Klein et al., 2018). Applications of these approaches for the assessment of replication success under various scenarios require further research.

Appendix A

Appendix

A.1 The threshold for unusual sceptical confidence set

The threshold for the unusual sceptical confidence set in the form of the standardized between-study conflict d (see Equation 2.8) depends only on the significance level α as discussed in Section 2.3.4. To this end, we can take the derivation of Equation 2.6 when the relative sample size is $c = 1$. The test statistics are $t_o = (\hat{\theta}_o - \mu)/\sigma$ and $t_r = (\hat{\theta}_r - \mu)/\sigma$, and the derivation goes as follows,

$$(t_o^2/z_S^2 - 1)(t_r^2/z_S^2 - 1) = 1,$$

for which, the derivation with respect to μ is

$$\frac{dz_S^2}{d\mu} = \frac{1}{\sigma^2}(\hat{\theta}_o + \hat{\theta}_r)\left\{\frac{1}{(\frac{\hat{\theta}_o-\mu}{\hat{\theta}_r-\mu} + \frac{\hat{\theta}_r-\mu}{\hat{\theta}_o-\mu})} - 1\right\}.$$

The solution for $dz_S^2/d\mu = 0$ is then

$$\mu = \frac{\hat{\theta}_o + \hat{\theta}_r}{2}.$$

Thus, the corresponding z_S^2 should be

$$z_S^2 = \left(\frac{\hat{\theta}_o - \hat{\theta}_r}{2\sqrt{\sigma_o^2 + \sigma_r^2}}\right)^2.$$

The threshold can be expressed in the form of the standardized between-study conflict d defined in Equation 2.8,

$$d = \frac{\hat{\theta}_o - \hat{\theta}_r}{\sqrt{\sigma_o^2 + \sigma_r^2}} = \pm 2z_{\alpha/2}.$$

In summary, the standardized between-study conflict d tells whether we could obtain a sceptical confidence interval or a sceptical confidence set. We could expect a sceptical confidence set, with two disjoint intervals, if there is a great conflict between the original and the replication study, say $|d| > z_{\alpha/2}$.

Even though this threshold only works under the equal sample size assumption, say the relative sample size $c = 1$. We are able to have some initial inferences under the more general unequal sample size assumption. Normally, we are more likely to come to an unusual sceptical confidence set under a relatively small relative sample size (i.e., $c<1$) as discussed in Section 2.3.6.

A.2 The narrowest sceptical confidence interval

In Figure 2.10, the equal sample size assumption holds, say, the shared standard error is $\sigma = \sigma_o = \sigma_r$. The narrowest sceptical confidence interval can be obtained at $t_o = t_r$, indicating a perfect replication. It corresponds to the solutions of μ for study pair A and study pair C as shown in Table A.1.

Table A.1: *Test statistics of these two study pairs*

study pair	t_o	t_r
A	$(\hat{\theta}_o - \mu_A)/\sigma$	$(\hat{\theta}_r - \mu_A)/\sigma$
C	$-(\hat{\theta}_o - \mu_C)/\sigma$	$-(\hat{\theta}_r - \mu_c)/\sigma$

In Table A.1, μ_A and μ_C are the confidence limits of the corresponding sceptical confidence interval at $t_o = t_r$. The distance between point A and point C in Figure 2.10 should be

$$d_{AC} = \sqrt{2}\frac{|\mu_A - \mu_c|}{\sigma},$$

since we assume that effect size estimates $\hat{\theta}_o$ and $\hat{\theta}_r$ are the same for study pair A and study pair C. The distance $d_{AC} = 4z_{\alpha/2}$ is obtainable intuitively, as we know the Cartesian coordinates of both two pairs of studies in the form of $z_{\alpha/2}$, say $A(\sqrt{2}z_{\alpha/2}, \sqrt{2}z_{\alpha/2})$ and $C(-\sqrt{2}z_{\alpha/2}, -\sqrt{2}z_{\alpha/2})$, according to Table 2.4. Therefore, the range of sceptical confidence interval is

$$|\mu_A - \mu_C| = 2\sqrt{2}z_{\alpha/2} \cdot \sigma,$$

which shows that we are likely to obtain a wider sceptical confidence interval with a larger shared standard error σ, if effect size estimates $\hat{\theta}_o$ and $\hat{\theta}_r$ are fixed.

A.3 Dominator of the requirements for replication paradox

There are two necessary but not sufficient requirements for the replication paradox, see Section 2.3.5. Namely, only if both the direction conflict and the unusual sceptical confidence set occur simultaneously, we may observe a replication paradox. In what follows, I will show that which requirement is the dominant one will depend on the test statistics of the original study t_o.

As demonstrated in Section A.1, the threshold in the form of standardized between-study conflict for an unusual sceptical confidence set is

$$d = \pm 2z_{\alpha/2}.$$

The boundary for the direction conflict is observed at $\hat{\theta}_r = 0$, where $\hat{\theta}_r$ is the effect size estimate of the replication study. This boundary can also be expressed in the form of the standardized between-study conflict,

$$d = \frac{\hat{\theta}_o}{\sqrt{2}\sigma},$$

where $\sigma = \sigma_o = \sigma_r$ holds under the equal sample size assumption.

Thus, the threshold and the boundary would be located at the same between-study conflict d if

$$\frac{\hat{\theta}_o}{\sqrt{2}\sigma} = \pm 2z_{\alpha/2}.$$

Consequently, an extremely significant result of the original study, say $|t_o| > 2\sqrt{2}z_{\alpha/2}$ will result in a boundary for the direction conflict more extreme than $2z_{\alpha/2}$. In such a case, we may come to a sceptical confidence set even there is no direction conflict. On the other hand, with a merely significant original study $|t_o| < 2\sqrt{2}z_{\alpha/2}$, we are supposed to obtain a sceptical confidence interval only if the direction conflict exists.

A.4 One-sided study-specific p_i for overall significance level

As mentioned in Section 2.5, the overall Type-I error concerning the assessment of replication success in the setting of a single replication should be $\widetilde{\alpha} = 1/1600$ (one-sided). The required study-specific p-value p_i is accessible based on the bound of critical values of the test statistics of the harmonic mean χ^2 test method.

The derivation of the bound for critical value defined in Equation 2.13 is taken from Held (2020a). There is a slight difference between the unweighted χ^2_ω and weighted test statistics χ^2, as defined in Equation 2.12 and 2.15. In what follows, I focus on the unweighted test statistic χ^2.

For the harmonic mean χ^2 test method, suppose $Z_i = z_i$ is the observed test statistics in the i-th study and assume that $z_i > 0$ for all i=1, ..., n, which means that all effects are in the same direction. The upper bound for the harmonic mean $z^2_H = n/\sum_{i=1}^n 1/z_i^2$ is as follows,

$$z^2_H \leq n z^2_{\text{min}} \leq n z_i^2, \tag{A.1}$$

where $z^2_{\text{min}} = \min\{z_1^2, ..., z_n^2\}$ and n is the number of studies. This inequality implies $x^2 \leq n^2 z_i^2$ for the test statistic x^2 we observed, where x is from Equation 2.13. Thus we obtain

$$\Pr\{\chi^2(1) \geq n^2 z_i^2\}/2^n \leq \widetilde{p}_H,$$

for the i-th study. The inequality $\widetilde{p}_H \leq \alpha_H$ is required for a claim of success at level α_H, thus $\Pr\{\chi^2(1) \geq n^2 z_i^2\}/2^n \leq \alpha_H$ must hold. Namely, $z_i \geq \sqrt{c_H}/n$ must hold. For this reason, the restriction on individual p_i is

$$p_i = 1 - \Phi(\sqrt{c_H}/n), \tag{A.2}$$

where c_H is the critical value for z^2_H with respect to α_H mentioned in Equation 2.14. This restriction on the individual p-values of the original and replication study is necessary but not sufficient for a claim of success. And the sufficient restriction bounds is

$$p_i = 1 - \Phi(\sqrt{c_H}/\sqrt{n}), \tag{A.3}$$

which is derived from the condition $\chi^2 = n z_i^2 \geq c_H$, equally $z_i \geq \sqrt{c_H}/\sqrt{n}$.

Under the one-sided 2.5% significance level $\alpha_H = 0.025^2$, the critical value is $c_H = 9.14$. The corresponding necessary and sufficient restriction for an individual p_i are 0.065 and 0.016 respectively. For the method of sceptical p-value, since the assumption of equal sample size could not always hold, the study-specific p_i based on the bound of critical value is not obtainable when the test statistics z_S^2 cannot be expressed by the harmonic mean of squared test statistics of individual studies. Thus, we just borrow the calculated p_i from the harmonic mean χ^2 test regarding to the same type-I error control. This necessary bounded one-sided individual $p_i = 0.065$ is where the calibrated level of α for sceptical p-value method comes from in Section 2.3.1.

Besides, the 'liberal' version based on the one-sided study-specific p-value at $p_i = 0.025$ and the corresponding critical value $c_H = 7.68$ could also be obtained from Equation A.3. In practice, we choose the liberal bound as the nominal level ($\widetilde{\alpha} = 0.025$) for the assessment of replication success, and the necessary bound as the calibrated level ($\widetilde{\alpha} = 0.065$) regarding to the type-I error control $\alpha = 1/1600$ in the setting of two investigated studies.

A.5 R code for functions

```
# ================================================================================
#  Sceptical $p$-value function with respect to mu
# ================================================================================
pScepValueFun <- function(thetao,thetar,
                          sigmao,sigmar,
                          alternative = "two.sided",...){
  require(ReplicationSuccess)
  mu_grid <-  seq(-5*abs(min(thetao,thetar)),
                  5*abs(max(thetao,thetar)),
                 length=1000)
  p_grid <- NULL
  for (i in 1:length(mu_grid)){
   p_grid[i] <- pSceptical(zo=(thetao-mu_grid[i])/sigmao,
                           zr=(thetar-mu_grid[i])/sigmar,
                           c=sigmao^2/sigmar^2,
                           alternative = alternative)
  }
  dat <- data.frame(cbind(mu=mu_grid,ps=p_grid))

  fig <- plot(dat$mu,dat$ps,...)
  return(fig)
}
# ================================================================================
#  Harmonic mean chi-squared $p$-value function with respect to mu
# ================================================================================
pHmeanValueFun <- function(thetahat, se,
                           w = rep(1, length(thetahat)),...){
  mu_grid <-  seq(-5*abs(min(thetahat)),
                 5*abs(max(thetahat)),
                 length=10000)
  n <- length(thetahat)
  p_grid <- NULL
  for (i in 1:length(mu_grid)){
    z <- (thetahat - mu_grid[i])/se
    zH2 <- sum(sqrt(w))^2/sum(w/z^2)
    res <- pchisq(zH2, df = 1, lower.tail = FALSE)
    p_grid[i] <- res/(2^(n-1))
  }
  dat <- data.frame(cbind(mu=mu_grid,ps=p_grid))
  plot(dat$mu,dat$ps,...)
  # real line for the cases without direction paradox
  points(x=dat[which(dat$mu<min(estimate) | dat$mu>max(estimate) ),"mu"],
         y=dat[which(dat$mu<min(estimate) | dat$mu>max(estimate)),]$ps,cex=0.4,
         col="black")
}
# ================================================================================
# Function to find out the roots of sceptical confidence interval or sets
# ================================================================================
rootSceptical <- function(thetao, thetar,
                          sigmao, sigmar,
                          c = 1,  alpha, alternative="two.sided"){
  require(rootSolve)
  require(ReplicationSuccess)
  stopifnot((alpha > 0) | (alpha < 1))
  stopifnot(min(sigmao,sigmar) > 0)
```

```
  stopifnot((!is.na(thetao)&&(!is.na(thetar))))
  f <- function(mu){
    zo <- (thetao-mu)/sigmao
    zr <- (thetar-mu)/sigmar
    ps <- pSceptical(zo = zo, zr = zr, c = c,
                     alternative = alternative)
    threshold <- thresholdSceptical(level = alpha/2,
                                    alternative = alternative)
    return(ps-threshold)
  }
  res <- uniroot.all(f,interval = c(-5,5))
  # this arbitrary value for interval is resonable
  # as we want to apply the function on fisher-z-transformed data
  return(res)
}
# ============================================================================
#  Plot function to show conventional confidence interval and sceptical CIs
# ============================================================================
plotciSceptical <- function(thetao, thetar,sigmao, sigmar,
                            alternative = "two.sided",ylab="Effect Size",
                            c = 1,alpha=0.05,...){
  require(plotrix)
  par(cex.axis=0.7,las=1)
  upperlimit <- c(thetao+sigmao*p2z(alpha,"two.sided"),
                  thetar+sigmar*p2z(alpha,"two.sided"))
  lowerlimit <- c(thetao-sigmao*p2z(alpha,"two.sided"),
                  thetar-sigmar*p2z(alpha,"two.sided"))
  mean <- c(mean(c(upperlimit[1],lowerlimit[1])),
            mean(c(upperlimit[2],lowerlimit[2])))
  df_study = data.frame(cbind(upperlimit,lowerlimit,mean))
  studies <-c("Original Study","Replication Study",
              "Sceptical (Nominal)","Sceptical (Calibrated)")
  # set ylim for final plot
  if(is.null(ylim)){
    ylims <- c(-1.5*max(abs(thetao),abs(thetar)),1.5*max(abs(thetao),abs(thetar)))
  }else{
    ylims <- ylim
  }
  # set ylim for final plot
  plot(1:2,df_study$mean,main=main,cex.main=cex.main,
       ylim=ylims,cex.axis=cex.axis,cex.lab=cex.lab,
       xlim = xlim, xaxt="n",ylab=ylab,xlab="")
  plotCI(1:2,y=df_study$mean,
         uiw=df_study$upperlimit-df_study$mean,
         liw=df_study$mean-df_study$lowerlimit,
         pch=20,pt.bg=par("bg"), add=TRUE)
  xtick <- seq(1, 4, by=1)
  axis(side=1, at=xtick, labels = studies,cex.axis=cex.axis)
  Sceptical_Nominal <- rootSceptical(thetao, thetar, sigmao, sigmar,
                                     c = c, alpha=0.05,
                                     alternative = alternative)
  Sceptical_Calibrated <- rootSceptical(thetao, thetar, sigmao, sigmar,
                                        c = c, alpha=0.13,
                                        alternative = alternative)
  if(length(Sceptical_Nominal)==2){
    ##### for the case of two roots for mu
    df_sceptical = data.frame(rbind(
```

```
      c(max(Sceptical_Nominal),min(Sceptical_Nominal),mean(Sceptical_Nominal))))
    colnames(df_sceptical) <- colnames(df_study)
    plotCI(3,y=df_sceptical$mean,
           uiw=df_sceptical$upperlimit-df_sceptical$mean,
           liw=df_sceptical$mean-df_sceptical$lowerlimit,
           err="y",add=TRUE)
  }else{
    #### for the case of four roots for mu
    df_sceptical <- data.frame(rbind(matrix(Sceptical_Nominal,ncol=4)))
    colnames(df_sceptical) <- c("l1","u1","l2","u2")
    df_sceptical$mean1 <- (df_sceptical$l1+df_sceptical$u1)/2
    df_sceptical$mean2 <- (df_sceptical$l2+df_sceptical$u2)/2
    df = data.frame(rbind(
      c(max(Sceptical_Nominal),min(Sceptical_Nominal),mean(Sceptical_Nominal))))
    plotCI(3,y=df_sceptical$mean1,
           liw=df_sceptical$mean1-df_sceptical$l1,
           uiw=df_sceptical$u1-df_sceptical$mean1,
           add=TRUE)
    plotCI(3,y=df_sceptical$mean2,
           liw=df_sceptical$mean2-df_sceptical$l2,
           uiw=df_sceptical$u2-df_sceptical$mean2,
           add=TRUE)
  }
  if(length(Sceptical_Calibrated)==2){
    ##### for the case of two roots for mu
    df_sceptical = data.frame(rbind(
      c(max(Sceptical_Calibrated),min(Sceptical_Calibrated),mean(Sceptical_Calibrated))))
    colnames(df_sceptical) <- colnames(df_study)
    plotCI(4,y=df_sceptical$mean,
           uiw=df_sceptical$upperlimit-df_sceptical$mean,
           liw=df_sceptical$mean-df_sceptical$lowerlimit,
           err="y",add=TRUE)
  }else{
    #### for the case of four roots for mu
    df_sceptical <- data.frame(rbind(matrix(Sceptical_Calibrated,ncol=4)))
    colnames(df_sceptical) <- c("l1","u1","l2","u2")
    df_sceptical$mean1 <- (df_sceptical$l1+df_sceptical$u1)/2
    df_sceptical$mean2 <- (df_sceptical$l2+df_sceptical$u2)/2
    df = data.frame(rbind(
      c(max(Sceptical_Calibrated),
        min(Sceptical_Calibrated),
        mean(Sceptical_Calibrated))))
    plotCI(4,y=df_sceptical$mean1,
           liw=df_sceptical$mean1-df_sceptical$l1,
           uiw=df_sceptical$u1-df_sceptical$mean1,
           add=TRUE)
    plotCI(4,y=df_sceptical$mean2,
           liw=df_sceptical$mean2-df_sceptical$l2,
           uiw=df_sceptical$u2-df_sceptical$mean2,
           add=TRUE)
  }
}
```

A.6 R Session Information

```
## R version 3.6.3 (2020-02-29)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 10 x64 (build 18362)
##
## Matrix products: default
##
## locale:
## [1] LC_COLLATE=English_United Kingdom.1252
## [2] LC_CTYPE=English_United Kingdom.1252
## [3] LC_MONETARY=English_United Kingdom.1252
## [4] LC_NUMERIC=C
## [5] LC_TIME=English_United Kingdom.1252
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
##  [1] ggrepel_0.8.2            dplyr_0.8.5
##  [3] rootSolve_1.8.2          meta_4.11-0
##  [5] xtable_1.8-4             ggpubr_0.2.5
##  [7] magrittr_1.5             ReplicationSuccess_0.1-2
##  [9] ggplot2_3.3.0            plotrix_3.7-7
## [11] latex2exp_0.4.0          knitr_1.28
##
## loaded via a namespace (and not attached):
##  [1] Rcpp_1.0.3         plyr_1.8.6         nloptr_1.2.2.1
##  [4] pillar_1.4.3       compiler_3.6.3     highr_0.8
##  [7] tools_3.6.3        digest_0.6.25      boot_1.3-24
## [10] lme4_1.1-21        nlme_3.1-145       evaluate_0.14
## [13] lifecycle_0.2.0    tibble_2.1.3       gtable_0.3.0
## [16] lattice_0.20-40    pkgconfig_2.0.3    rlang_0.4.5
## [19] Matrix_1.2-18      CompQuadForm_1.4.3 xfun_0.12
## [22] metafor_2.1-0      withr_2.1.2        stringr_1.4.0
## [25] cowplot_1.0.0      grid_3.6.3         tidyselect_1.0.0
## [28] glue_1.3.2         R6_2.4.1           minqa_1.2.4
## [31] farver_2.0.3       purrr_0.3.3        codetools_0.2-16
## [34] MASS_7.3-51.5      scales_1.1.0       splines_3.6.3
## [37] assertthat_0.2.1   colorspace_1.4-1   ggsignif_0.6.0
## [40] labeling_0.3       stringi_1.4.6      munsell_0.5.0
## [43] crayon_1.3.4
```

Bibliography

Anderson, S. F. and Maxwell, S. E. (2016). There's more than one way to conduct a replication study: Beyond statistical significance. *Psychological Methods*, 21(1):1. 2

Austin, P. C. (2009). Using the standardized difference to compare the prevalence of a binary variable between two groups in observational research. *Communications in statistics-simulation and computation*, 38(6):1228–1234. 16

Bailar, J. C. and Mosteller, F. (1988). Guidelines for statistical reporting in articles for medical journals: Amplifications and explanations. *Annals of internal medicine*, 108(2):266–273. 13

Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. *Nature News*, 533(7604):452. 2

Balduzzi, S., Rucker, G., and Schwarzer, G. (2019). How to perform a meta-analysis with R: A practical tutorial. *Evidence Based Mental Health*, 22:ebmental–2019. 37

Bender, R., Berg, G., and Zeeb, H. (2005). Tutorial: Using confidence curves in medical research. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, 47(2):237–247. 13

Benjamin, D., Berger, J., Johannesson, M., Nosek, B., Wagenmakers, E.-J., Berk, R., Bollen, K., Brembs, B., Brown, L., Camerer, C., Cesarini, D., Chambers, C., Clyde, M., Cook, T., De Boeck, P., Dienes, Z., Dreber, A., Easwaran, K., Efferson, C., and Johnson, V. (2017). Redefine statistical significance. 2(1):6. 2, 6

Berrar, D. (2017). Confidence curves: An alternative to null hypothesis significance testing for the comparison of classifiers. *Machine Learning*, 106(6):911–949. 13

Birnbaum, A. (1961). Confidence curves: An omnibus technique for estimation and testing statistical hypotheses. *Journal of the American Statistical Association*, 56(294):246–249. 13

Blaker, H. (2000). Confidence curves and improved exact confidence intervals for discrete distributions. *Canadian Journal of Statistics*, 28(4):783–798. 13

Blaker, H. and Spjøtvoll, E. (2000). Paradoxes and improvements in interval estimation. *The American Statistician*, 54(4):242–247. 13

Bland, J. M. and Altman, D. G. (1998). Bayesians and frequentists. *BMJ*, 317(7166):1151–1160. 6

Bobka, M. S. (1993). The 21 CFR Online Database: Food and Drug Administration Regulations Full-Text. *Medical reference services quarterly*, 12(1):7–15. 34

Bolles, R. and Messick, S. (1958). Statistical utility in experimental inference. *Psychological Reports*, 4(3):223–227. 13

Borenstein, M., Hedges, L. V., Higgins, J. P., and Rothstein, H. R. (2010). A basic introduction to fixed-effect and random-effects models for meta-analysis. *Research synthesis methods*, 1(2):97–111. 33

Box, G. E. (1980). Sampling and Bayes' inference in scientific modelling and robustness. *Journal of the Royal Statistical Society: Series A (General)*, 143(4):383–404. 7, 8, 10

Bracey, G. W. (1991). Sense, non-sense, and statistics. *Phi Delta Kappan*, 73(4):335–335. 6

Camerer, C. F., Dreber, A., Forsell, E., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Almenberg, J., Altmejd, A., Chan, T., et al. (2016). Evaluating replicability of laboratory experiments in economics. *Science*, 351(6280):1433–1436. 2, 37

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Nave, G., Nosek, B. A., Pfeiffer, T., et al. (2018). Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour*, 2(9):637–644. 2, 37

Carey, B. (2015). Psychology's fears confirmed: Rechecked studies don't hold up. *New York Times*, 27:A1. 2

Cohen, J. (1992). Things I have learned (so far). In *Annual Convention of the American Psychological Association, 98th, Aug, 1990, Boston, MA, US; Presented at the aforementioned conference.* American Psychological Association. 13

Cohen, J. (1994). The earth is round ($p <$. 05). *American psychologist*, 49(12):997. 6, 13

Cohen, J. (2013). *Statistical power analysis for the behavioral sciences.* Academic press. 16

Colquhoun, D. (2017). The reproducibility of research and the misinterpretation of p-values. *Royal society open science*, 4(12):171085. 8

Colquhoun, D. (2019). The false positive risk: A proposal concerning what to do about p-values. *The American Statistician*, 73(sup1):192–201. 8

Cova, F., Strickland, B., Abatista, A., Allard, A., Andow, J., Attie, M., Beebe, J., Berniūnas, R., Boudesseul, J., Colombo, M., Cushman, F., Diaz, R., N'Djaye Nikolai van Dongen, N., Dranseika, V., Earp, B. D., Torres, A. G., Hannikainen, I., Hernández-Conde, J. V., Hu, W., Jaquet, F., Khalifa, K., Kim, H., Kneer, M., Knobe, J., Kurthy, M., Lantian, A., Liao, S.-y., Machery, E., Moerenhout, T., Mott, C., Phelan, M., Phillips, J., Rambharose, N., Reuter, K., Romero, F., Sousa, P., Sprenger, J., Thalabard, E., Tobia, K., Viciana, H., Wilkenfeld, D., and Zhou, X. (2019). Estimating the reproducibility of experimental philosophy. *Review of Philosophy and Psychology*, 6:1–36. 2, 37

Cox, D. R. (1956). Some problems connected with statistical inference. Technical report, North Carolina State University. Dept. of Statistics. 13

Cristea, I. and Ioannidis, J. (2018). p values in display items are ubiquitous and almost invariably significant: A survey of top science journals. *PLOS ONE*, 13:e0197440. 8

Dahl, D. B., Scott, D., Roosen, C., Magnusson, A., and Swinton, J. (2019). *xtable: Export Tables to LaTeX or HTML.* R package version 1.8-4. 37

DerSimonian, R. and Laird, N. (1986). Meta-analysis in clinical trials. *Controlled clinical trials*, 7(3):177–188. 3, 7, 33, 34

Ebersole, C., Atherton, O., Belanger, A., Skulborstad, H., Allen, J., Banks, J., Baranski, E., Bernstein, M., Bonfiglio, D., Boucher, L., Brown, E., Budiman, N., Cairo, A., Capaldi, C., Chartier, C., Cicero, D., Coleman, J., Davis, W., and Nosek, B. (2016). Many Labs 3: Evaluating participant pool quality across the academic semester via replication. 50

Evans, M., Moshonov, H., et al. (2006). Checking for prior-data conflict. *Bayesian analysis*, 1(4):893–914. 10

Fayers, P. M., Ashby, D., and Parmar, M. k. (1997). Tutorial in biostatistics: Bayesian data monitoring in clinical trials. *Statistics in medicine*, 16(12):1413–1430. 7

Fisher, L. D. (1999). Carvedilol and the Food and Drug Administration (FDA) approval process: The FDA paradigm and reflections on hypothesis testing. *Controlled clinical trials*, 20(1):16–39. 31

Fisher, R. A. (1925). Theory of statistical estimation. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 22, pages 700–725. Cambridge University Press. 6

Fletcher, A., Spiegelhalter, D., Staessen, J., Thijs, L., and Bulpitt, C. (1993). Implications for trials in progress of publication of positive results. *The Lancet*, 342(8872):653–657. 7

Food and Drug Administration (1998). Guidance for industry: Providing clinical evidence of effectiveness for human drug and biological products. *Maryland: United States Food and Drug Administration*, 5. 34

Foster, D. and Sullivan, K. (1987). Computer program produces p-value graphics. *American journal of public health*, 77(7):880–881. 13

Freedman, L. S., Spiegelhalter, D. J., and Parmar, M. K. (1994). The what, why and how of Bayesian clinical trials monitoring. *Statistics in medicine*, 13(13-14):1371–1383. 7

Gardner, M. J. and Altman, D. G. (1986). Confidence intervals rather than p values: Estimation rather than hypothesis testing. *Br Med J (Clin Res Ed)*, 292(6522):746–750. 13

Gelman, A. and Loken, E. (2014). The statistical crisis in scienced: Data-dependent analysis–a "garden of forking paths" – explains why many statistically significant comparisons don't hold up. *American scientist*, 102(6):460–466. 6

Gilbert, D. T., King, G., Pettigrew, S., and Wilson, T. D. (2016). Comment on "Estimating the reproducibility of psychological science". *Science*, 351(6277):1037–1037. 2

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational researcher*, 5(10):3–8. 33

Gonzalez, R. (1994). The statistics ritual in psychological research. *Psychological Science*, 5(6):321–325. 13

Good, I. J. (1983). *Good thinking: The foundations of probability and its applications*. U of Minnesota Press. 8

Good, I. J. (1992). The Bayes/non-Bayes compromise: A brief review. *Journal of the American Statistical Association*, 87(419):597–606. 3, 12

Goodman, S. N. (1992). A comment on replication, p-values and evidence. *Statistics in medicine*, 11(7):875–879. 3, 13

Greenland, S. (2006). Bayesian perspectives for epidemiological research: I. foundations and basic methods. *International journal of epidemiology*, 35(3):765–775. 8

Greenland, S. (2011). Null misinterpretation in statistical testing and its impact on health risk assessment. *Preventive medicine*, 53(4-5):225–228. 8

Haidich, A.-B. (2010). Meta-analysis in medical research. *Hippokratia*, 14(Suppl 1):29. 3

Haile, S. R., Held, L., Meyer, S., Rueeger, S., Rufibach, K., and Schwab, S. (2019). *biostatUZH: Misc Tools of the Department of Biostatistics, EBPI, University of Zurich*. R package version 1.8.0/r82. 37

Heitjan, D. F. (1997). Bayesian interim analysis of phase II cancer clinical trials. *Statistics in medicine*, 16(16):1791–1802. 7

Held, L. (2019). The assessment of intrinsic credibility and a new argument for $p < 0.005$. *Royal Society open science*, 6(3):181534. 9

Held, L. (2020a). The harmonic mean χ^2 test to substantiate scientific findings. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*. 3, 12, 23, 25, 29, 31, 49, 53

Held, L. (2020b). A new standard for the analysis and design of replication studies (with discussion). *Journal of the Royal Statistical Society, Series A (Statistics in Society)*, 183(2):431 – 448. 3, 8, 11, 12, 36, 37, 38, 41, 49

Held, L. and Sabanés Bové, D. (2014). Applied statistical inference. *Springer, Berlin Heidelberg, doi*, 10(978-3):16. 5

Higgins, J. P. and Spiegelhalter, D. J. (2002). Being sceptical about meta-analyses: A Bayesian perspective on magnesium trials in myocardial infarction. *International journal of epidemiology*, 31(1):96–104. 7

Higgins, J. P., Thompson, S. G., and Spiegelhalter, D. J. (2009). A re-evaluation of random-effects meta-analysis. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 172(1):137–159. 34

Hunter, J. E. (1997). Needed: A ban on the significance test. *Psychological science*, 8(1):3–7. 13

Ioannidis, J. P. (2005). Why most published research findings are false. *PLos med*, 2(8):e124. 2, 6

Johnson, V. E., Payne, R. D., Wang, T., Asher, A., and Mandal, S. (2017). On the reproducibility of psychological science. *Journal of the American Statistical Association*, 112(517):1–10. 36

Kass, R. E. and Greenhouse, J. B. (1989). Comments on "Investigating therapies of potentially great benefit: ECMO" (by JH Ware). *Statistical Science*, 4(3):1–3. 7

Kassambara, A. (2020). *ggpubr: 'ggplot2' Based Publication Ready Plots*. R package version 0.2.5. 37

Kay, R. (2014). *Statistical thinking for non-statisticians in drug regulation*. Wiley Online Library. 2, 34, 35

Keppel, G. (1991). *Design and analysis: A researcher's handbook*. Prentice-Hall, Inc. 2

Keren, G. E. and Lewis, C. E. (1993). *A handbook for data analysis in the behavioral sciences: Methodological issues*. Lawrence Erlbaum Associates, Inc. 6

Killeen, P. R. (2005). An alternative to null-hypothesis significance tests. *Psychological science*, 16(5):345–353. 3, 13

Klein, R., Vianello, M., Hasselman, F., Adams, B., Adams, R., Alper, S., Aveyard, M., Axt, J., Babalola, M., Bahník, Š., Barlow, F., Berkics, M., Bernstein, M., Berry, D., Bialobrzeska, O., Bocian, K., Brandt, M., Busching, R., and Cai, H. (2018). Many Labs 2: Investigating Variation in Replicability Across Sample and Setting. 50

Klein, R. A., Ratliff, K. A., Vianello, M., Adams Jr, R. B., Bahník, Š., Bernstein, M. J., Bocian, K., Brandt, M. J., Brooks, B., Brumbaugh, C. C., et al. (2014). Investigating variation in replicability. 50

Lakens, D., Adolfi, F. G., Albers, C. J., Anvari, F., Apps, M. A., Argamon, S. E., Baguley, T., Becker, R. B., Benning, S. D., Bradford, D. E., et al. (2018). Justify your alpha. *Nature Human Behaviour*, 2(3):168–171. 6

Lemon J (2006). Plotrix: A package in the red light district of R. *R-News*, 6(4):8–12. 37

Lilford, R. and Braunholtz, D. (1996). For Debate: The statistical basis of public policy: A paradigm shift is overdue. *BMJ*, 313(7057):603–607. 6

Ly, A., Etz, A., Marsman, M., and Wagenmakers, E.-J. (2019). Replication Bayes factors from evidence updating. *Behavior research methods*, 51(6):2498–2508. 3, 21

Lykken, D. T. (1968). Statistical significance in psychological research. *Psychological bulletin*, 70(3p1):151. 2

Matthews, R. A. (2001a). Methods for assessing the credibility of clinical trial outcomes. *Drug Information Journal*, 35(4):1469–1478. 8

Matthews, R. A. (2001b). Why should clinicians care about Bayesian methods? *Journal of Statistical Planning and Inference*, 94(1):43–58. 3, 6, 8, 9

Matthews, R. A. (2018). Beyond 'significance': Principles and practice of the Analysis of Credibility. *Royal Society Open Science*, 5(1):171047. 3, 8

Mau, J. (1988). A statistical assessment of clinical equivalence. *Statistics in Medicine*, 7(12):1267–1277. 13

Maxwell, S. E., Lau, M. Y., and Howard, G. S. (2015). Is psychology suffering from a replication crisis? What does "failure to replicate" really mean? *American Psychologist*, 70(6):487. 23, 24, 50

McShane, B. B., Gal, D., Gelman, A., Robert, C., and Tackett, J. L. (2019). Abandon statistical significance. *The American Statistician*, 73(sup1):235–245. 6

Milner, J. and Good, I. (1951). Probability and the weighing of evidence. *The University of Toronto Law Journal*, 9:159. 8

Neyman, J. and Pearson, E. S. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference: Part I. *Biometrika*, pages 175–240. 6

Nosek, B. A. and Errington, T. M. (2017). Reproducibility in cancer biology: Making sense of replications. *Elife*, 6:e23383. 2

Nosek, B. A. and Lakens, D. (2014). Registered reports. 2

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251):aac4716. 2, 3, 16, 24, 36, 39

Parmar, M. K., Spiegelhalter, D. J., Freedman, L. S., and Committee, C. S. (1994). The CHART trials: Bayesian design and monitoring in practice. *Statistics in medicine*, 13(13-14):1297–1312. 7

Pernet, C. (2015). Null hypothesis significance testing: A short tutorial. *F1000Research*, 4. 6

Poole, C. (1987a). Beyond the confidence interval. *American Journal of Public Health*, 77(2):195–199. 13

Poole, C. (1987b). Confidence intervals exclude nothing. *American journal of public health*, 77(4):492–493. 13

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. 37

Rothman, K. J. (1998). Writing for epidemiology. *Epidemiology*, pages 333–337. 13

Rothman, K. J., Greenland, S., and Lash, T. L. (2008). *Modern epidemiology*. Lippincott Williams & Wilkins. 13

Sargent, C. L. (1981). The repeatability of significance and the significance of repeatability. *European Journal of Parapsychology*, 3(4):423–443. 2

Schmidt, S. (2009). Shall we really do it again? The powerful concept of replication is neglected in the social sciences. *Review of general psychology*, 13(2):90–100. 2

Senn, S. S. (2008). *Statistical issues in drug development*, volume 69. John Wiley & Sons. 2

Shakespeare, T. P., Gebski, V. J., Veness, M. J., and Simes, J. (2001). Improving interpretation of clinical studies by use of confidence levels, clinical significance curves, and risk-benefit contours. *The Lancet*, 357(9265):1349–1353. 13, 14

Simons, D. J., Holcombe, A. O., and Spellman, B. A. (2014). An introduction to registered replication reports at perspectives on psychological science. *Perspectives on Psychological Science*, 9(5):552–555. 44

Simonsohn, U. (2015). Small telescopes: Detectability and the evaluation of replication results. *Psychological science*, 26(5):559–569. 3, 13

Slowikowski, K. (2020). *ggrepel: Automatically Position Non-Overlapping Text Labels with 'ggplot2'*. R package version 0.8.2. 37

Smith, A. H. and Bates, M. N. (1992). Confidence limit analyses should replace power calculations in the interpretation of epidemiologic studies. *Epidemiology*, pages 449–452. 13

Spiegelhalter, D. J., Abrams, K. R., and Myles, J. P. (2004). *Bayesian approaches to clinical trials and health-care evaluation*, volume 13. John Wiley & Sons. 6, 7

Spiegelhalter, D. J., Myles, J. P., Jones, D. R., and Abrams, K. R. (1999). An introduction to Bayesian methods in health technology assessment. *BMJ*, 319(7208):508–512. 6

Spieglehalter, D., Freedman, L., and Parmar, M. (1994). Bayesian approaches to randomized trials (with discussion). *JR Stat. Soc. Ser. A*, 157:357–416. 7

Stene, J. and Miettinen, O. (1987). Theoretical Epidemiology: Principles of Occurrence Research in Medicine. *Biometrics*, 43:482. 13

Stodden, V., Leisch, F., and Peng, R. D. (2014). *Implementing reproducible research*. CRC Press. 3

Sullivan, K. M. and Foster, D. A. (1990). Use of the confidence interval function. *Epidemiology*, 1:39–42. 13

Szucs, D. and Ioannidis, J. (2017). When null hypothesis significance testing is unsuitable for research: A reassessment. *Frontiers in human neuroscience*, 11:390. 6

Verhagen, J. and Wagenmakers, E.-J. (2014). Bayesian tests to quantify the result of a replication attempt. *Journal of Experimental Psychology: General*, 143(4):1457. 3

Veroniki, A. A., Jackson, D., Viechtbauer, W., Bender, R., Bowden, J., Knapp, G., Kuss, O., Higgins, J. P., Langan, D., and Salanti, G. (2016). Methods to estimate the between-study variance and its uncertainty in meta-analysis. *Research synthesis methods*, 7(1):55–79. 34

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. 37