University of Zurich[UZH]

Master Thesis

Sceptical Confidence Interval for Evidential Assessment of Replication Studies

Xijin Chen

Supervisors: Prof. Leonhard Held, Charlotte Micheloud

June 25, 2020

Overview

- Introduction
- New standard for the assessment
- Application on replication projects
- Further research

Introduction

- Reproducibility is a core principle for scientific research
- Large-scale reproducibility projects (i.e. Open Science Collaboration)
- The problem of replication crisis

Standards for the replication success assessment

Open Science Collaboration used a method that severely underestimates the actual rate of replication (*Gilbert et al., 2016*).

– **Comparison** of the original and replication studies

Standards for the replication success assessment

Open Science Collaboration used a method that severely underestimates the actual rate of replication (*Gilbert et al., 2016*).

- **Comparison** of the original and replication studies
- **Standard significance** of the replication study
- Combination of evidence by **meta-analysis**
- Quantitative **Bayesian methods**
  (i.e. Replication Bayes Factor, *Ly et al., 2019*)

New standard for the assessment

- Two-trials rule
- Sceptical p-value & P-value function
- Sceptical p-value function & Sceptical confidence interval
- Alternative methods

Two-trials rule

At least two primary studies testing the same medical product, is required by drug regulations to make decisions for drug approval (*FDA*).

– ***P*-value**

Significant studies in the same direction ($p_o < 0.05, p_r < 0.05$)

Two-trials rule

At least two primary studies testing the same medical product, is required by drug regulations to make decisions for drug approval (*FDA*).

– ***P*-value**

Significant studies in the same direction ($p_o < 0.05, p_r < 0.05$)

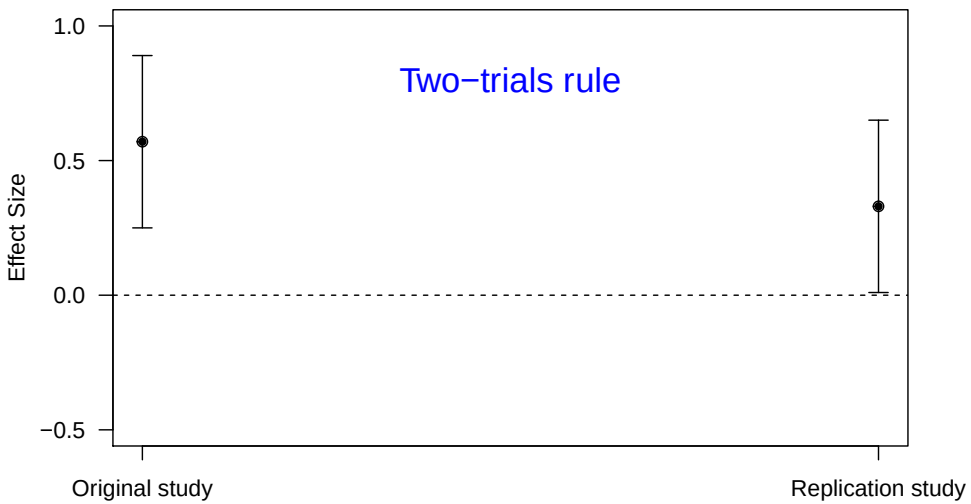
– **Confidence interval**

Effect size estimates: $\hat{\theta}_o$, $\hat{\theta}_r$

Variances: σ_o^2, σ_r^2

Two-trials rule

At least two primary studies testing the same medical product, is required by drug regulations to make decisions for drug approval (*FDA*).

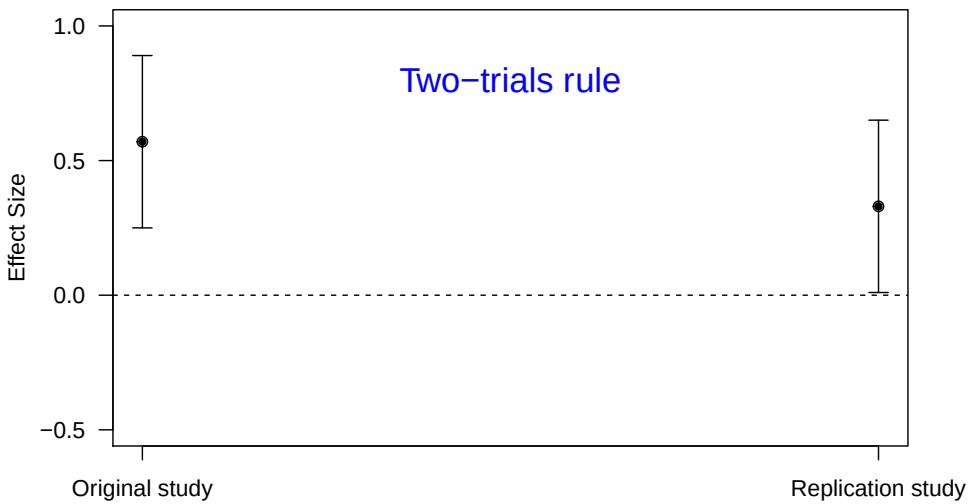


Effect size estimates: $\hat{\theta}_o$, $\hat{\theta}_r$

Variances: σ_o^2, σ_r^2

Two-trials rule

At least two primary studies testing the same medical product, is required by drug regulations to make decisions for drug approval (*FDA*).



Effect size estimates: $\hat{\theta}_o$, $\hat{\theta}_r$

Variances: σ_o^2, σ_r^2

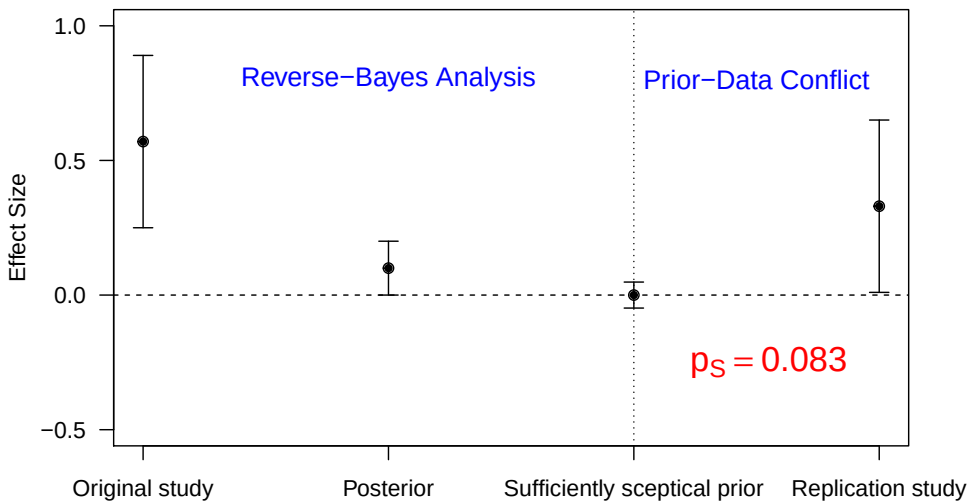
$\Downarrow$

Test statistics: t_o, t_r

Relative sample size: $c = n_r/n_o = \sigma_o^2/\sigma_r^2$

Sceptical *p*-value

A new standard for the analysis and design of replication studies (*Held, 2020a*)

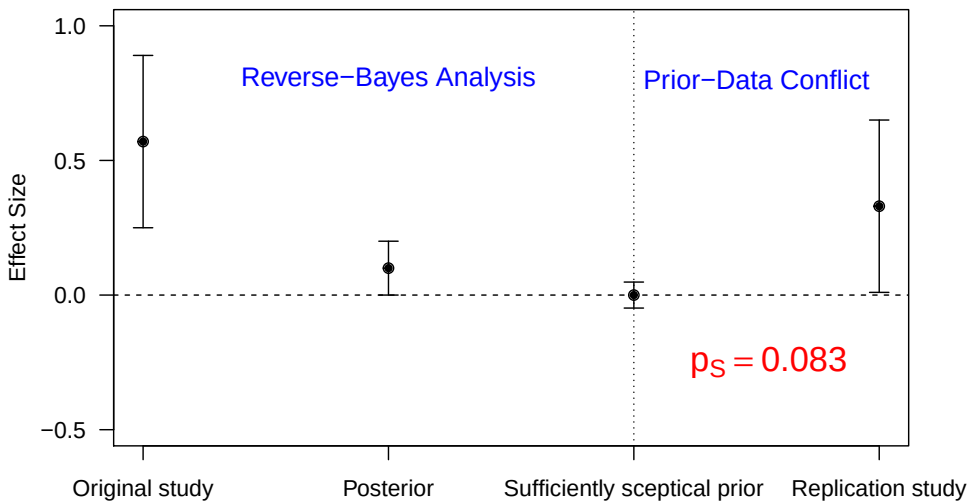


Analysis of Credibility
(*Matthews, 2001, 2018*)

Assessment of Prior-Data Conflict
(*Box, 1980*)

Sceptical *p*-value

A new standard for the analysis and design of replication studies (*Held, 2020a*)



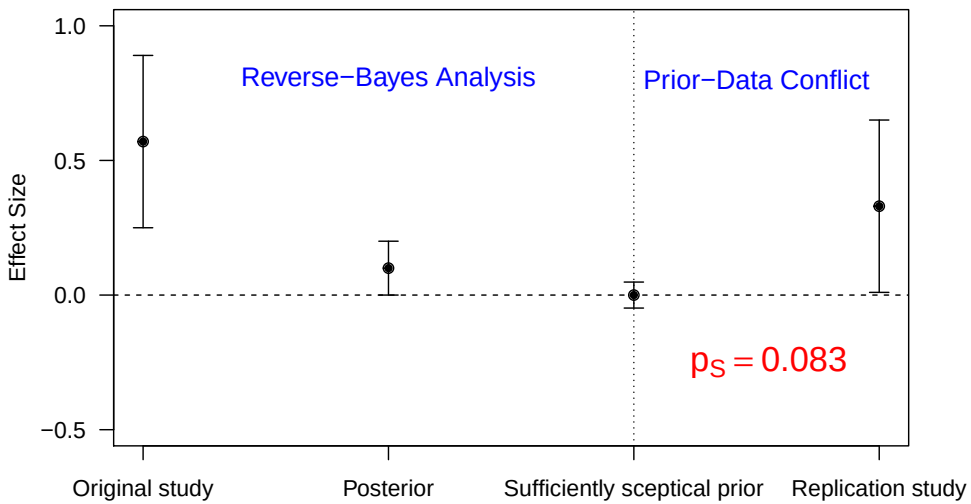
Analysis of Credibility
(*Matthews, 2001, 2018*)

Assessment of Prior-Data Conflict
(*Box, 1980*)

– A standard based on the Bayes-non-Bayes compromise (*Good, 1992*)

Sceptical *p*-value

A new standard for the analysis and design of replication studies (*Held, 2020a*)



Analysis of Credibility
(*Matthews, 2001, 2018*)

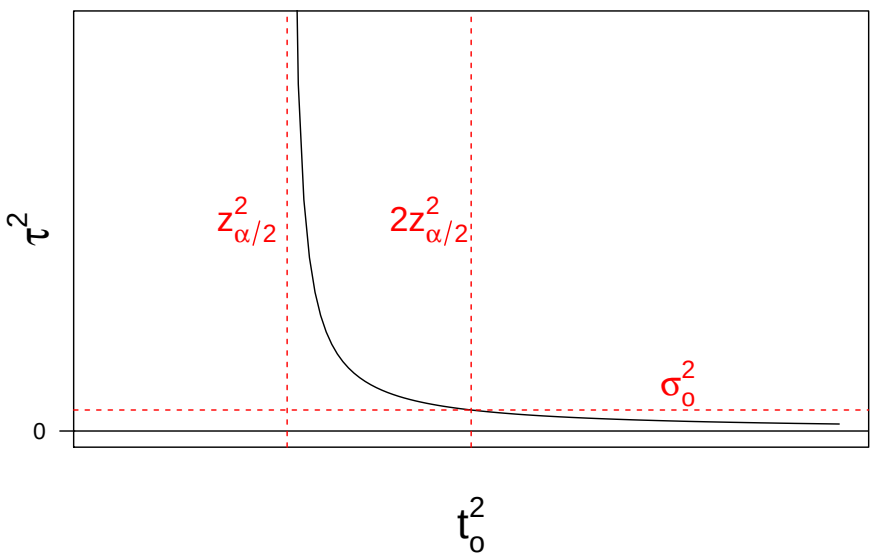
Assessment of Prior-Data Conflict
(*Box, 1980*)

- **A standard based on the Bayes-non-Bayes compromise (*Good, 1992*)**
- **A claim of replication success at the α if $p_S \leq \alpha$ or $z_S^2 \geq z_{\alpha/2}^2$**

  Specification of α: $\alpha = 0.05$ (nominal) and $\alpha = 0.13$ (calibrated)

Sufficiently sceptical prior

Connecting the evidence from both studies, with mean $\mu = 0$ and variance τ^2



Original study with $t_o = (\hat{\theta}_o - \mu)/\sigma_o$:

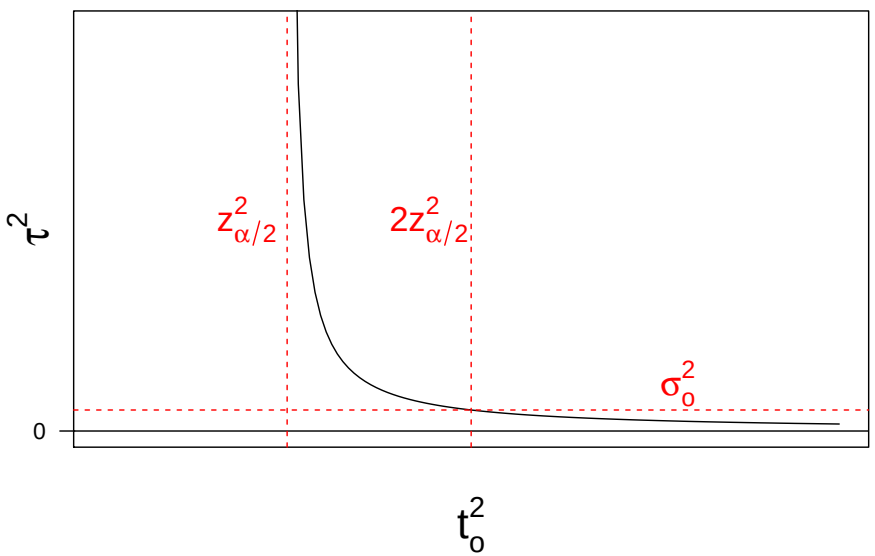
$$\tau^2 = \frac{\sigma_o^2}{t_o^2/z^2_{\alpha/2} - 1},$$

Replication study with $t_r = (\hat{\theta}_r - \mu)/\sigma_r$:

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

Sufficiently sceptical prior

Connecting the evidence from both studies, with mean $\mu = 0$ and variance τ^2



Original study with $t_o = (\hat{\theta}_o - \mu)/\sigma_o$:

$$\tau^2 = \frac{\sigma_o^2}{t_o^2/z_{\alpha/2}^2 - 1},$$

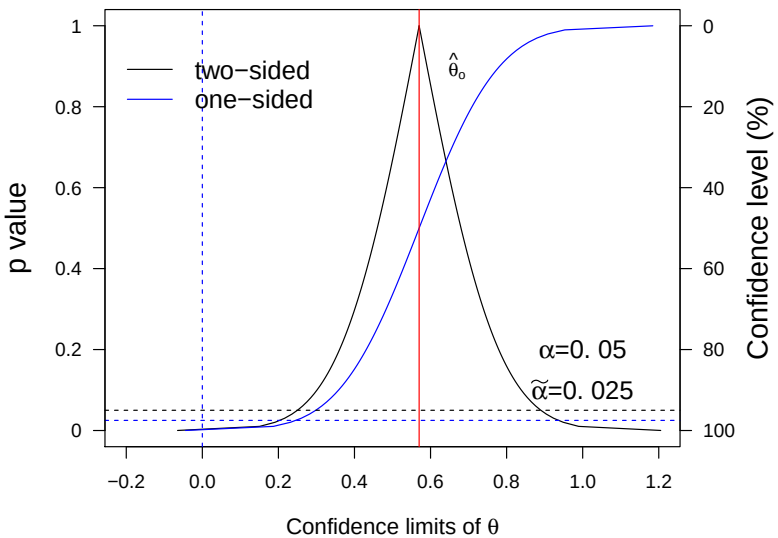
Replication study with $t_r = (\hat{\theta}_r - \mu)/\sigma_r$:

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

– **A nice expression of our scepticism about the previous finding: τ^2**

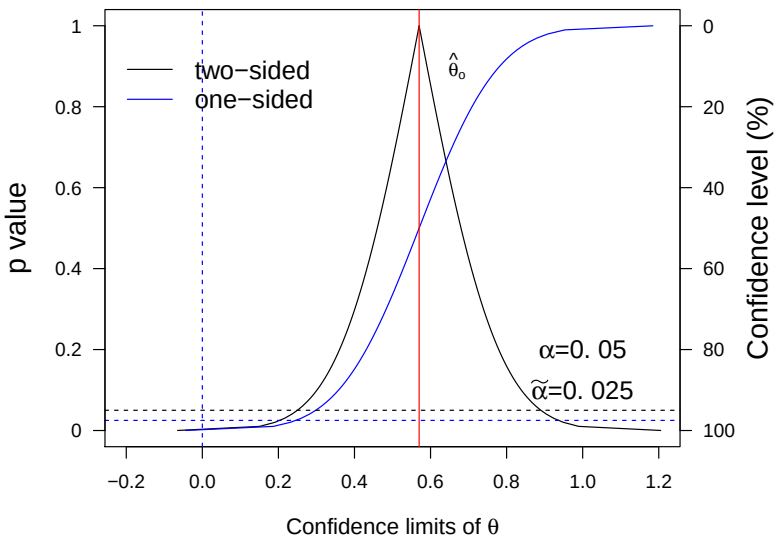
P-value function & Confidence interval

It is a good idea to calculate confidence intervals for various levels instead of using only the conventional fixed 95% level (*Cox, 1958*).



P-value function & Confidence interval

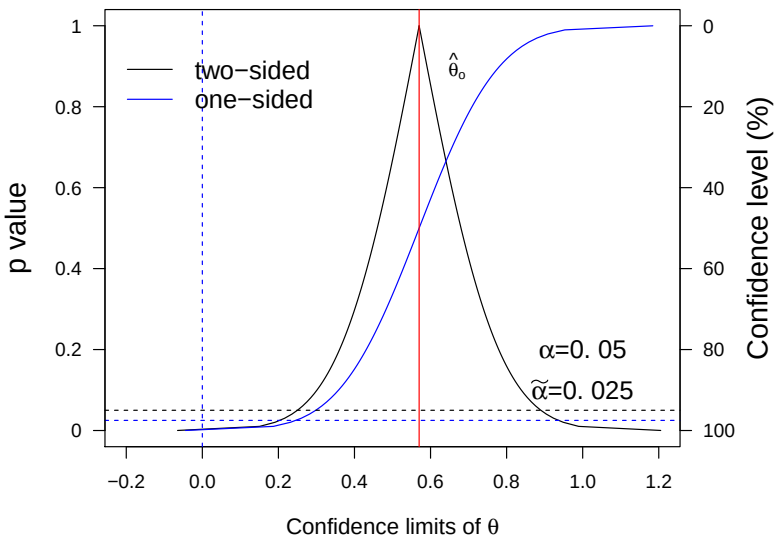
It is a good idea to calculate confidence intervals for various levels instead of using only the conventional fixed 95% level (*Cox, 1958*).



confidence distribution *(Cox, 1958)*
confidence curves *(Birnbaum, 1961)*
p-value function *(Miettinen, 1985)*
confidence interval function *(Sullivan and Foster, 1990)*
confidence distribution function *(Mau, 1988)*
clinical significance curve *(Shakespeare et al., 2001)*

P-value function & Confidence interval

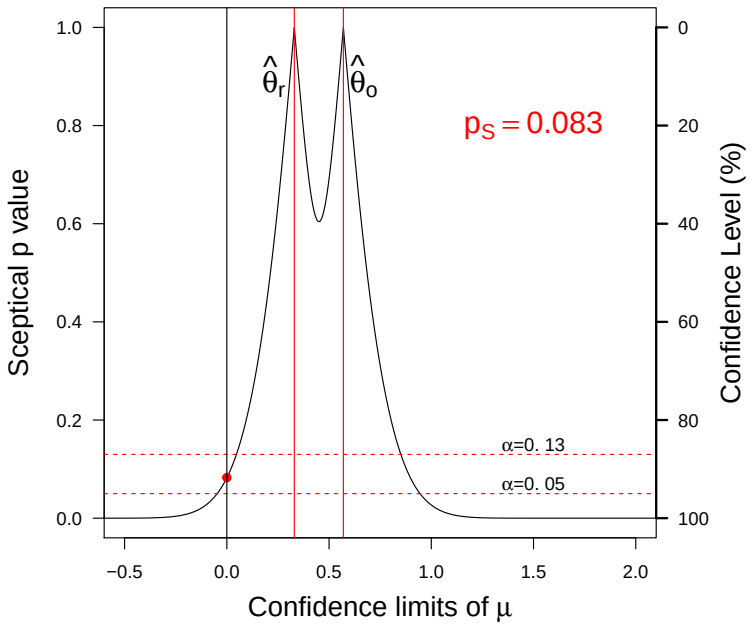
It is a good idea to calculate confidence intervals for various levels instead of using only the conventional fixed 95% level (*Cox, 1958*).

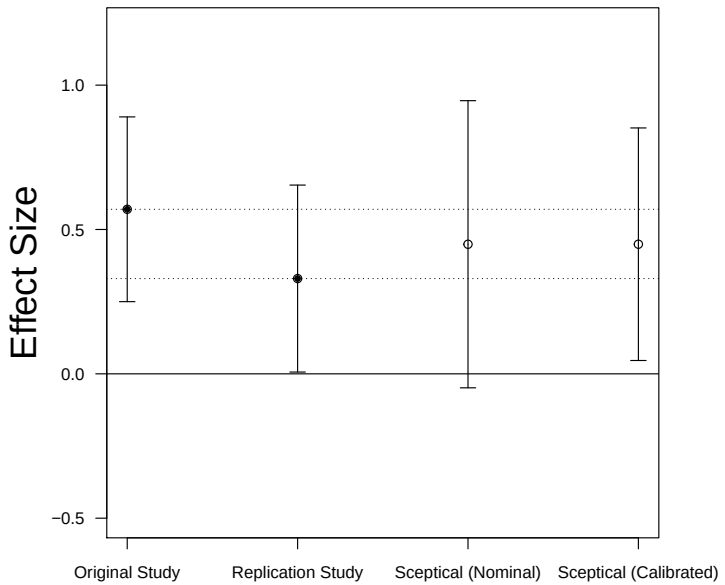


confidence distribution *(Cox, 1958)*
confidence curves *(Birnbaum, 1961)*
p-value function *(Miettinen, 1985)*
confidence interval function *(Sullivan and Foster, 1990)*
confidence distribution function *(Mau, 1988)*
clinical significance curve *(Shakespeare et al., 2001)*

– **The sceptical *p*-value function: an integration of sceptical *p*-values under varying mean values μ of the sufficiently sceptical prior**

Sceptical *p*-value function & Sceptical confidence interval

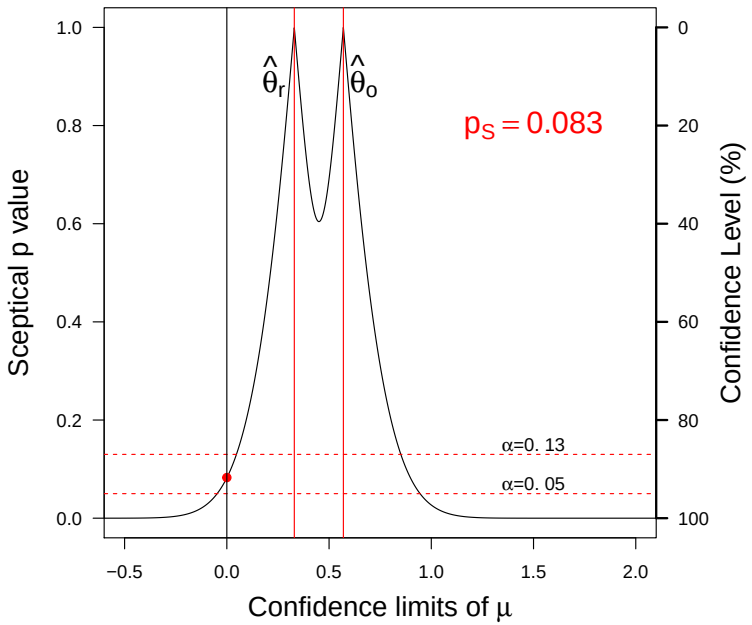


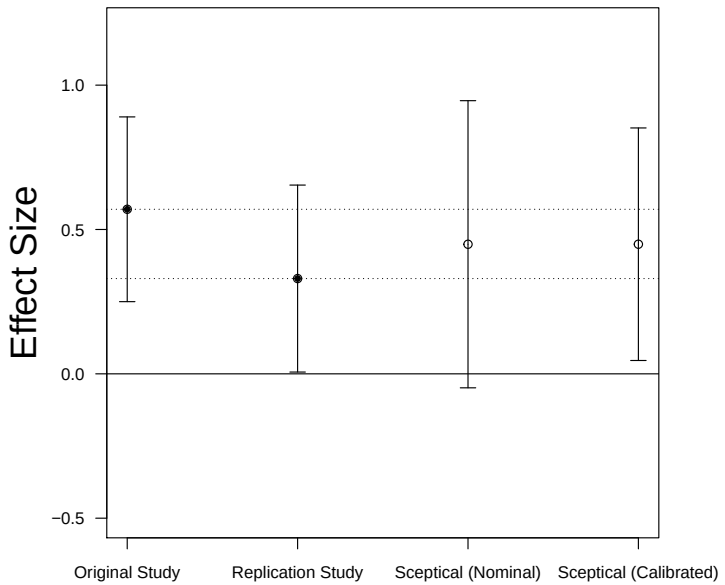


– **A claim of replication success:**

$p_S < \alpha$ **at** $\mu = 0$ **or** $\mu = 0$ **not included in** CI_S

Sceptical *p*-value function & Sceptical confidence interval

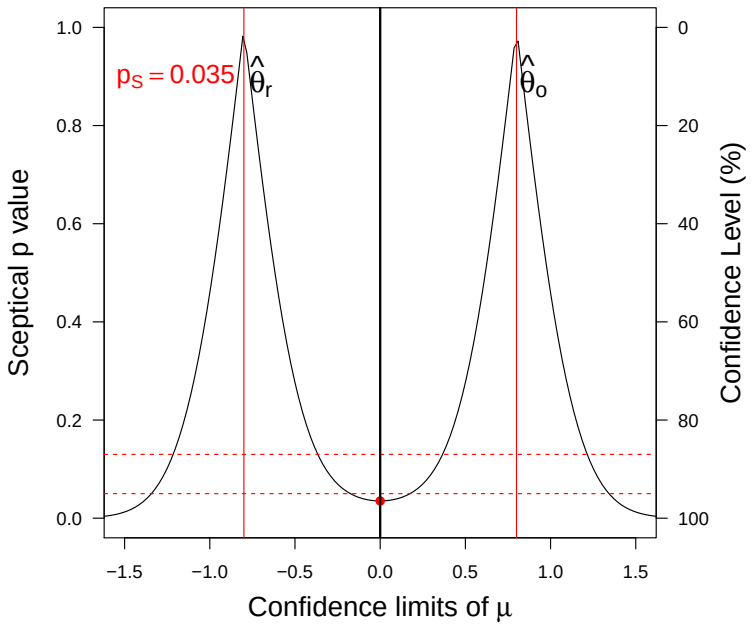


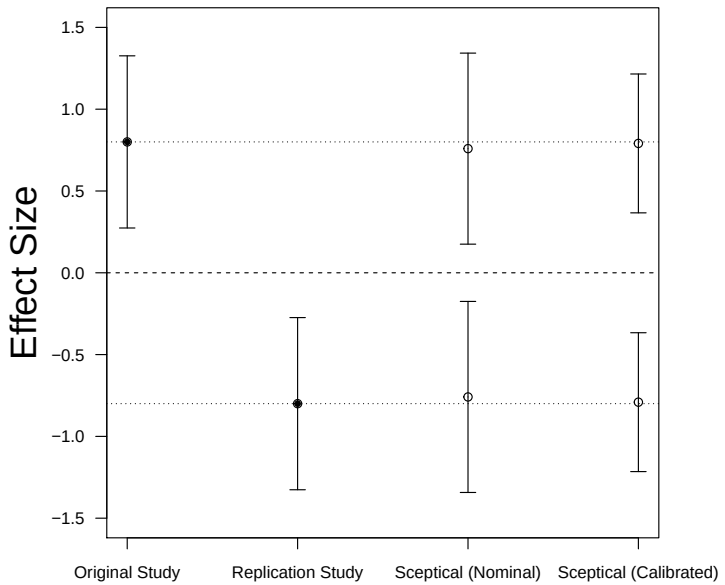


- **A claim of replication success:**

  $p_S < \alpha$ **at** $\mu = 0$ **or** $\mu = 0$ **not included in** CI_S
- **The bimodal function could lead to a set of unusual disjoint intervals**

Sceptical confidence set & Replication paradox





- The **sceptical confidence set** comes from a large **between-study conflict**
- A claim of replication success even under a **direction conflict** leads to the **replication paradox**

Replication paradox

The method may yield compelling evidence to support replication success even there is a direction conflict between two studies (*Ly et al., 2019*).

- Two necessary but not sufficient requirements
  1. Direction conflict
  2. Disjoint intervals due to a large between-study conflict
- Two alternatives for the assessment of replication success
  1. One-sided sceptical p-value (*Held, 2020a*)
  2. Harmonic mean χ^2 p-value (*Held, 2020b*)

Adjustments for the replication paradox

Avoid unrealistically small p-values or unusual confidence sets under the direction conflict

	Test statistic	Replication paradox	Assessment
I. $p_S = 2\{1 - \Phi(z_S)\}$	$z_S^2 = [\Phi^{-1}(1-\alpha_S)]^2$	No adjustment	$p_S \leq 2\alpha_S$
II. $\widetilde{p}_S = 1 - \Phi(z_S)$	$z_S^2 = [\Phi^{-1}(1-\alpha_S)]^2$	$\widetilde{p}_S = 1 - p_S/2$	$\widetilde{p}_S \leq \alpha_S$
III. $p_H = 2\{1 - \Phi(2z_S)\}$	$z_H^2 = \left[\Phi^{-1}(1-2\alpha_H)\right]^2$	$p_H > 0.5$	$p_H \leq 2\alpha_H$

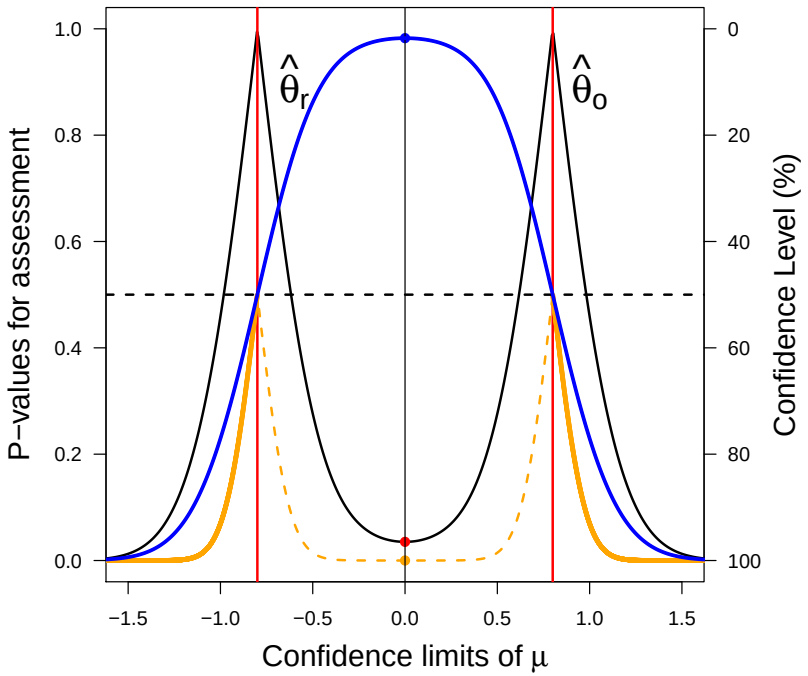
Note: $z_S^2 = 1/(1/t_o^2 + 1/t_r^2)$ and $z_H = 2z_S$. α_S and α_H are one-sided levels

– **Adjust unrealistically small p-values when there is a direction conflict**
– **No unusual confidence set, but only confidence interval**

Three approaches for the assessment

- Different p-value functions for the assessment (I,II&III)
- Empirical results for corresponding p-values and confidence intervals (I&II)
- Analytical results for the criteria (I&II)
- Summary & Comparison(I,II&III)

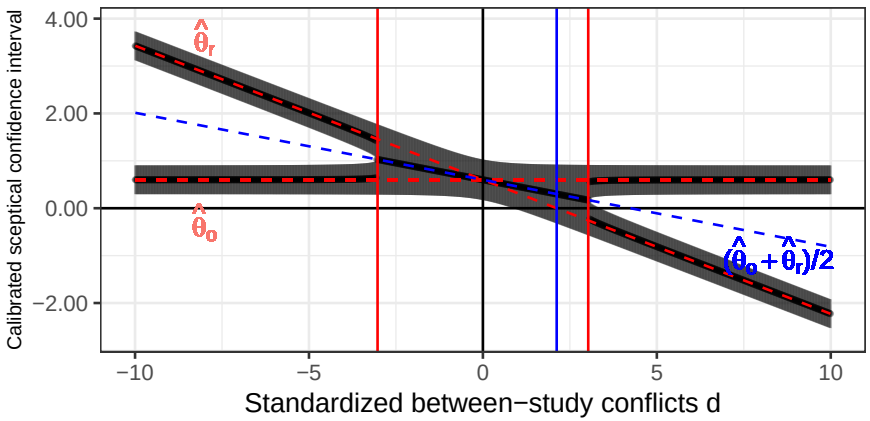
Different p-value functions

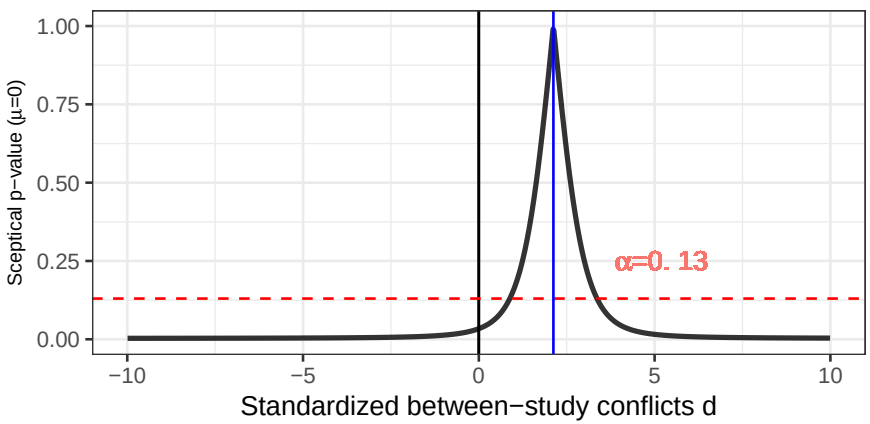


- Adjustment for the direction conflict
  - I. No adjustment
  - II. an adjusted value: $\widetilde{p}_S = 1 - p_S/2$
  - III. an equation: $p_H > 0.5$
- Range of p-values:
  - I. $p_S \in [0, 1]$
  - II. $\widetilde{p}_S \in [0, \text{arbitrary value}]$
  - III. $p_H \in [0, 0.5]$

- **Assessment p-values at $\mu = 0$**
- **Assessment confidence intervals at a specified level α or $\widetilde{\alpha}$**

I. Sceptical p-value methods: CI_S & p_S





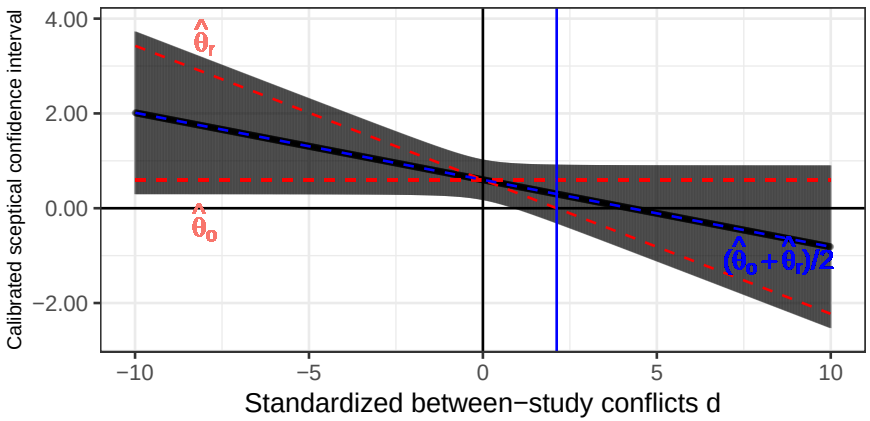
✓illustration of the replication paradox

✓unusual set if $|d| > 2z_{\alpha/2}$

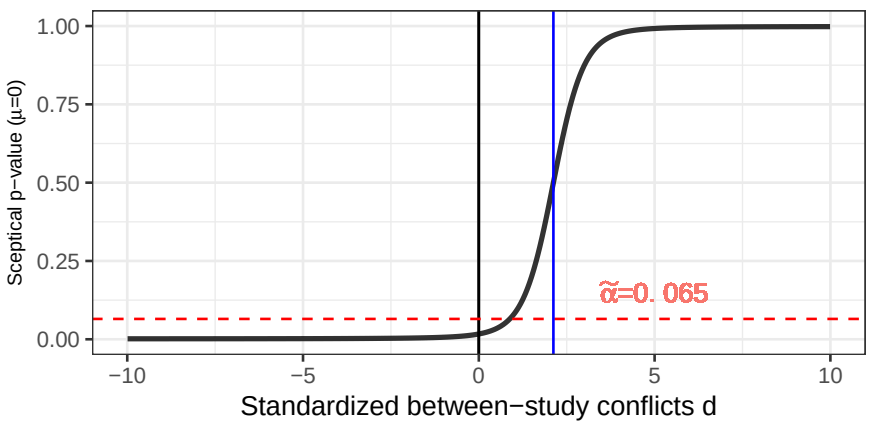
✓direction conflict if $t_o t_r < 0$

✓$p_S = 1$ at $t_r = 0$

II. Sceptical p-value methods: $\widetilde{CI}_S$ & $\tilde{p}_S$



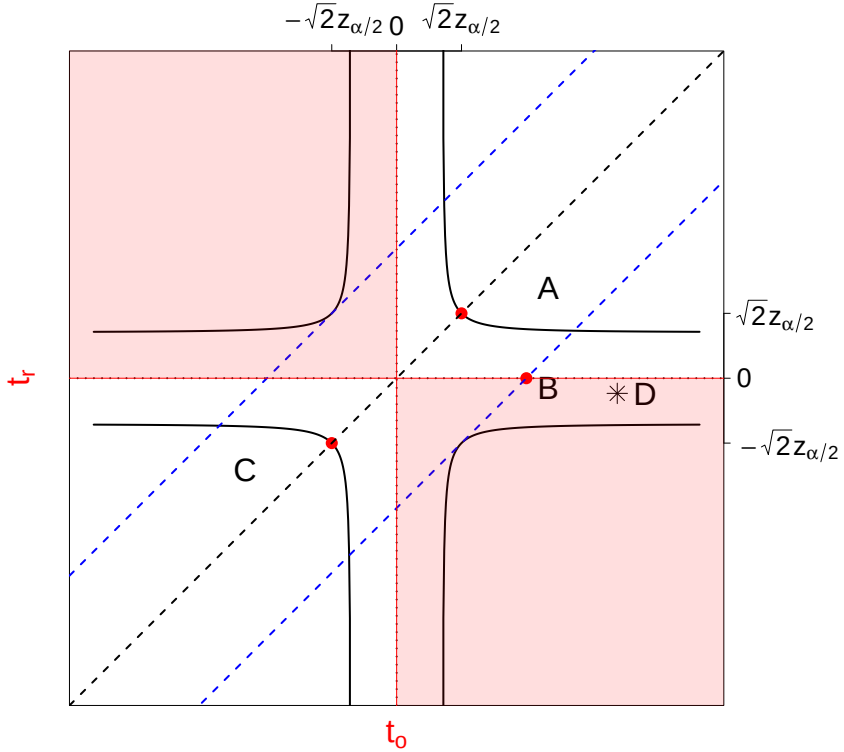
- ✓no unusual confidence set
- ✓no replication paradox
- ✓direction conflict if $t_o t_r < 0$



- ✓$\tilde{p}_S = 0.5$ at $t_r = 0$
- ✓no replication paradox
- ✓direction conflict if $\tilde{p}_S > 0.5$

I&II. Sceptical p-value method

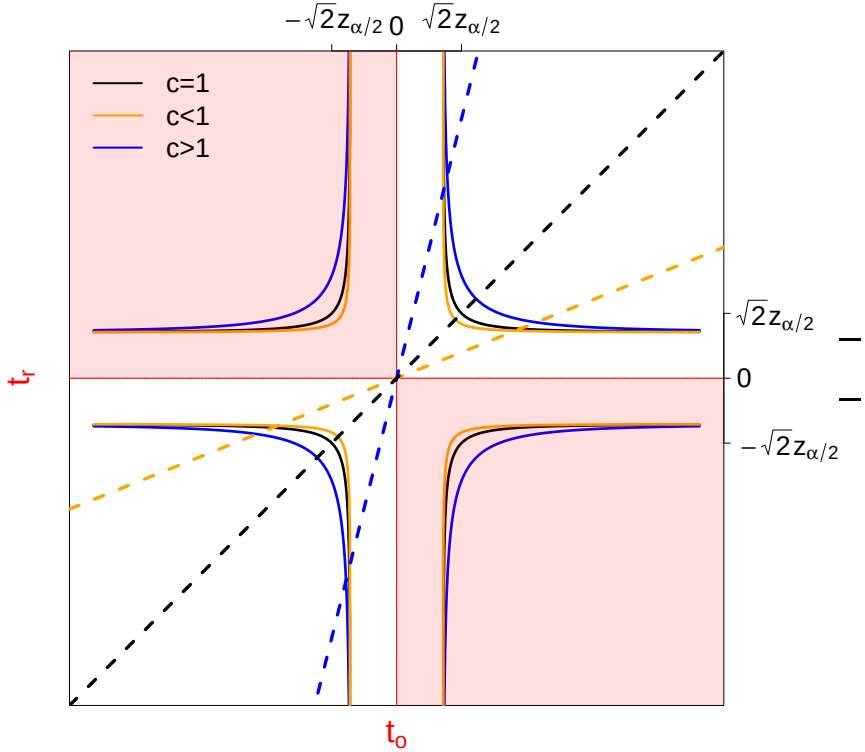
Under the equal sample size assumption ($\sigma = \sigma_o = \sigma_r$)



	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	$-\sqrt{2}z_{\alpha/2}$	$-\sqrt{2}z_{\alpha/2}$

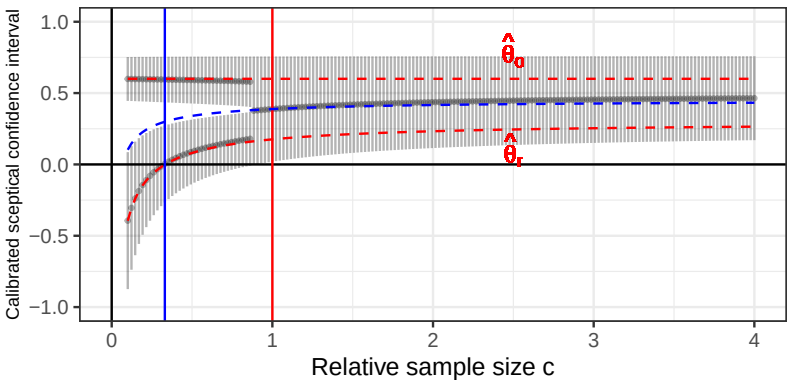
I&II. Sceptical p-value method

Under the general sample size assumption $c \neq 1$

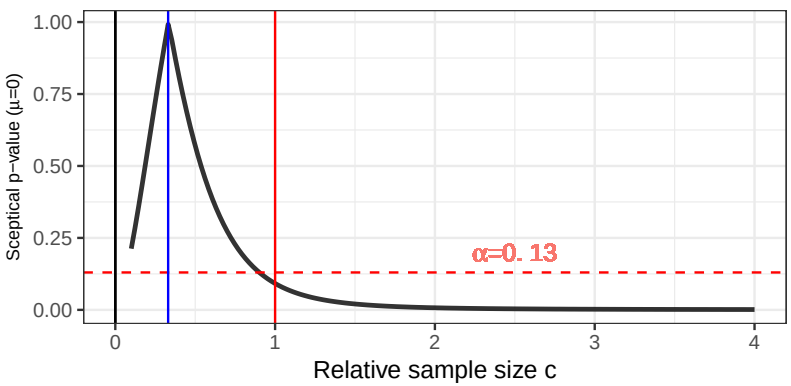


- $c > 1$: guarantee of the statistical power
- $c < 1$: limited resources/too large original study

I&II. Influence of relative sample size c



- Larger relative sample size is favorable
- Unusual set might be due to small c

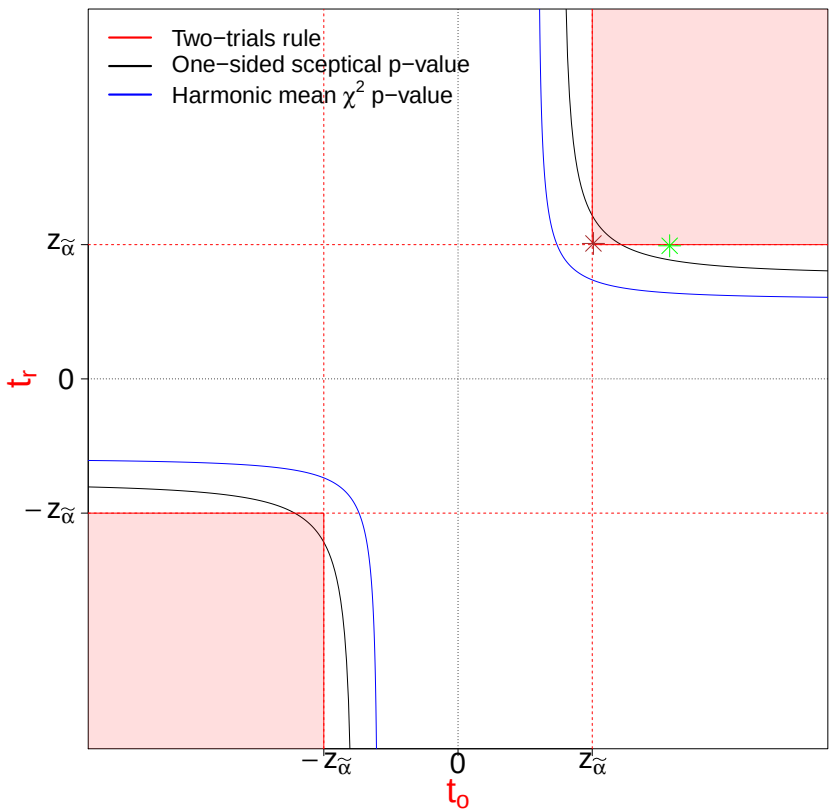


- $p_S = 1$ at $t_r = 0$
- Unwanted large p_S might be due to small c

Summary for different criteria

Method	Sceptical p-value				Harmonic mean χ^2 p-value	
Explanation for μ	mean value of sufficiently sceptical prior				point estimate of θ	
Test	Two-sided test (I)		One-sided test (II)		Two-sided test (III)	
P-value & CI	p_S	CI_S	$\widetilde{p}_S$	$\widetilde{CI}_S$	p_H	CI_H
Result	[0,1]	interval/set	[0,X]	interval	[0,0.5]	interval
Replication paradox	paradox		no paradox		no paradox	
Explanation	easy		hard		easy	
	confusing					
Information						
∗replication paradox	✗	✓	–	–	–	–
∗direction conflict	✗	✓	✓	✓	✓	✓
∗between-study conflict	✗	✓	✗	✗	✗	✗
∗individual $\hat{\theta}_i$	✗	✓	✗	✓	✗	✓
Multiple replications	not applicable				applicable	

Comparison with the two-trials rule



Method	$\tilde{\alpha}$	$p_o = 0.024$ $p_r = 0.024$	$p_o = 0.001$ $p_r = 0.026$
Two-trials	0.025	Success	No Success
$\tilde{p}_S$	0.065	No Success	Success
p_H	0.025	Success	Success

Note: p_o, p_r and $\tilde{\alpha}$ are in one-sided versions.

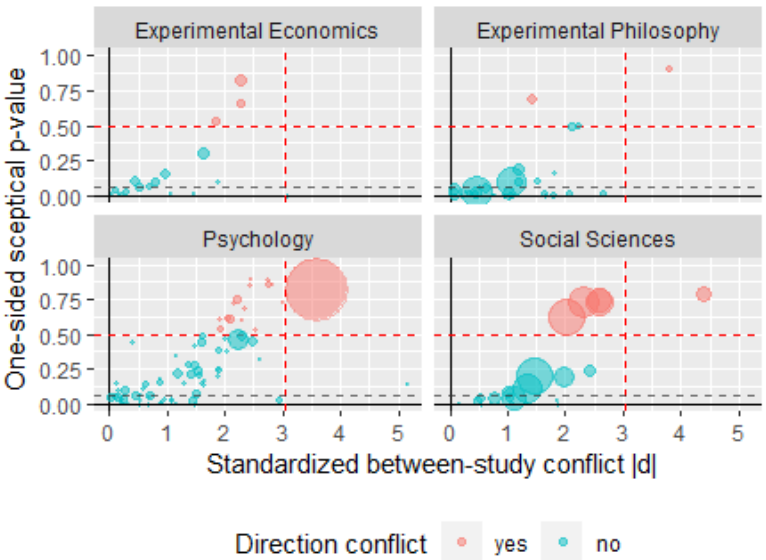
Replication project results

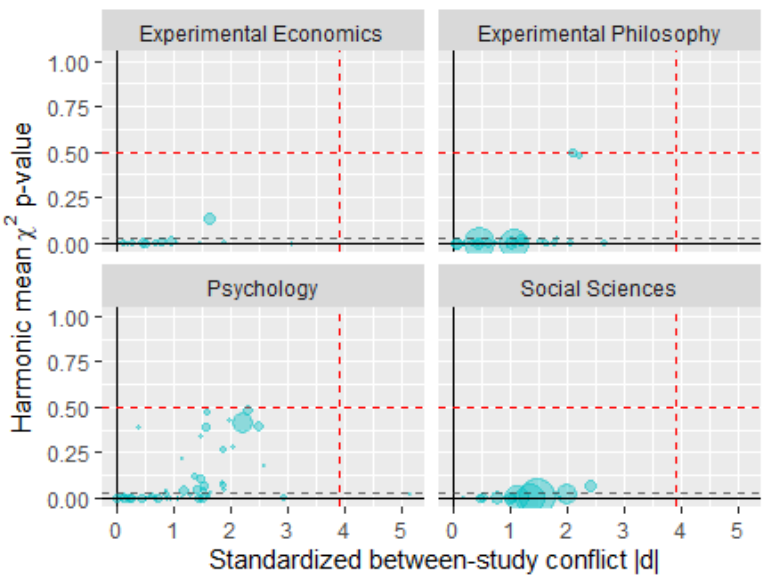
```
library(ReplicationSuccess)
data("RProjects")
Leonhard Held, Charlotte Micheloud, Samuel Pawel
```

Replication projects

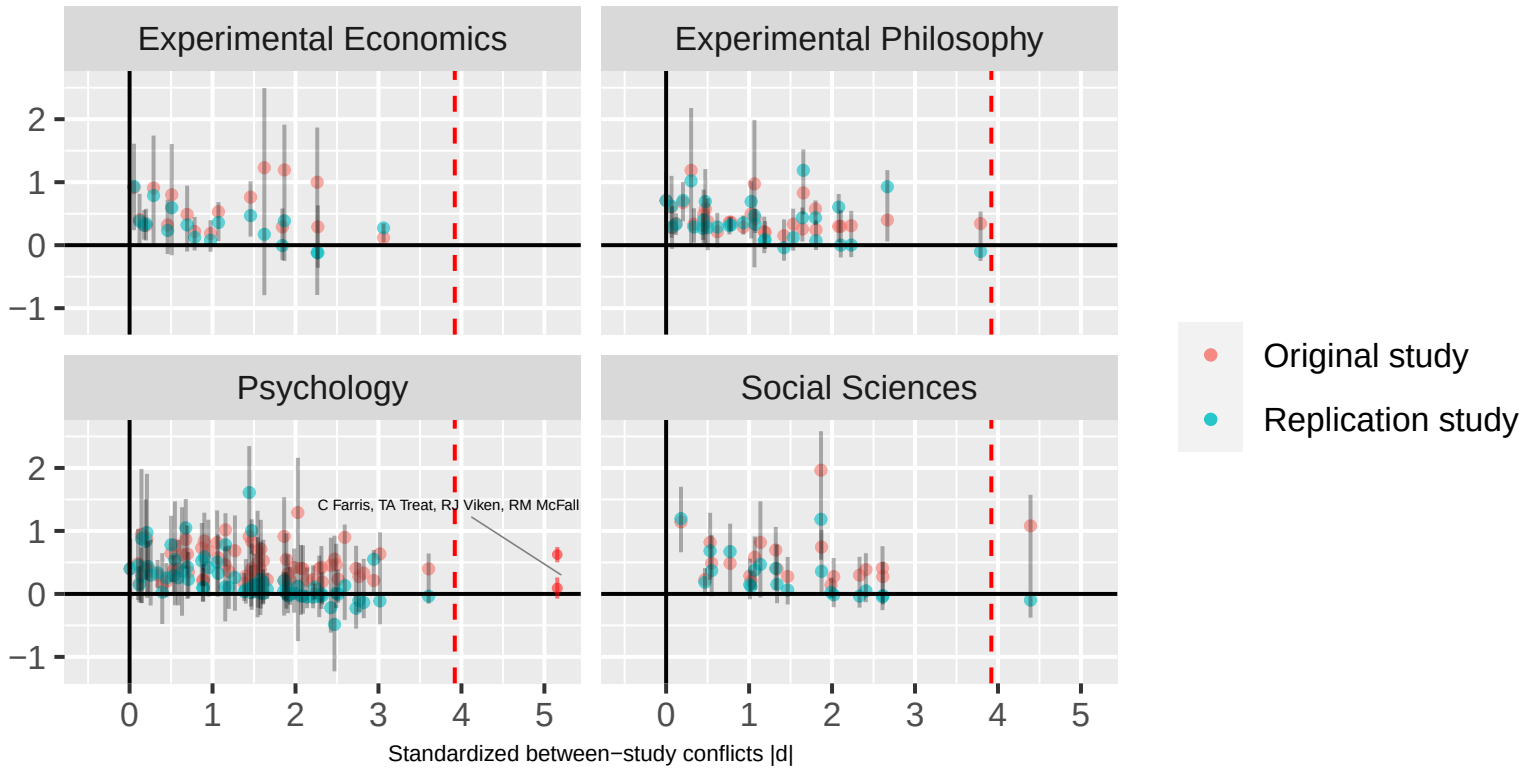
*R*eproducibility Project Psychology (73)
*E*xperimental Economics Replication Project (18)
*S*ocial Sciences Replication Project (21)
*E*xperimental Philosophy Replicability Project (31)

Sceptical p-value & Harmonic mean χ^2 p-value



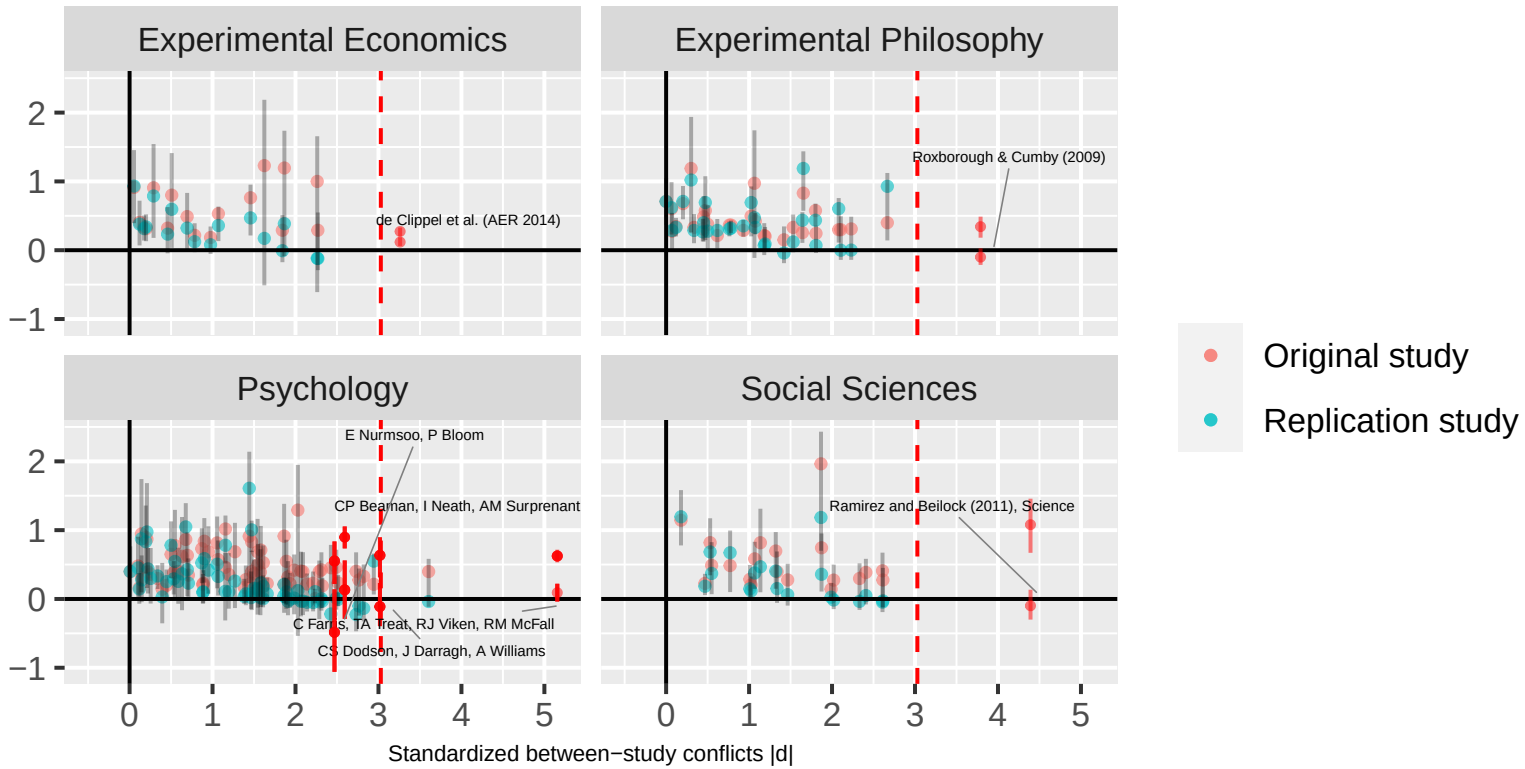


Nominal sceptical confidence intervals



– 1 **unusual sceptical confidence set due to the large between-study conflict**

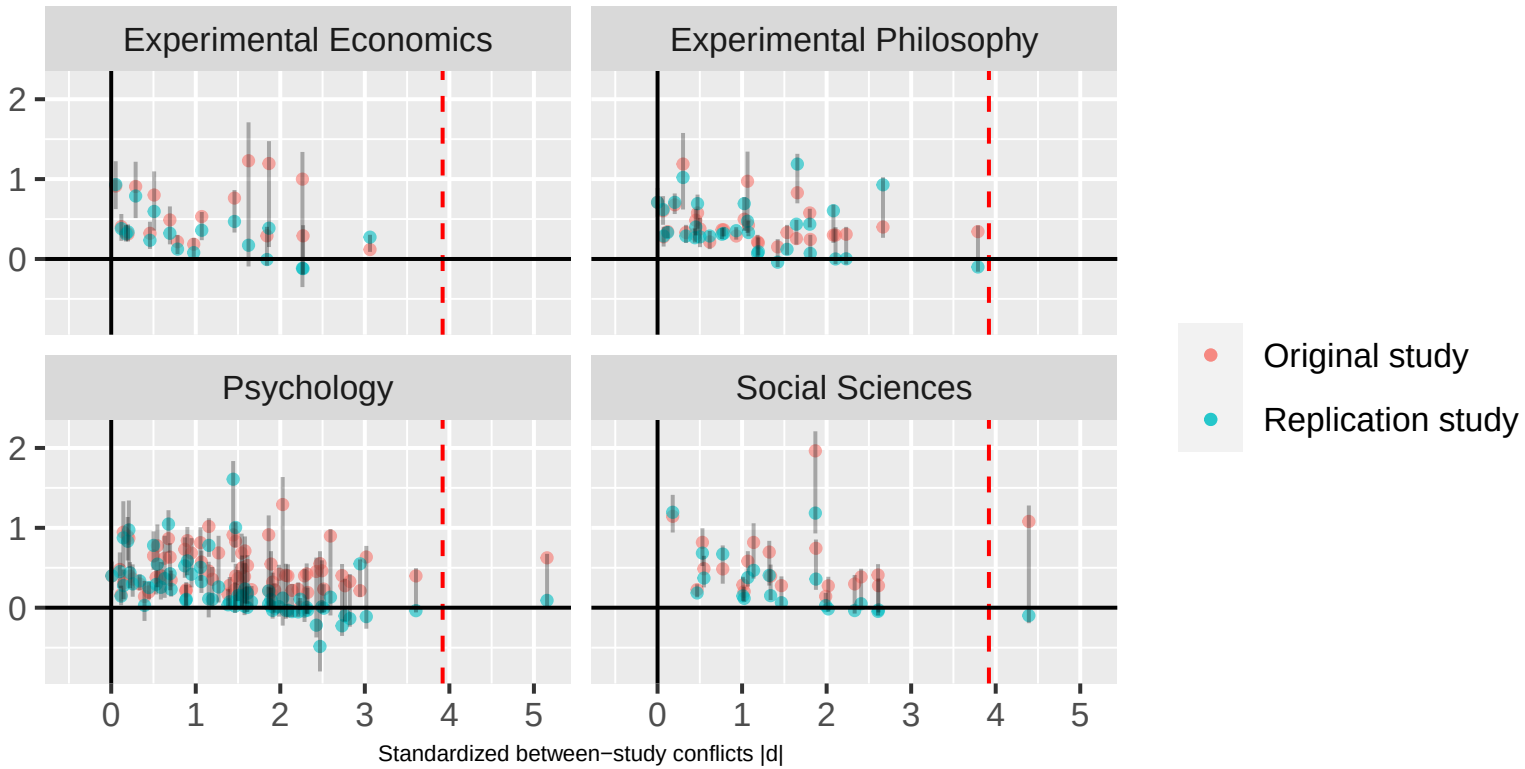
Calibrated sceptical confidence intervals



– 4 **unusual sceptical confidence set due to the large between-study conflict**

– 3 **unusual sceptical confidence set due to the small relative sample size** c

Harmonic mean χ^2 confidence interval



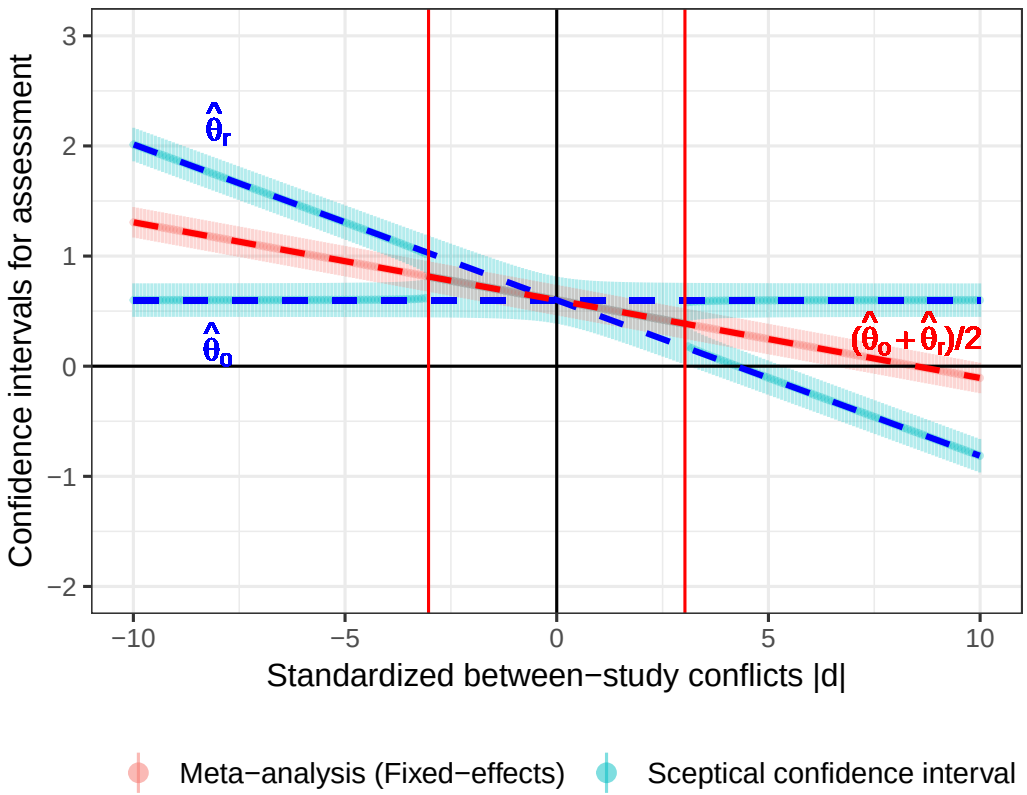
– **No unusual confidence set**

Meta-analysis

Fixed-effects model VS. Random-effects model

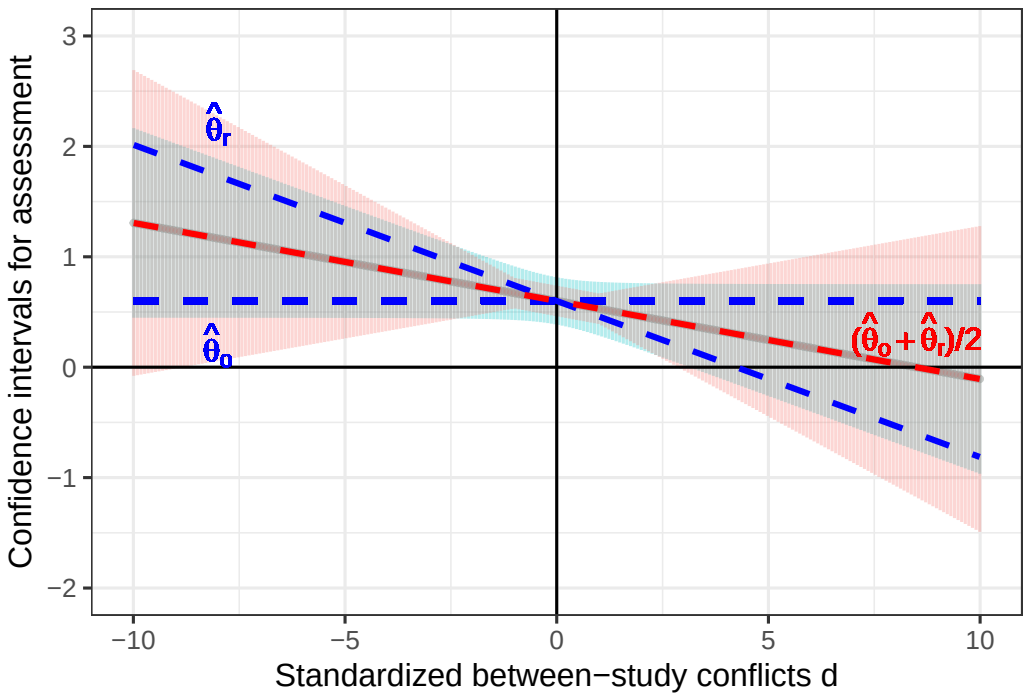
In the RPP, even though replication studies are reproduced in an extremely close way as in original studies, researcher might not be measuring exactly the same effect as studies from a variety of laboratories that all followed an identical in this project (*Simons et al., 2014*).

Meta-analysis & Sceptical confidence interval (CI_S)



- **Fixed-effects model is inappropriate**
- **Replication paradox**

Meta-analysis & Sceptical confidence interval ($\widetilde{CI}_S$)



- **Sensitivity difference**
- **No replication paradox**
- **Different conclusion**

? **The intrinsic difference: when can we expect to obtain different results**

Further research

? The intrinsic difference in contrast to Meta-analysis:
When can we expect to obtain different results

? From interval estimation to interval hypothesis:
Equivalence test

References

[1] A new standard for the analysis and design of replication studies. Held, Leonhard. (2020a). Journal of the Royal Statistical Society: Series A (Statistics in Society).

[2] The harmonic mean χ^2 test to substantiate scientific findings. Held, Leonhard. (2020b). Journal of the Royal Statistical Society: Series C (Applied Statistics).

[3] Beyond 'significance': Principles and practice of the Analysis of Credibility. Matthews, Robert AJ. (2018). Royal Society Open Science.

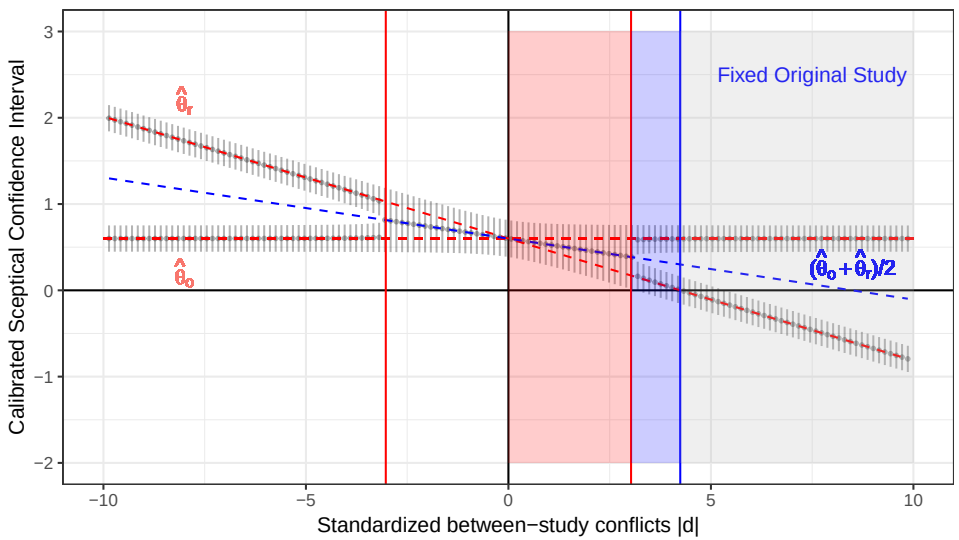
[4] Sampling and Bayes' inference in scientific modelling and robustness. Box, George EP. (1980). Journal of the Royal Statistical Society: Series A (General).

University of Zurich[UZH]

Appendix

Discussion about three cases

Extremely significant original studies ($t_o > 2\sqrt{2}z_{\alpha/2}$)



Sceptical confidence interval

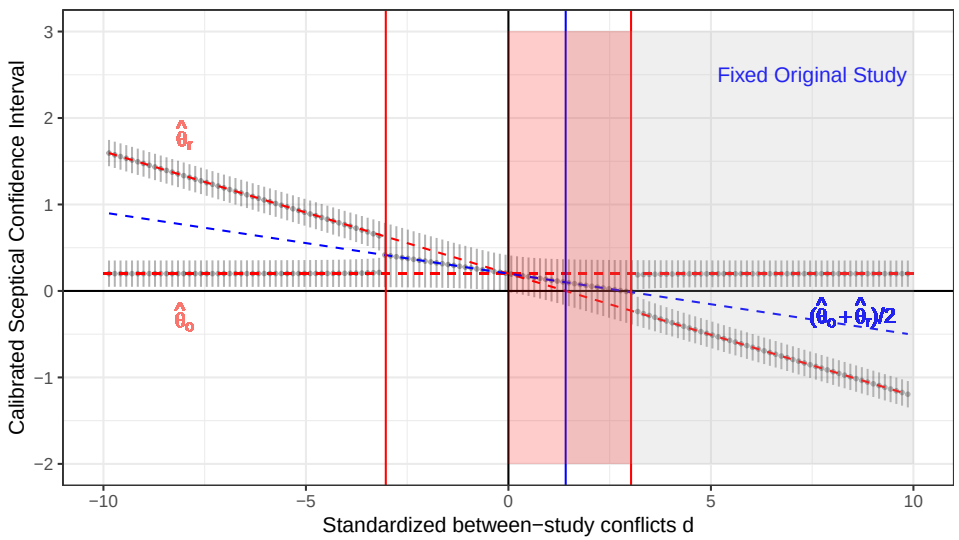
Sceptical confidence sets

Replication paradox (Possible)

– Result in a sceptical confidence set even there is no direction conflict
– Dominator for replication: **direction conflict boundary**

Discussion about three cases

Merely significant original studies ($t_o < 2\sqrt{2}z_{\alpha/2}$)



Sceptical confidence interval

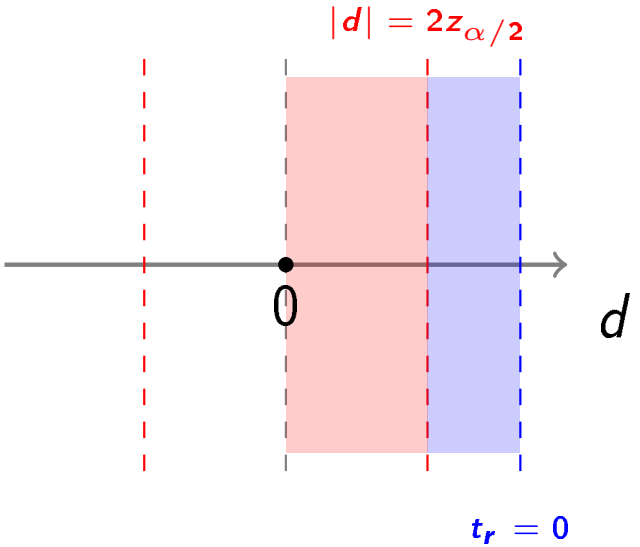
Direction conflict

Replication paradox (Possible)

– Result in a sceptical confidence set only when there is a direction conflict
– Dominator for replication: **unusual set threshold**

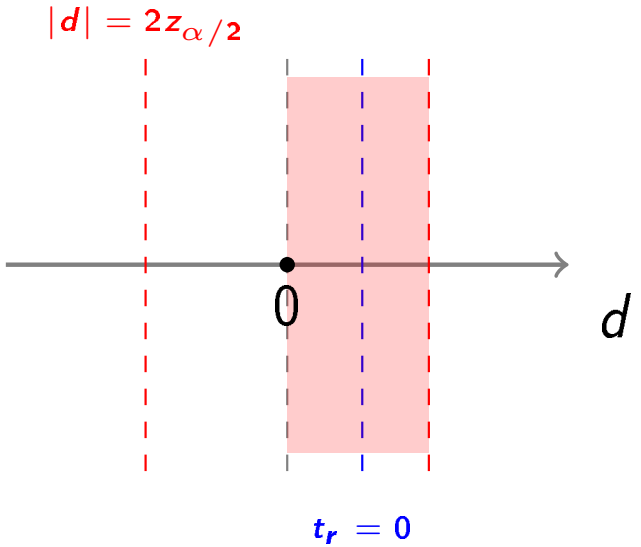
Dominator of the two requirements

Necessary but not sufficient for the replication paradox



$t_o > 2\sqrt{2}z_{\alpha/2}$

Dominator: direction conflict



$t_o < 2\sqrt{2}z_{\alpha/2}$

Dominator: unusual set

– Under a merely significant original study $t_o = \hat{\theta}_o/\sigma_o < 2\sqrt{2}z_{\alpha/2}$, a sceptical confidence set is only obtainable under the direction conflict.

Sceptical p-value method

Analysis of credibility + Prior-data conflict

The construction of sceptical confidence interval (CI_S): by non-zero mean μ of sufficiently sceptical prior:

$$Z_S^2 = \frac{1}{1/t_o^2 + 1/t_r^2}$$

$$t_o = \hat{\theta}_o/\sigma_o \qquad \Rightarrow \qquad t_o = (\hat{\theta}_o - \mu)/\sigma_o$$

$$t_r = \hat{\theta}_r/\sigma_r \qquad\qquad t_r = (\hat{\theta}_r - \mu)/\sigma_r$$

$$c = n_r/n_o = 1 \qquad\qquad c = n_r/n_o = 1$$

$\Rightarrow$ **A claim of replication success:**

$z_S^2 \geq z_{\alpha/2}^2$, **with a set of** μ **values where corresponding** $p_S < \alpha$.

Harmonic mean χ^2 test method

Evidence synthesis via harmonic mean of squared test statistics

The construction of harmonic mean χ^2 confidence interval (CI_S): by non-zero μ point estimates of individual effect size:

$$Z_H^2 = \frac{2}{\sum_{i=1}^{n} 1/Z_i^2} \overset{n=2}{=\!=} \frac{2}{1/Z_o^2 + 1/Z_r^2}$$

$$z_o = \hat{\theta}_o/\sigma_o \qquad \Rightarrow \qquad z_o = (\hat{\theta}_o - \mu)/\sigma_o$$

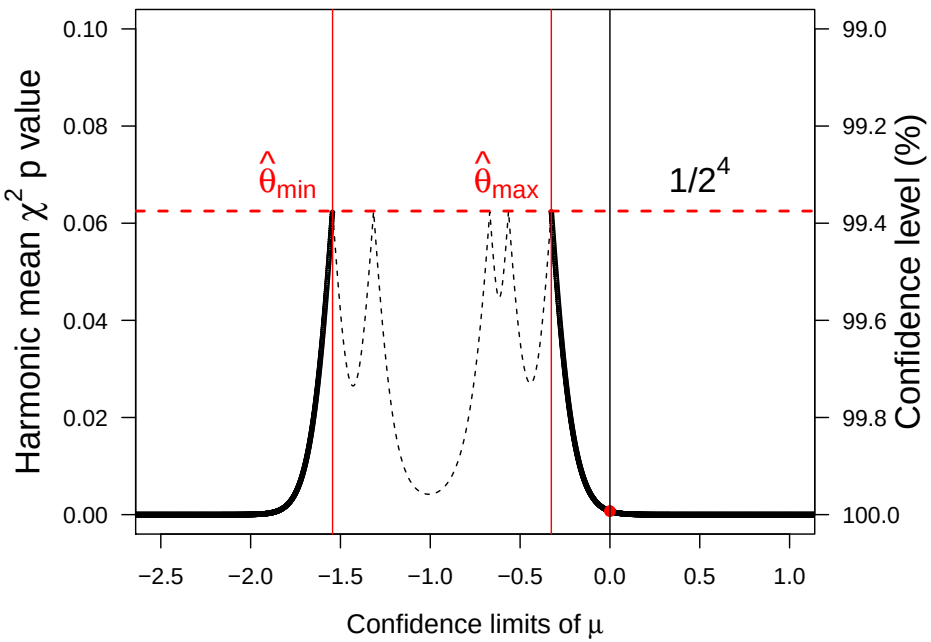
$$z_r = \hat{\theta}_r/\sigma_r \qquad\qquad z_r = (\hat{\theta}_r - \mu)/\sigma_r$$

$\Rightarrow$ **A claim of replication success:**

$z_H^2 \geq z_{\alpha/2}^2$, **with a set of** μ **values where corresponding** $p_H < \alpha$.

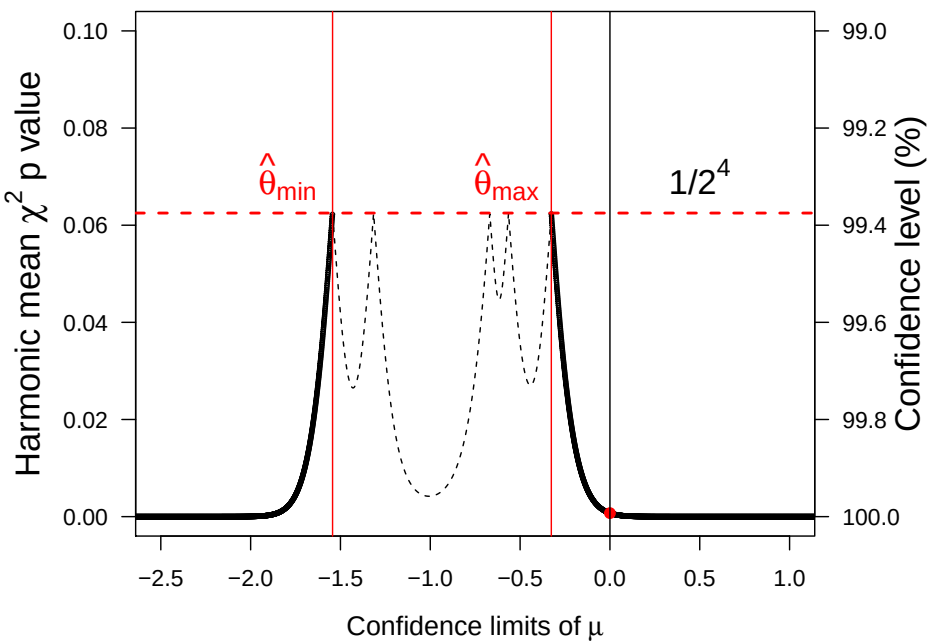
III. Harmonic mean χ^2 p-value function

Applicable for multiple replications



III. Harmonic mean χ^2 p-value function

Applicable for multiple replications



– VS. I&II (Sceptical p-value methods)

✓Applicable for multiple replications

✗No exact value under a direction conflict

– VS. I (p_S & CI_S)

✓No replication paradox

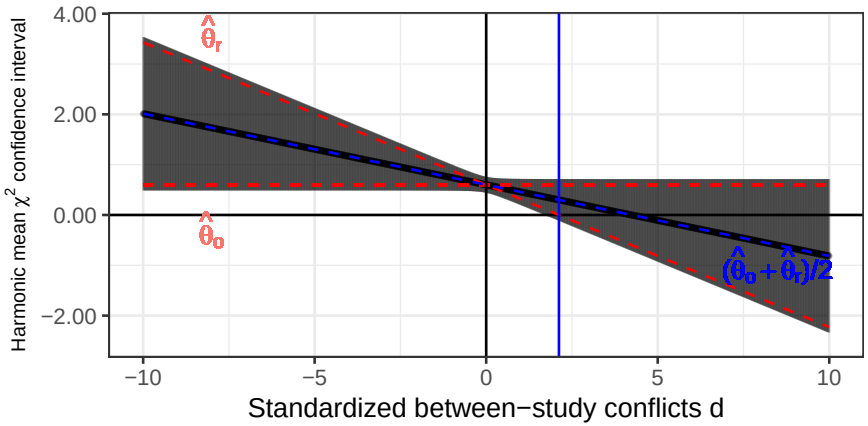
✓No unusual confidence set

– VS. II ($\widetilde{p}_S$ & $\widetilde{CI}_S$)

✓Two-sided test for both positive and negative results

✓Fixed upper bound of p-value: $1/2^{n-1}$

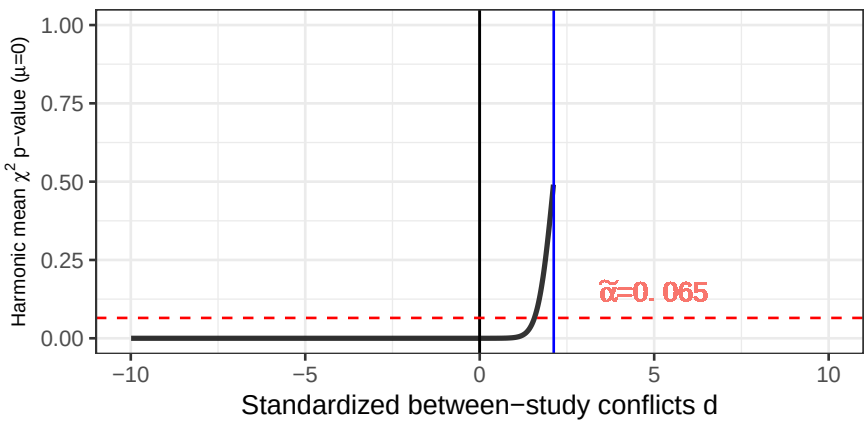
III. Harmonic mean χ^2 methods: p_H & CI_H



✓no replication paradox

✓no unusual confidence set

– narrower than sceptical confidence intervals

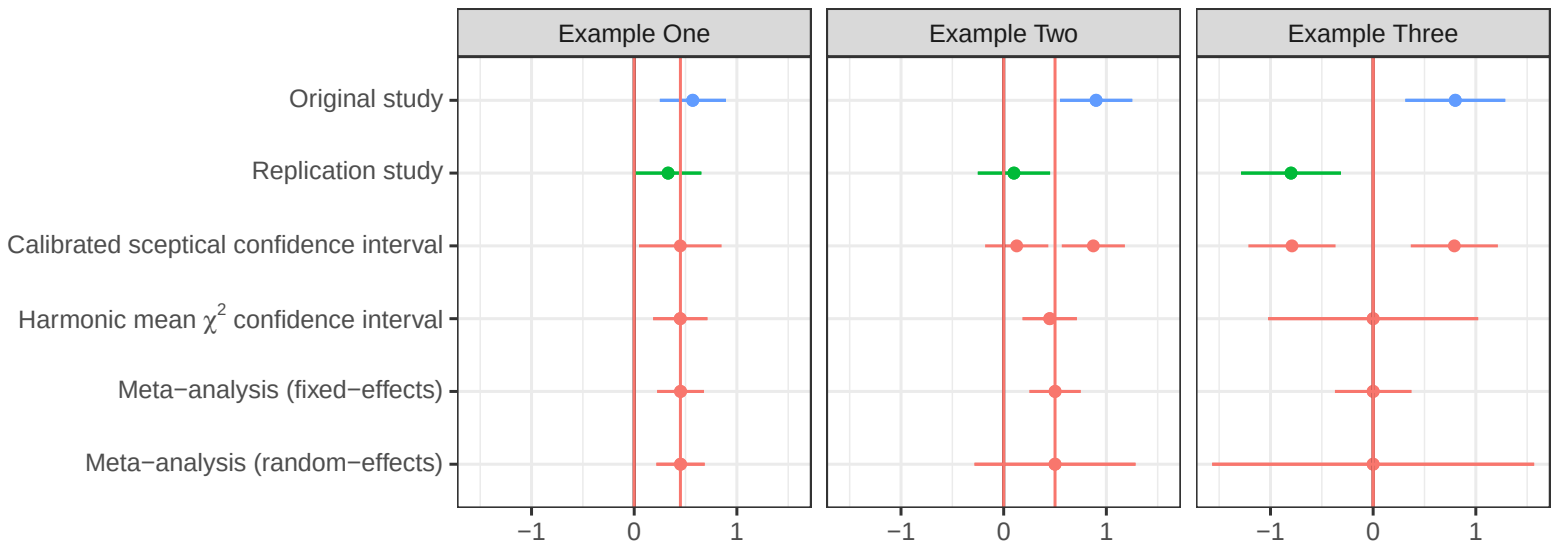


✓no replication paradox

✓direction conflict if no p_H value

✗between-study conflict d

Different methods under different cases



– **Random-effects model** is more appropriate, especially under a large between-study conflict.