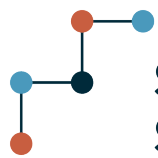
Design and Analysis of Replication Studies

Leonhard Held, Charlotte Micheloud

University of Zurich



Swiss National
Science Foundation

Winter School 2022, Les Diablerets

Introduction

Replication studies

Direct replication

- Repeating original study using the same methodology

→ Tool to assess credibility of scientific discoveries

→ Regulatory requirement

Replication crisis

- Low replicability of many scientific discoveries

→ Large-scale replication projects

Large-scale replication projects

- 2015: Reproducibility project psychology
- 2016: Experimental economics replication project
- 2018: Experimental philosophy replicability project
- 2018: Social sciences replication project
- 2021: Reproducibility Project: Cancer Biology

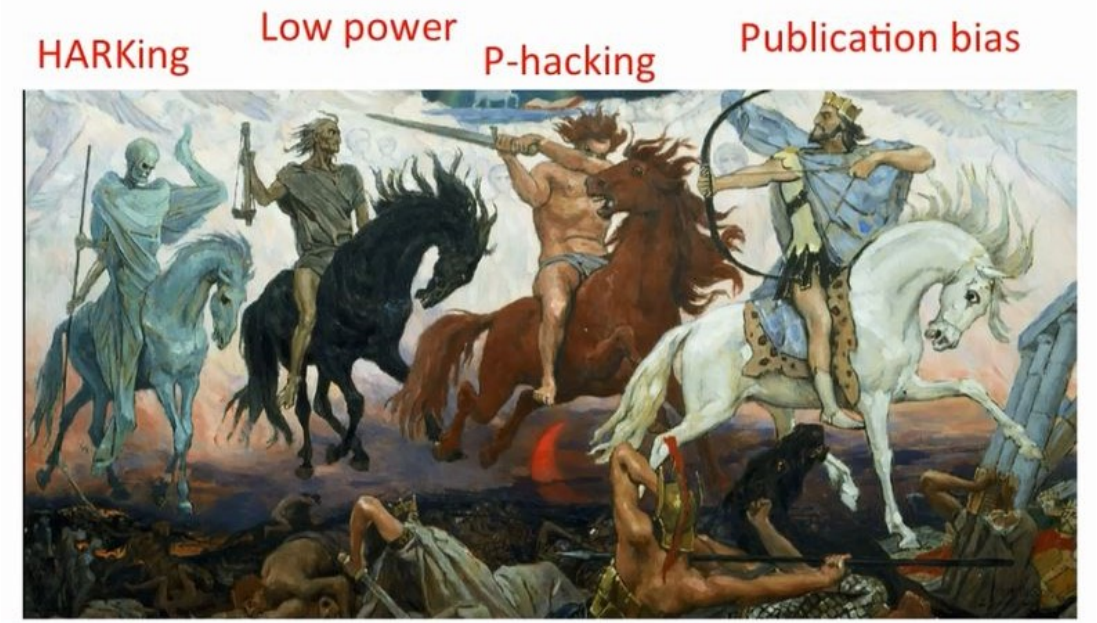
Estimating the reproducibility of psychological science

Open Science Collaboration

Science **349** (6251), aac4716.
DOI: 10.1126/science.aac4716

The four horsemen of irreproducibility

Dorothy Bishop (2019) in Nature



Questionable research practices (QRPs)

Large-scale replication projects

- 2015: Reproducibility project psychology
- 2016: Experimental economics replication project
- 2018: Experimental philosophy replicability project
- 2018: Social sciences replication project
- 2021: Reproducibility Project: Cancer Biology

nature human behaviour

Letter | Published: 27 August 2018

Evaluating the replicability of social science experiments in *Nature* and *Science* between 2010 and 2015

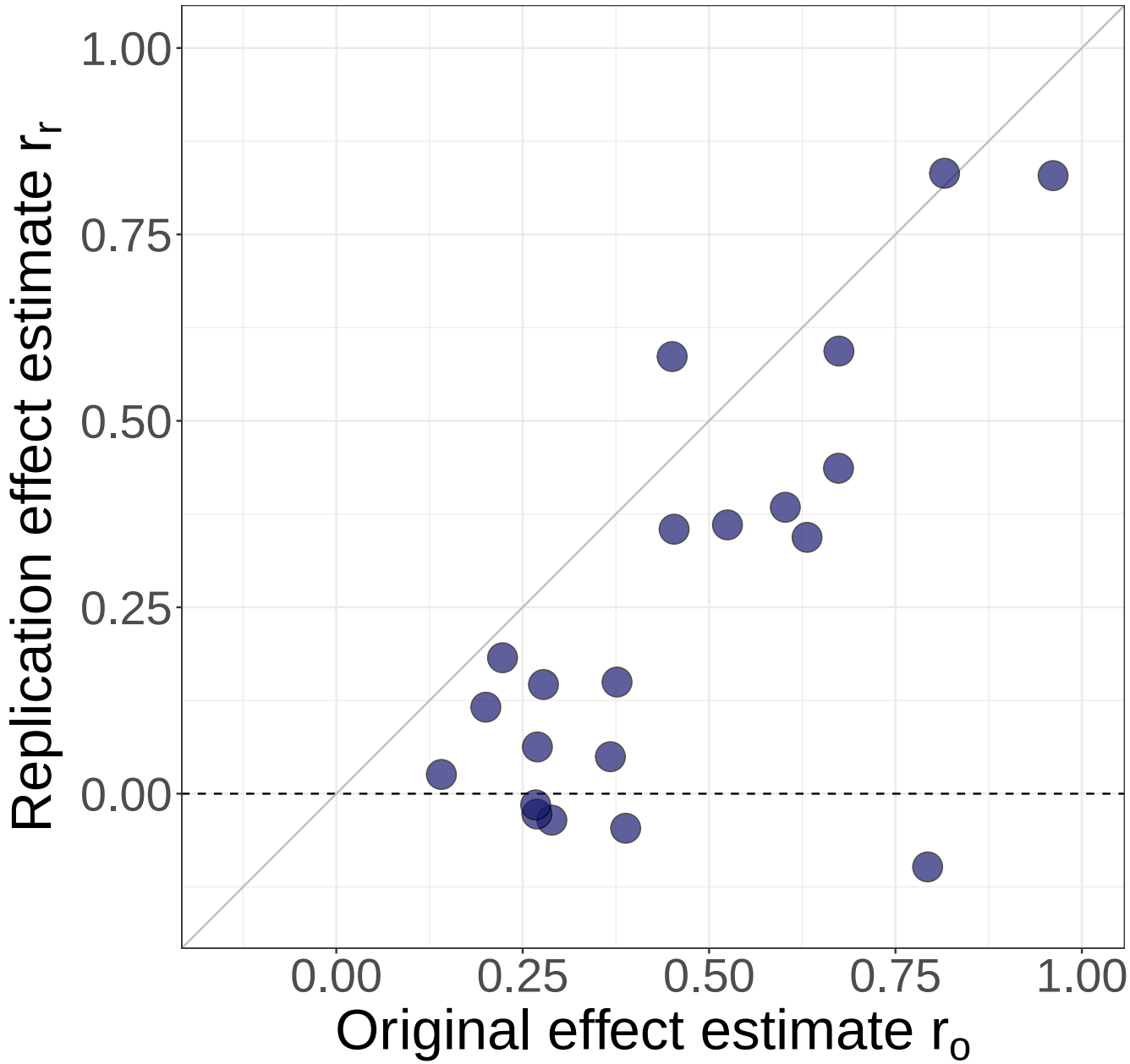
Colin F. Camerer, Anna Dreber, Felix Holzmeister, Teck-Hua Ho, Jürgen Huber, Magnus Johannesson, Michael Kirchler, Gideon Nave, Brian A. Nosek ✉, Thomas Pfeiffer, Adam Altmejd, Nick Buttrick, Taizan Chan, Yiling Chen, Eskil Forsell, Anup Gampa, Emma Heikensten, Lily Hummer, Taisuke Imai, Siri Isaksson, Dylan Manfredi, Julia Rose, Eric-Jan Wagenmakers & Hang Wu

Social sciences replication project

```
library(ReplicationSuccess)
data("RProjects")
social <- subset(RProjects,
                 project == "Social Sciences")
```

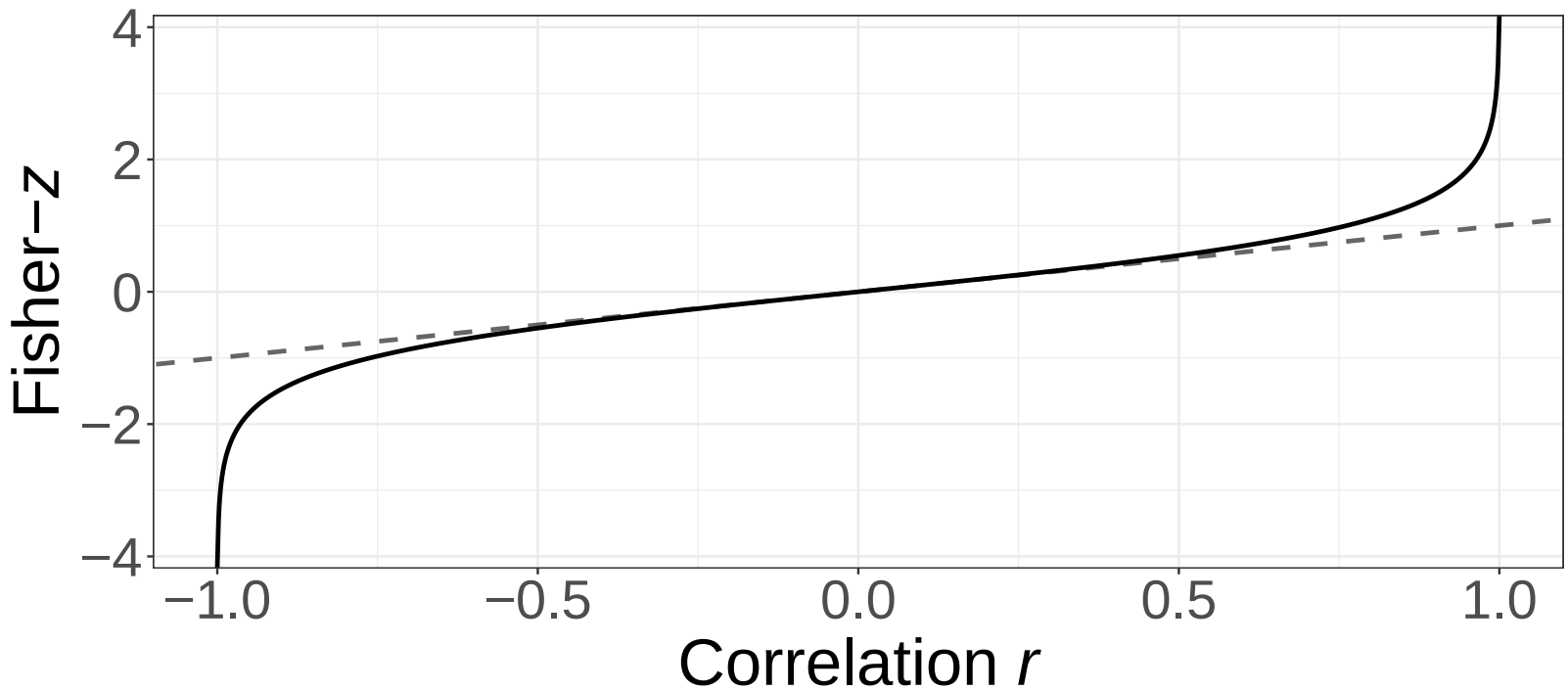
Social sciences replication project

Correlation scale



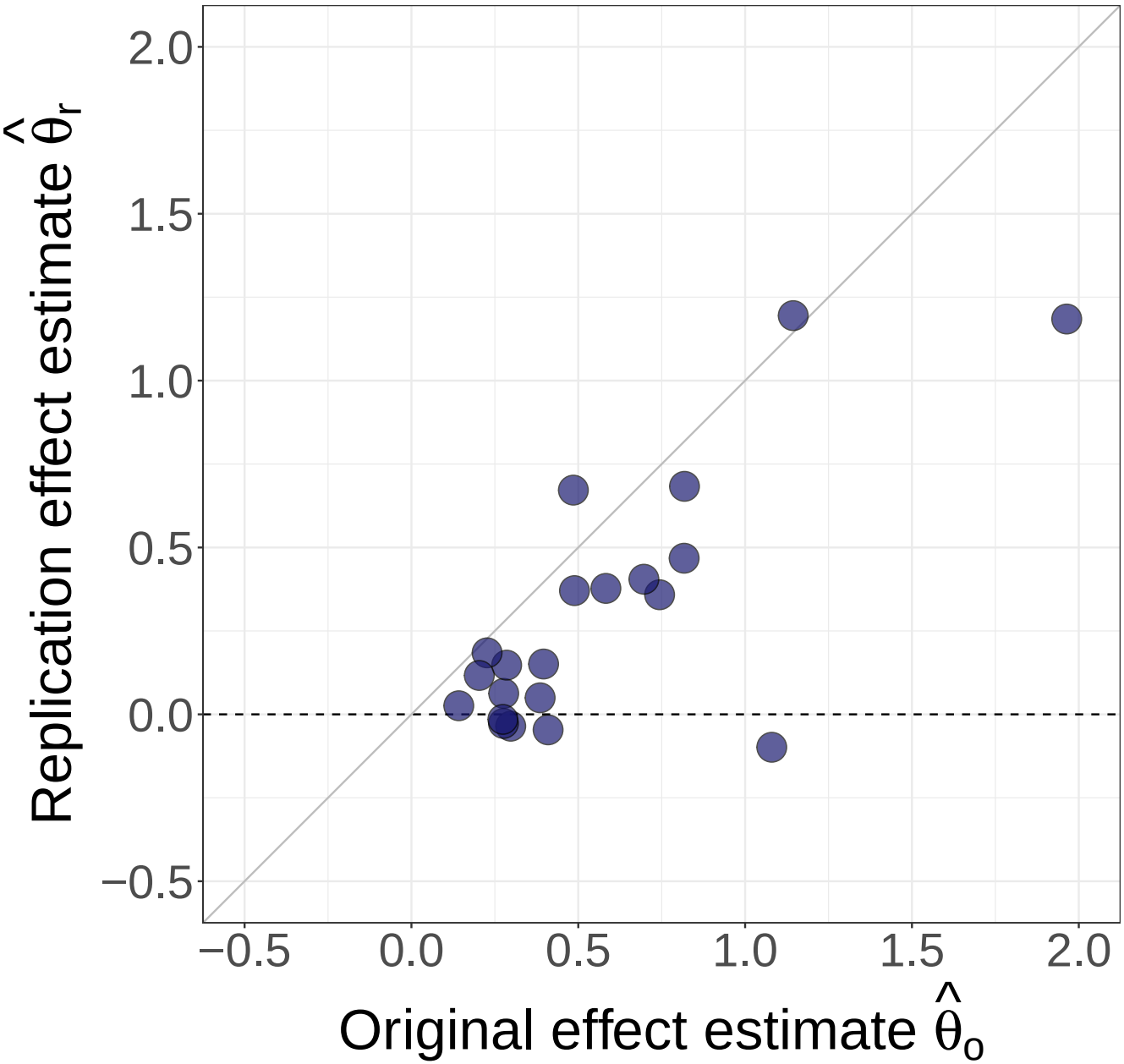
Statistical framework of package

– Effect estimates are assumed to be normally distributed after suitable transformation
→ Fisher's z-transformation for correlation coefficients r with (effective) sample size $n - 3$



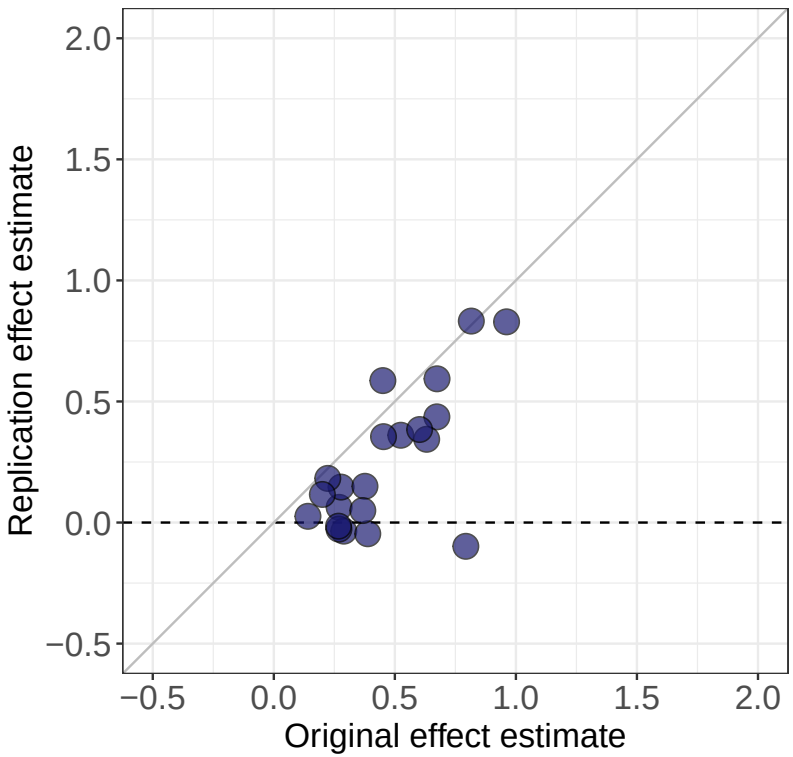
Social sciences replication project

Fisher-*z* scale

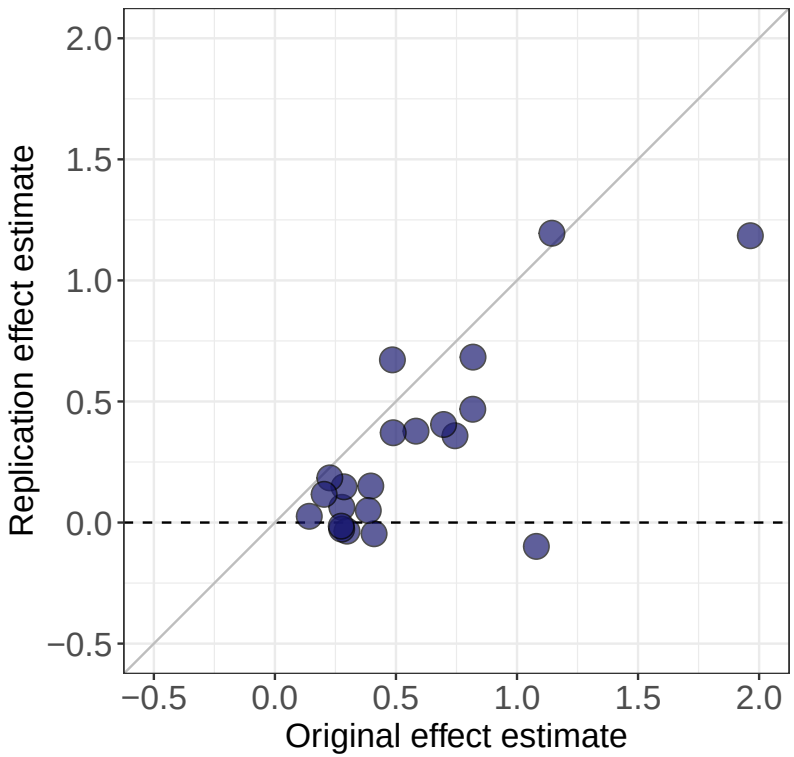


Correlation vs Fisher-z scale

Correlation scale

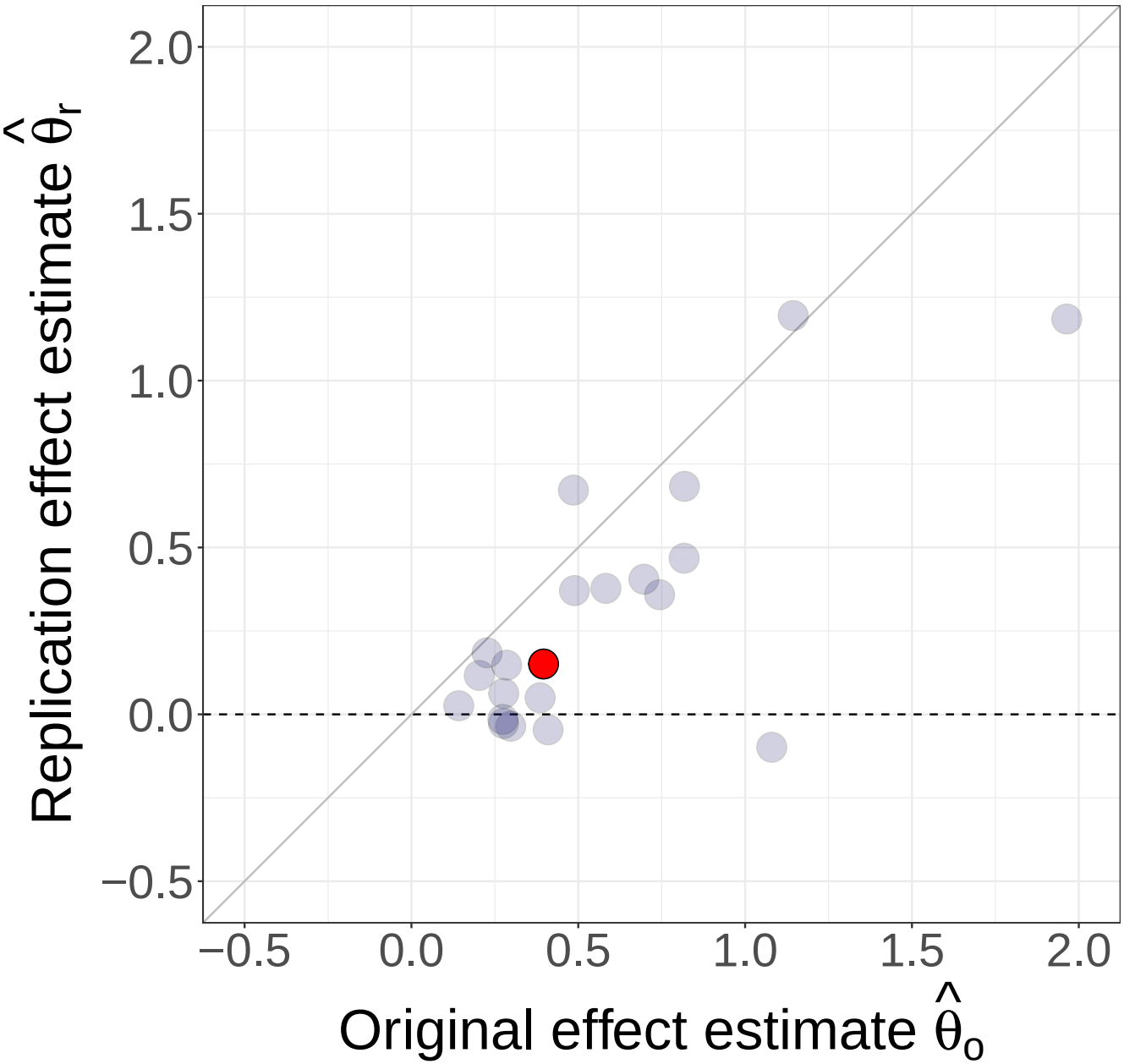


Fisher−z scale



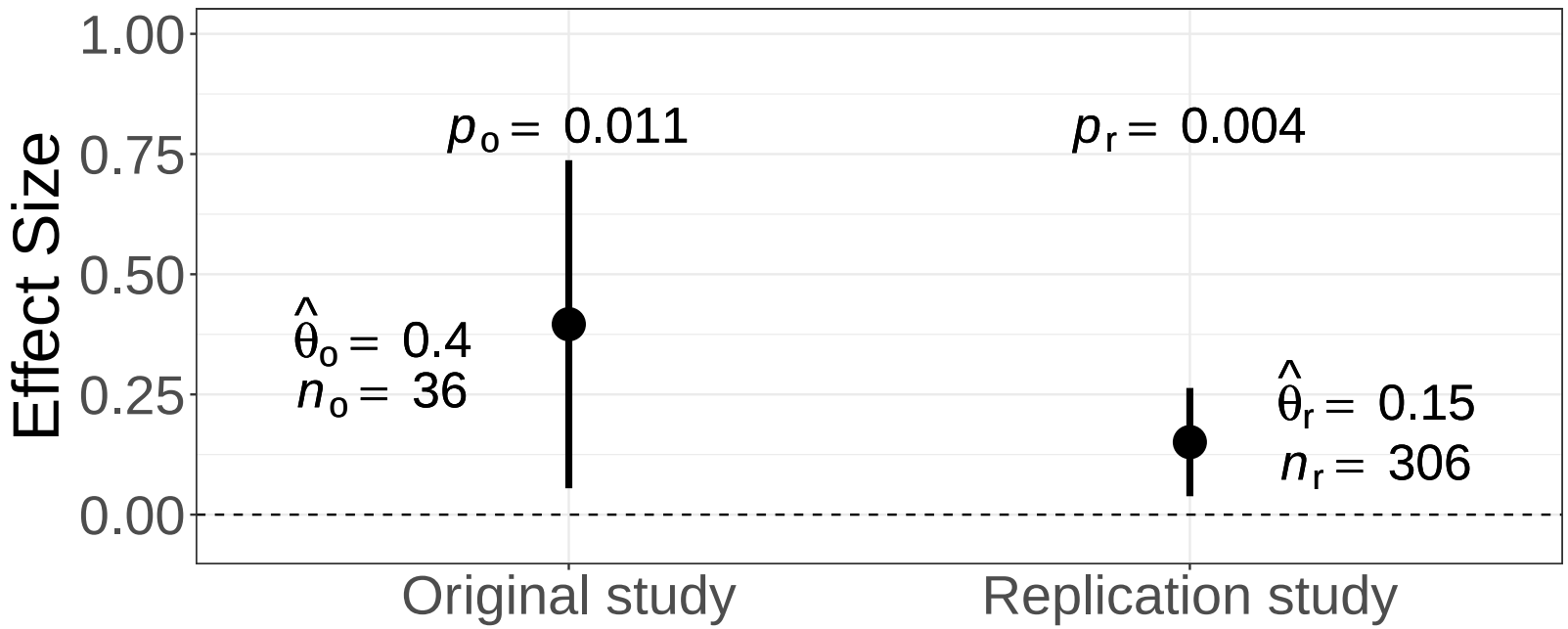
Social sciences replication project

Pyc and Rawson (2010), Science



Pyc and Rawson (2010). Science

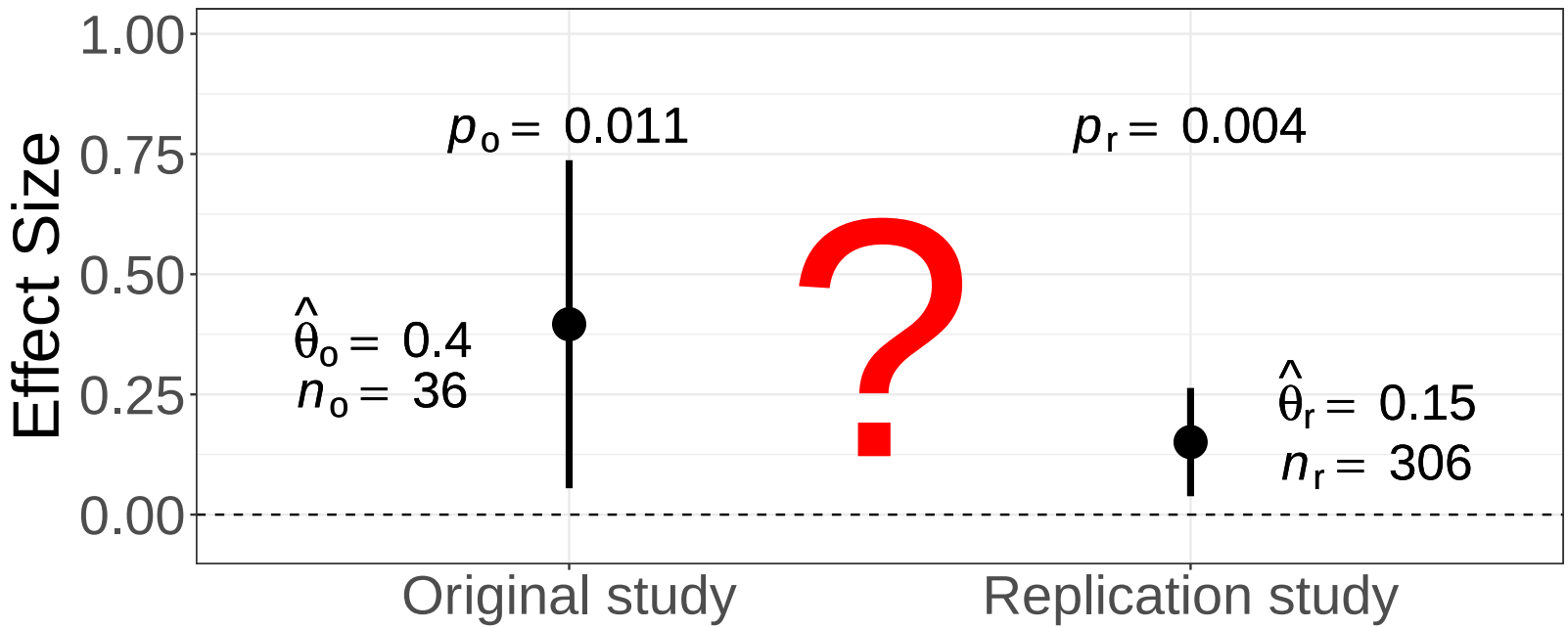
Original discovery: “Testing improves memory”



n_o	Sample size	n_r
$\hat{\theta}_o$	Effect estimate	$\hat{\theta}_r$
σ_o	Standard error	σ_r
p_o	one-sided *p*-value	p_r

Pyc and Rawson (2010). Science

Original discovery: "Testing improves memory"



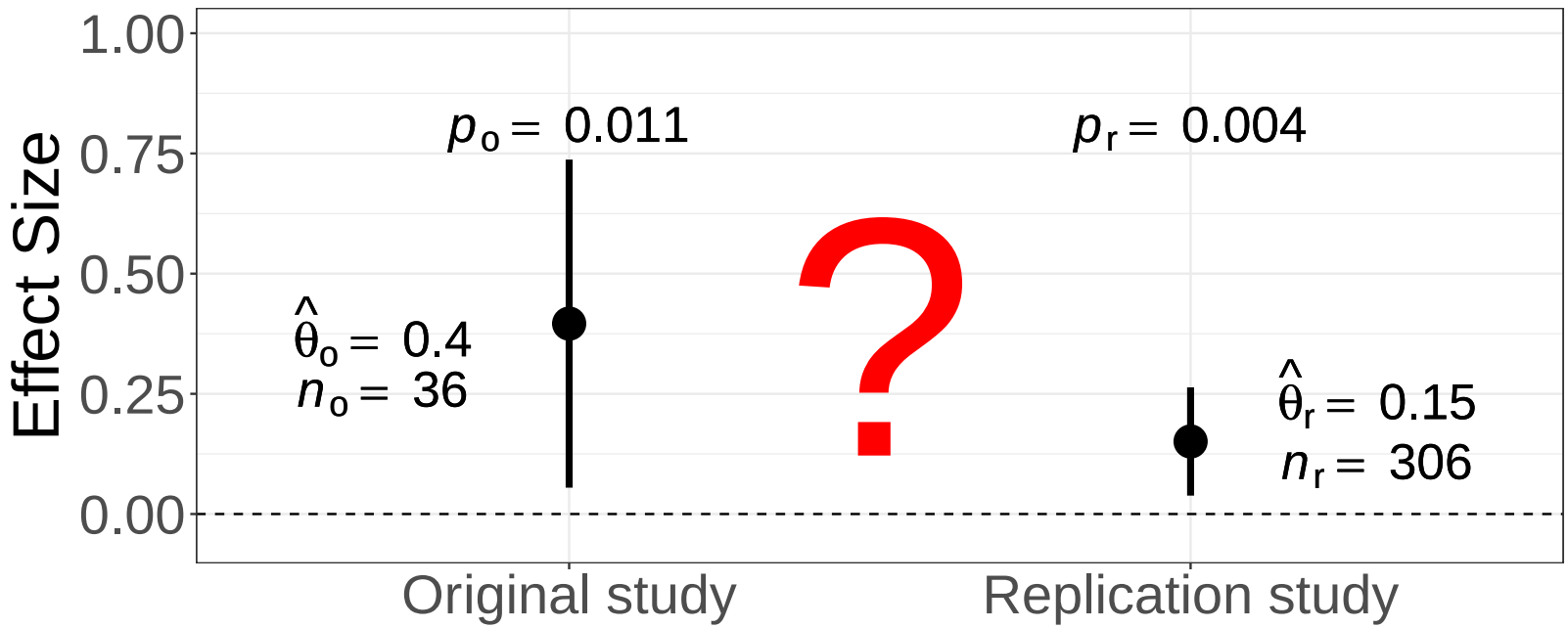
Relative effect size $d = \hat{\theta}_r / \hat{\theta}_o = 0.38$

Relative sample size $c = n_r / n_o = 9$

When is a replication successful?

Some proposed criteria

1. Two-trials rule (statistical significance)
2. Compatibility of effect estimates
3. Meta-analysis of estimates
4. Sceptical *p*-value

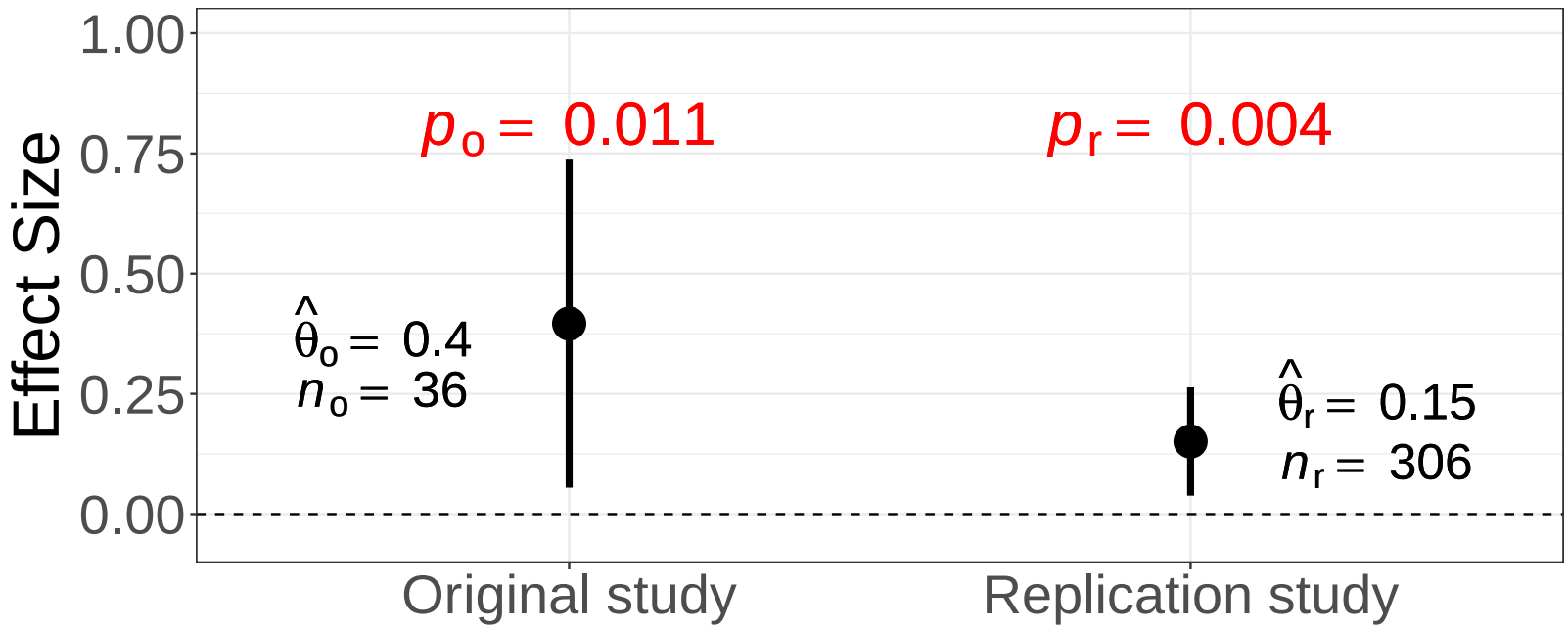


1. Two-trials rule

Are both estimates statistically significant in the same direction?

$\rightarrow$ Which threshold?

$\rightarrow$ one-sided $\alpha = 0.025$

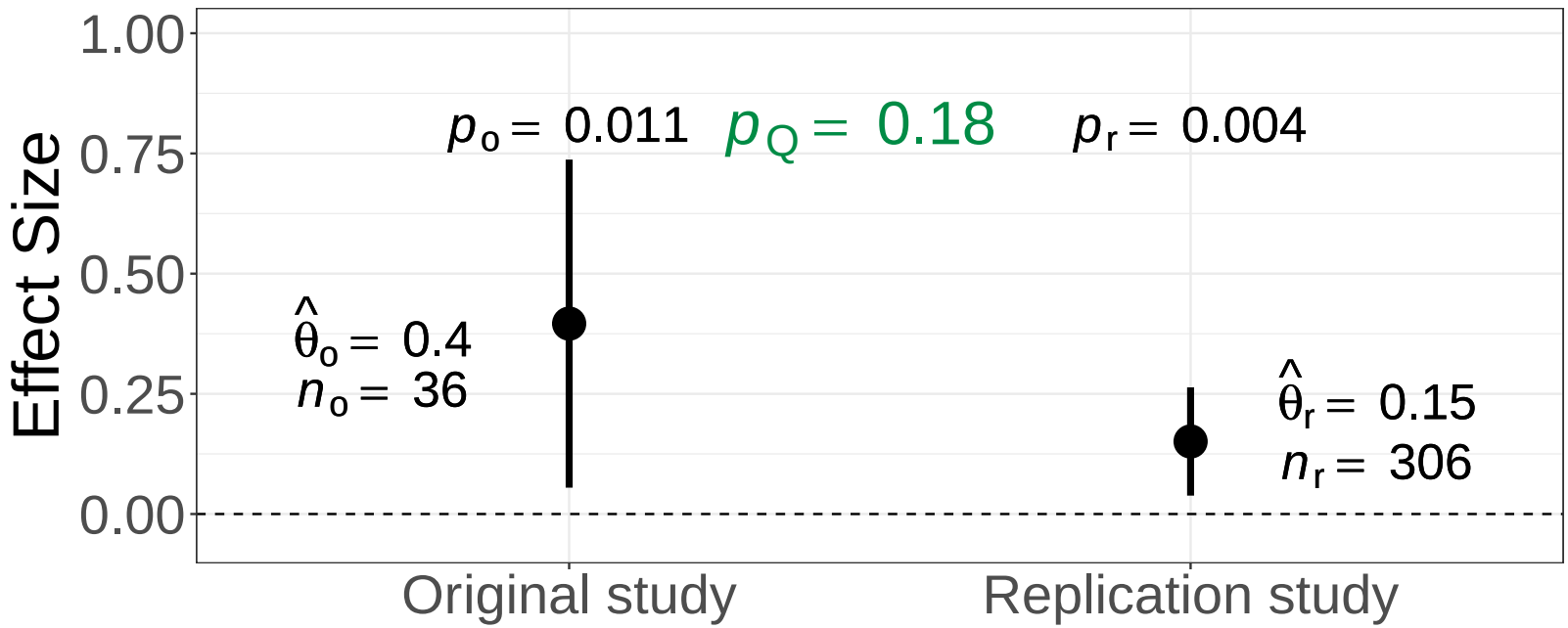


2. Compatibility of effect estimates

Is the meta-analytic Q-test of the estimates statistically significant?

→ Which threshold?

→ two-sided $\alpha = 0.05$

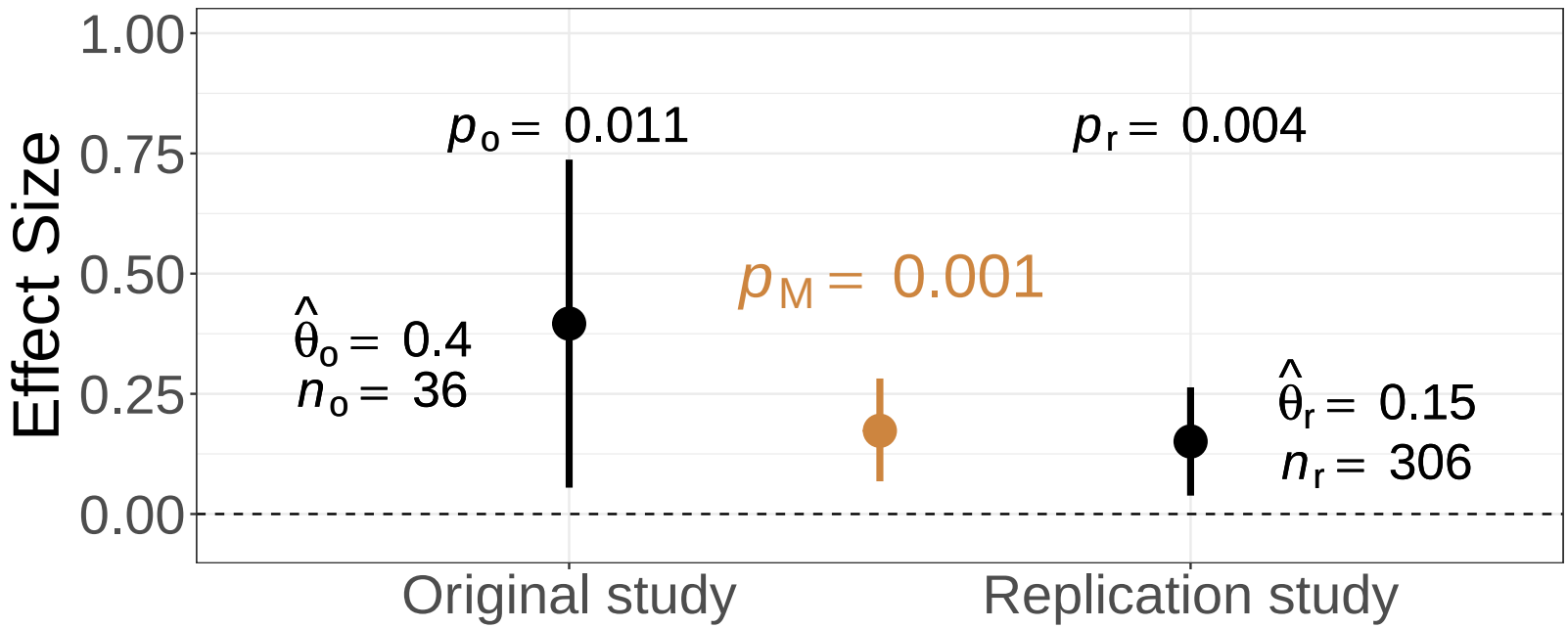


3. Meta-analysis of effect estimates

Is a meta-analytic estimate statistically significant?

$\rightarrow$ Which threshold?

$\rightarrow$ one-sided $\alpha = 0.025$ or 0.0025 used in practice



4. Sceptical *p*-value

New definition of replication success

J. R. Statist. Soc. A (2020)
183, *Part* 2, *pp.* 431–448

A new standard for the analysis and design of replication studies

Leonhard Held

University of Zurich, Switzerland

https://doi.org/10.1111/rssa.12493

THE ASSESSMENT OF REPLICATION SUCCESS BASED ON RELATIVE EFFECT SIZE

By Leonhard Held, Charlotte Micheloud and Samuel Pawel

Epidemiology, Biostatistics and Prevention Institute, Center for Reproducible Science, University of Zurich,
leonhard.held@uzh.ch; charlotte.micheloud@uzh.ch; samuel.pawel@uzh.ch

Replication studies are increasingly conducted in order to confirm original findings. However, there is no established standard how to assess replication success and in practice many different approaches are used. The purpose of this paper is to refine and extend a recently proposed reverse-Bayes approach for the analysis of replication studies. We show how this method is directly related to the relative effect size, the ratio of the replication to the original effect estimate. This perspective leads to a new proposal to recalibrate the assessment of replication success, the golden level. The recalibration ensures that for borderline significant original studies replication success can only be achieved if the replication effect estimate is larger than the original one. Conditional power for replication success can then take any desired value if the original study is significant and the replication sample size is large enough. Compared to the standard approach to require statistical significance of both the original and replication study, replication success at the golden level offers uniform gains in project power and controls the Type-I error rate if the replication sample size is not smaller than the original one. An application to data from four large replication projects shows that the new approach leads to more appropriate inferences, as it penalizes shrinkage of the replication estimate compared to the original one, while ensuring that both effect estimates are sufficiently convincing on their own.

https://arxiv.org/abs/2009.07782

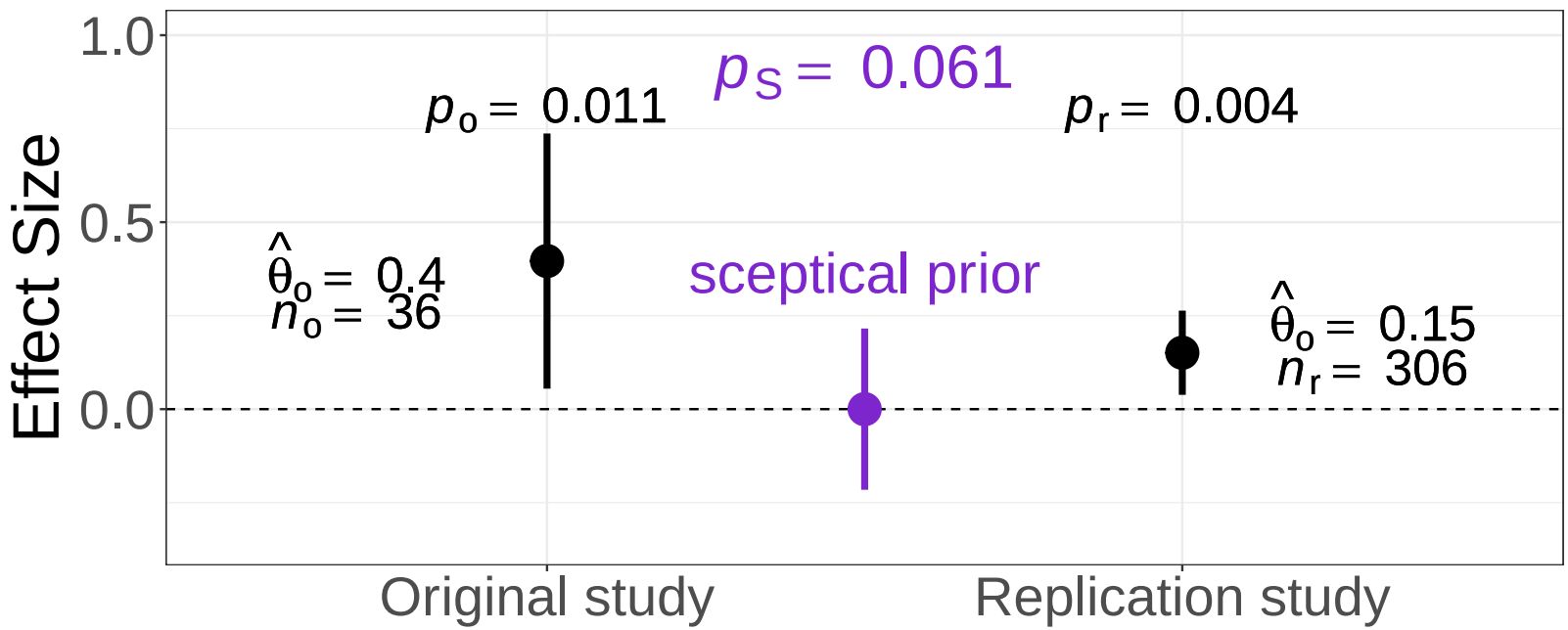
AOAS, to appear

4. Sceptical *p*-value

New definition of replication success

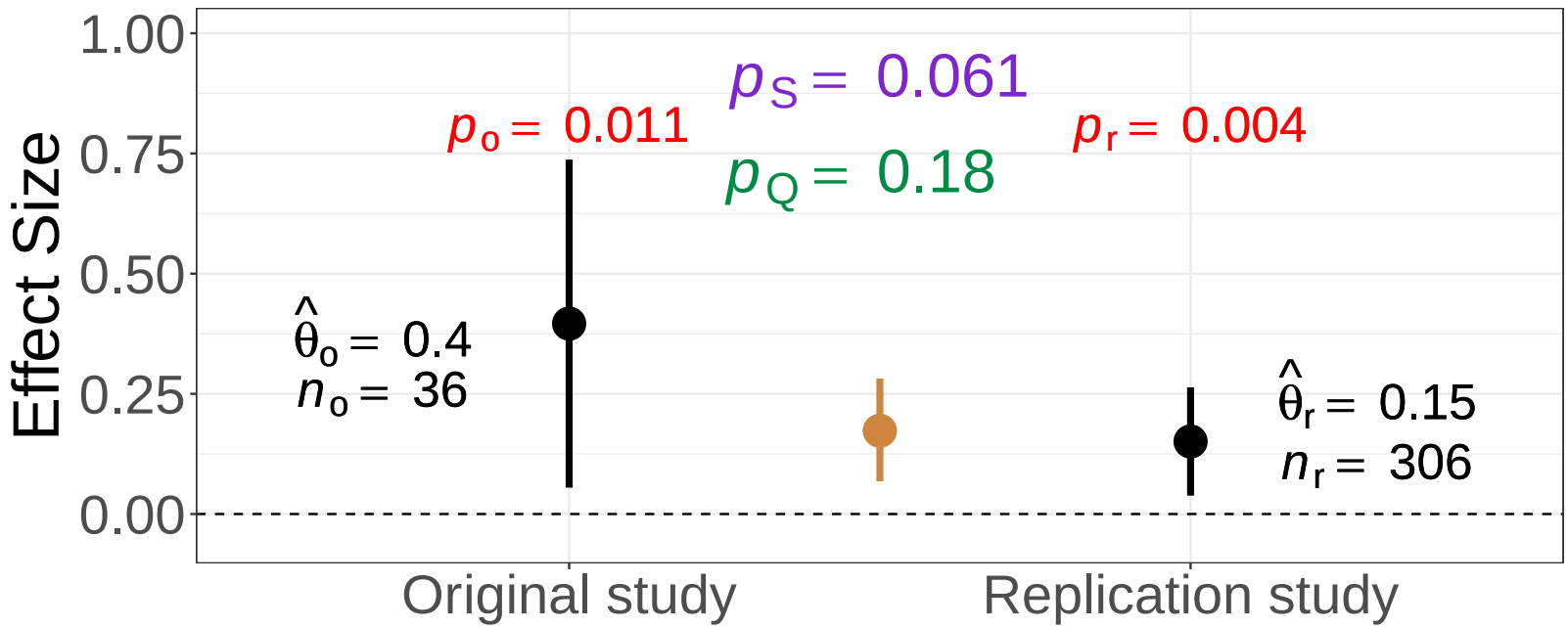
Can we convince a sceptic whose priof beliefs make the original study not significant?

If $p_S \leq \alpha$ we have replication success at level α

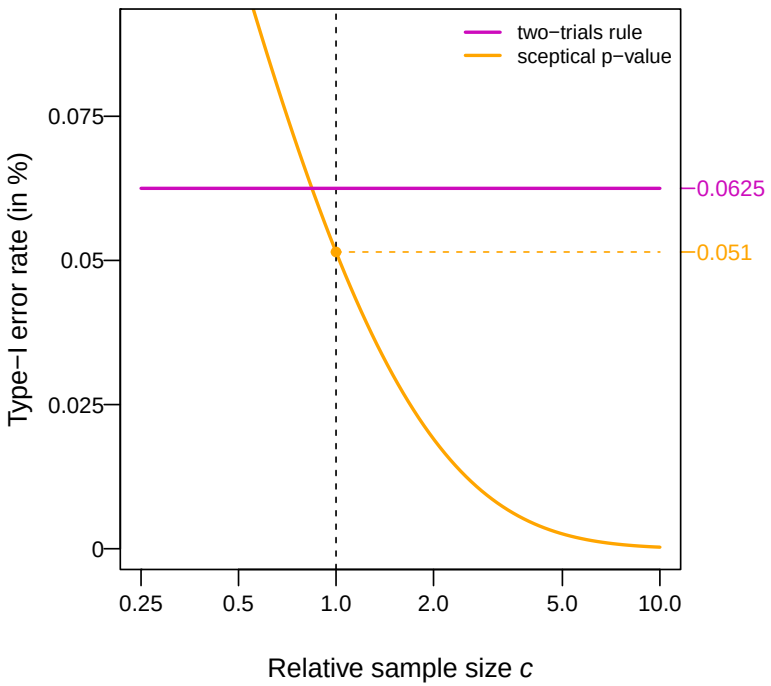


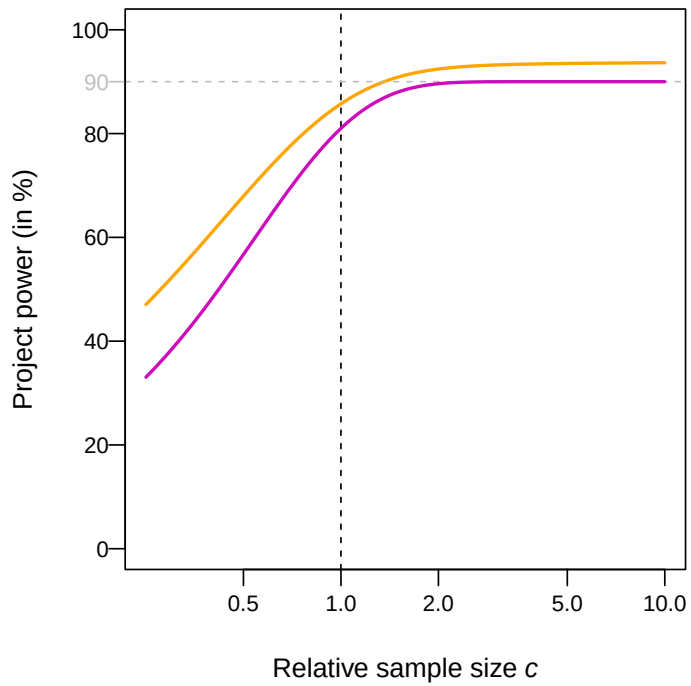
Assessement of replication success

- Two-trials rule doesn't take into account effect size
- *Q*-test doesn't take into account significance
- Meta-analysis assumes exchangeability
- Sceptical *p*-value takes into account effect size and significance



Type-I error rate and project power





The relative effect size

For the assessment of replication success

- Relative effect size $d = \hat{\theta}_r / \hat{\theta}_o$ can be used in the assessment of replication success

  $\rightarrow$ success if $d > d_{\text{min}}$, the minimum relative effect size
- d_{min} depends on
  - relative sample size c
  - original p-value p_o
  - one-sided level α
- Equivalent to assessment based on (sceptical) p-value

The relative effect size

For the assessment of replication success

Pyc and Rawson (2010) example

- $d = 0.15/0.4 = 0.38$
- Minimum relative effect size
  - With the two-trials rule:

    $d_{\text{min}} = 0.28 < 0.38$ Replication Success
  - With the sceptical p-value:

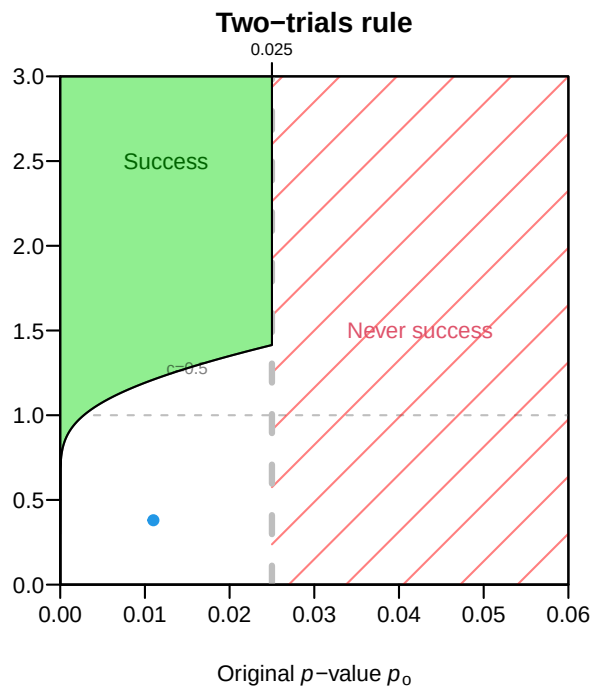
    $d_{\text{min}} = 0.66 > 0.38$ No Replication Success

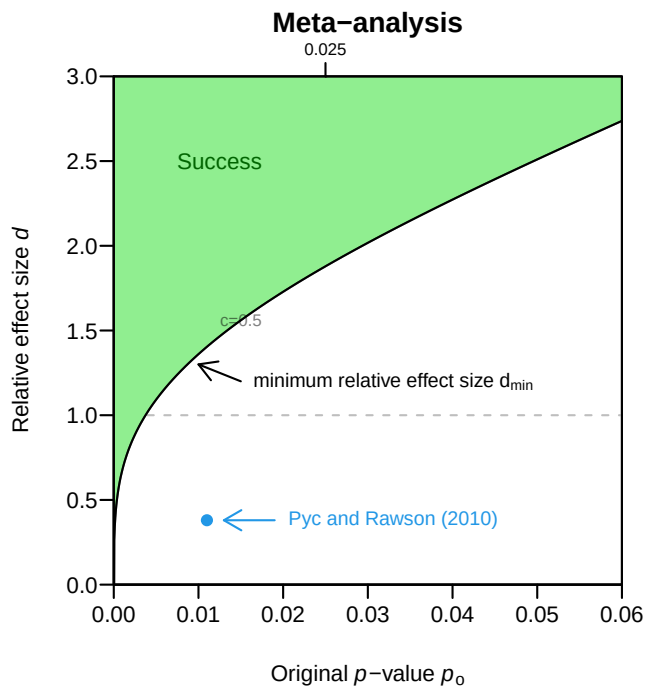
Why do the two methods disagree?

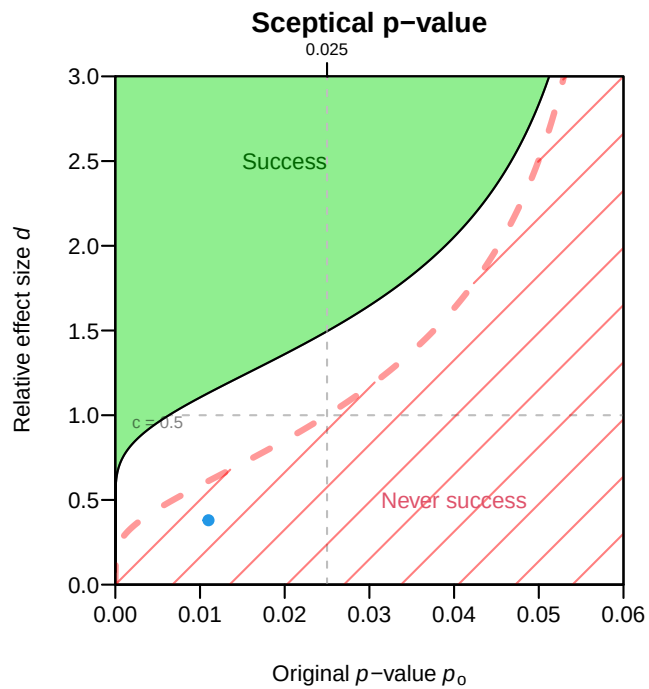
Comparison of success regions

Relative sample size $c = 0.5$

Success if $d > d_{\text{min}}$



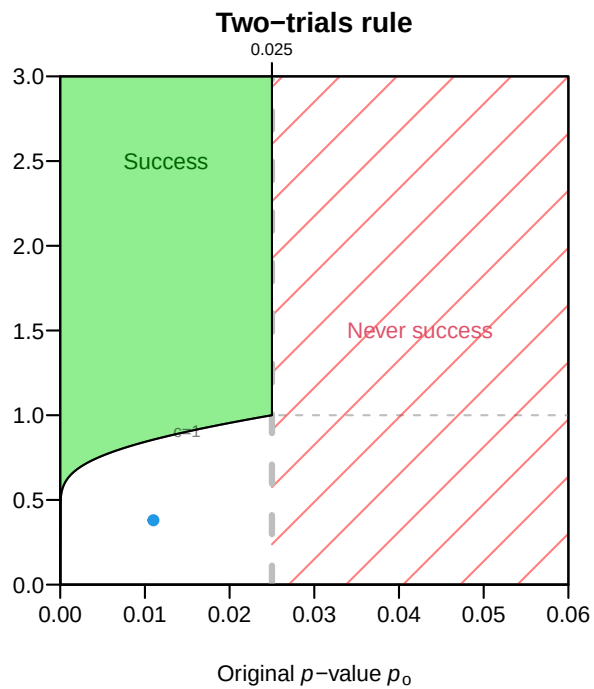


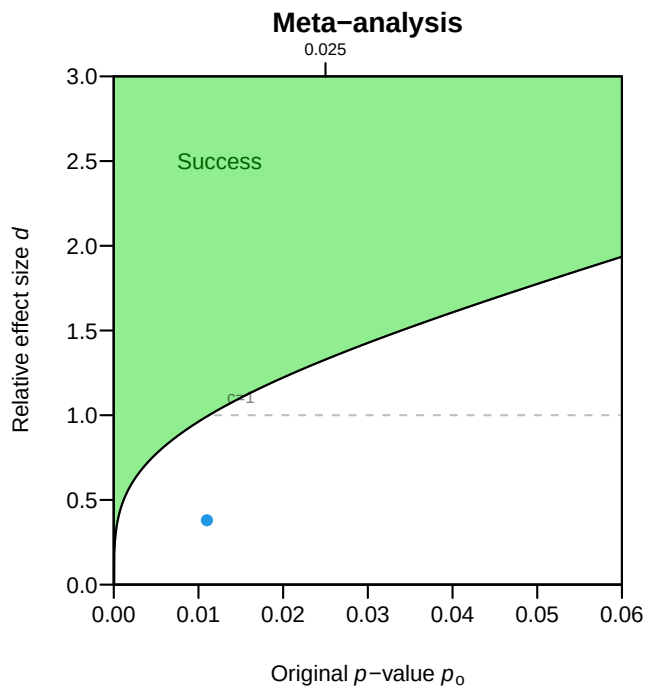


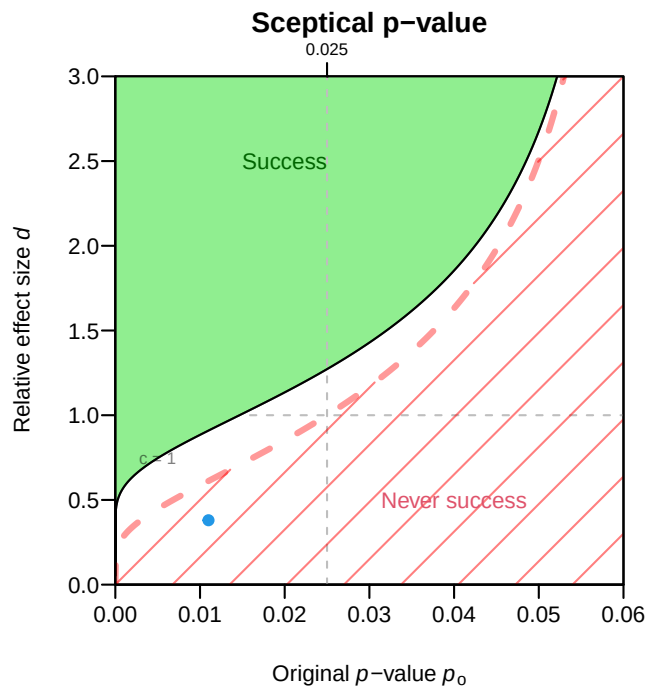
Comparison of success regions

Relative sample size $c = 1$

Success if $d > d_{\text{min}}$



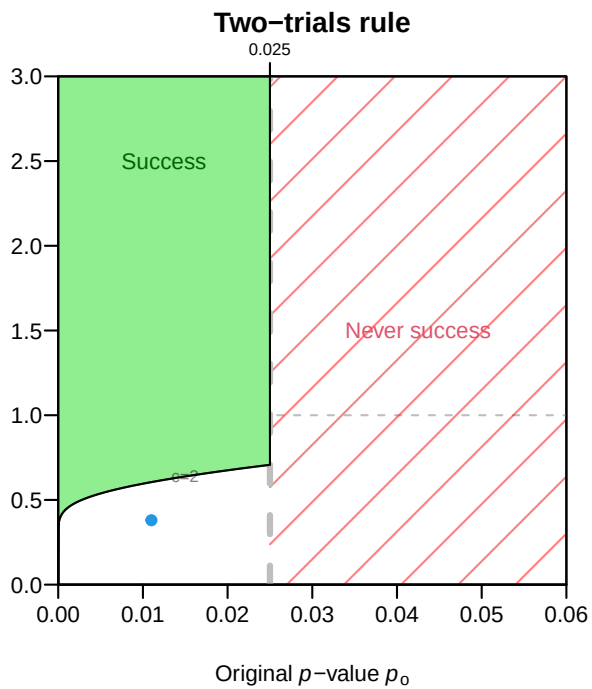


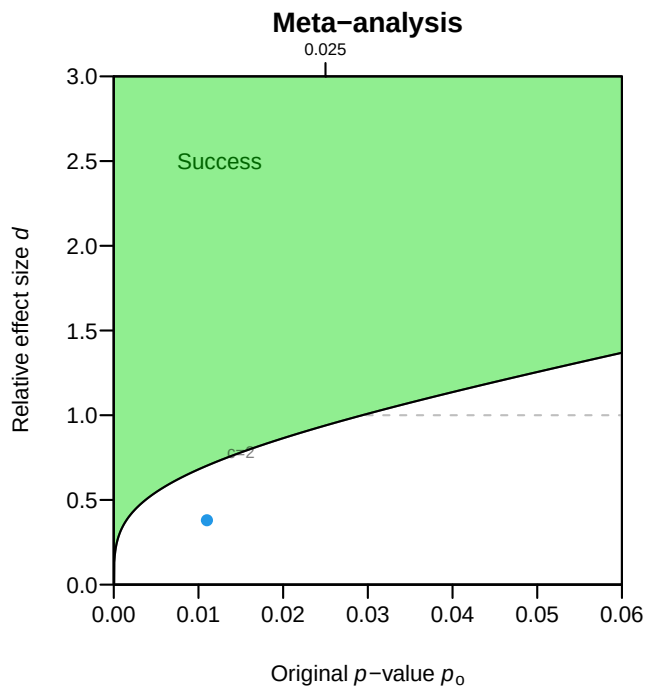


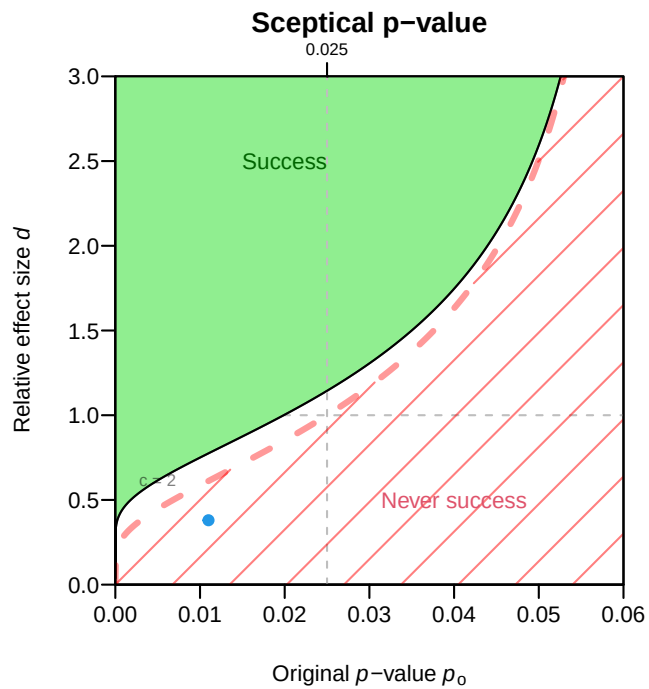
Comparison of success regions

Relative sample size $c = 2$

Success if $d > d_{\text{min}}$



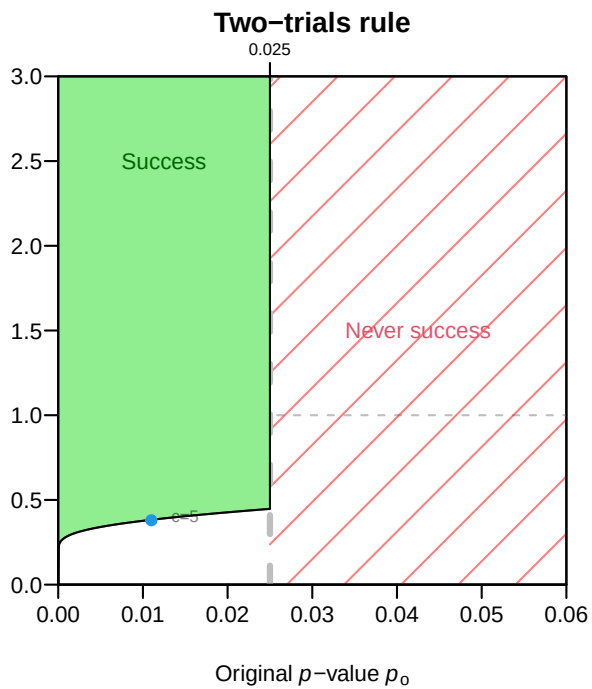


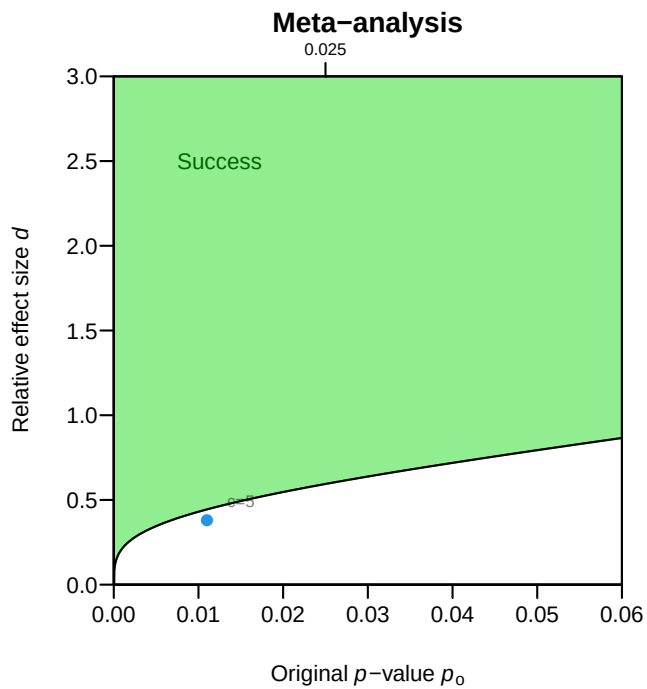


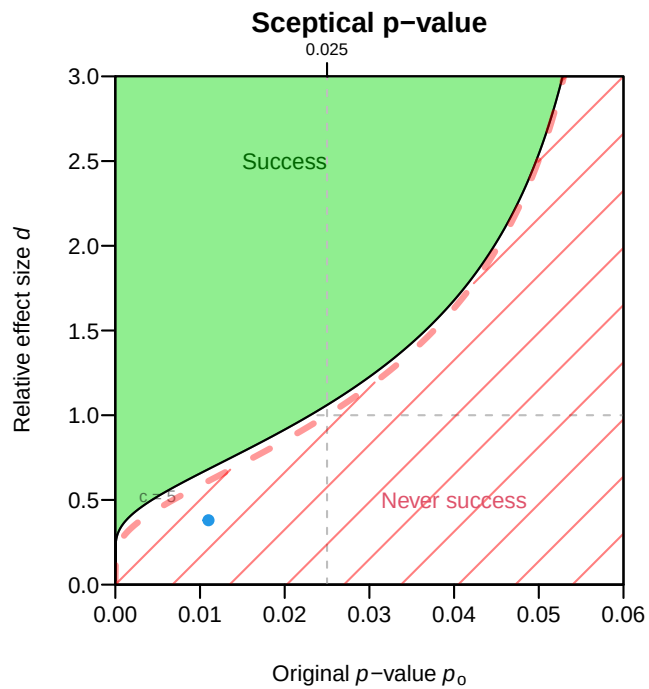
Comparison of success regions

Relative sample size $c = 5$

Success if $d > d_{\text{min}}$



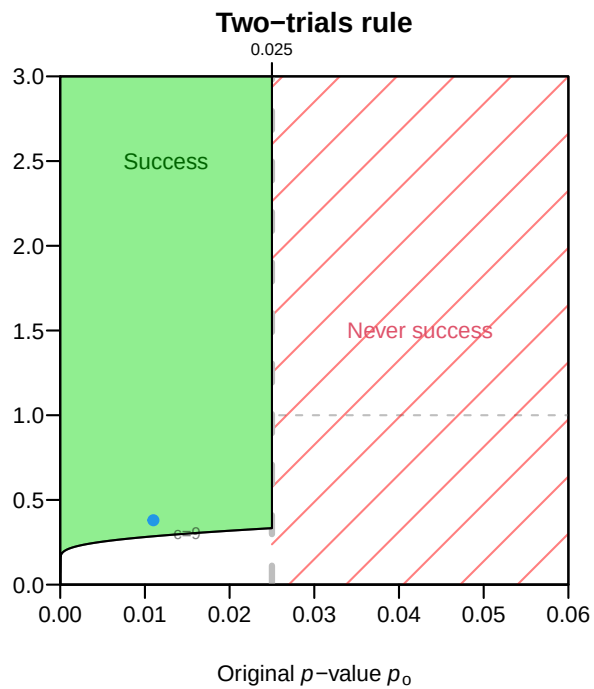


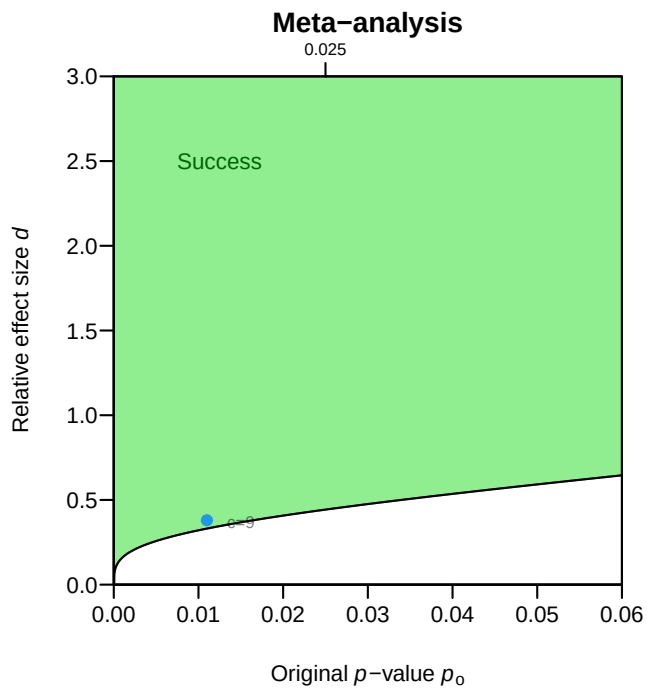


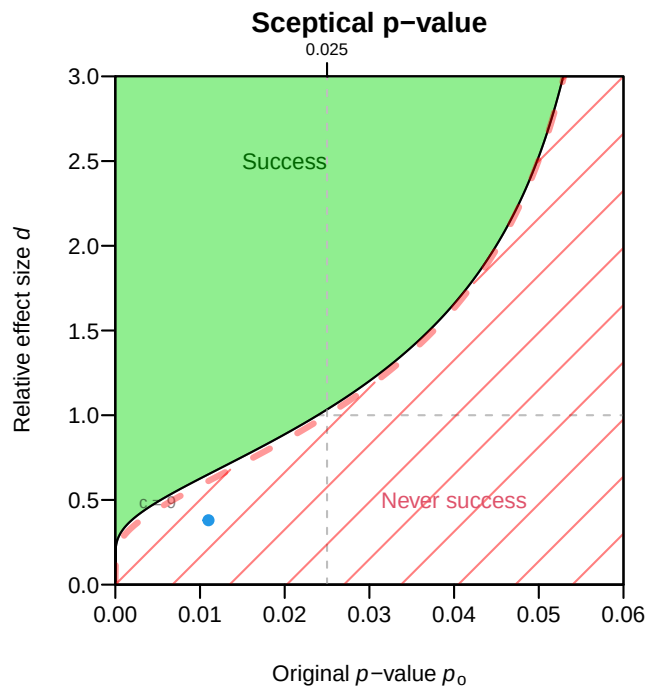
Comparison of success regions

Relative sample size $c = 9$

Success if $d > d_{\text{min}}$



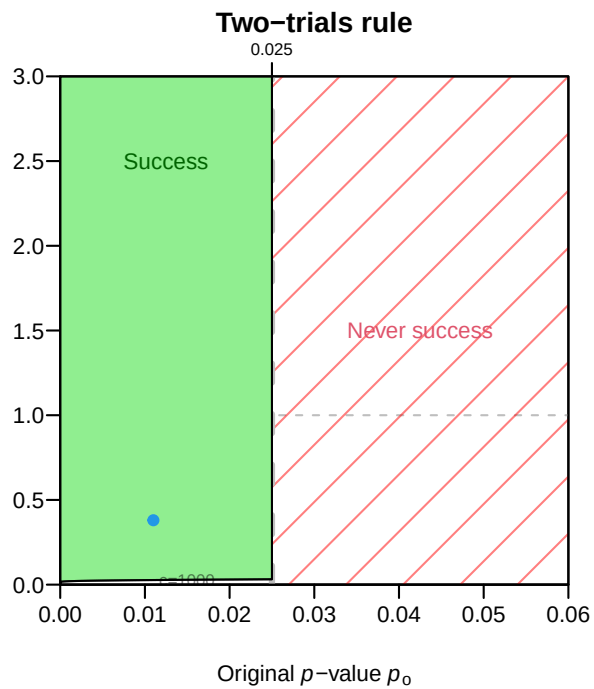


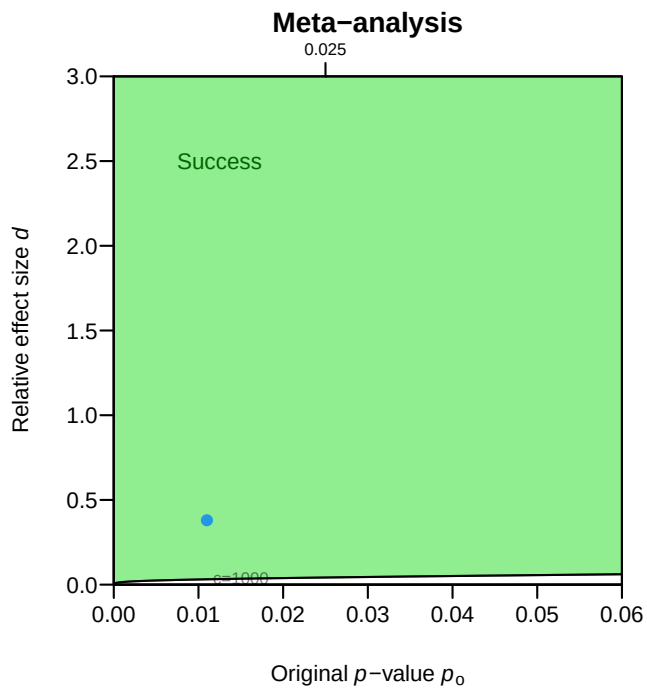


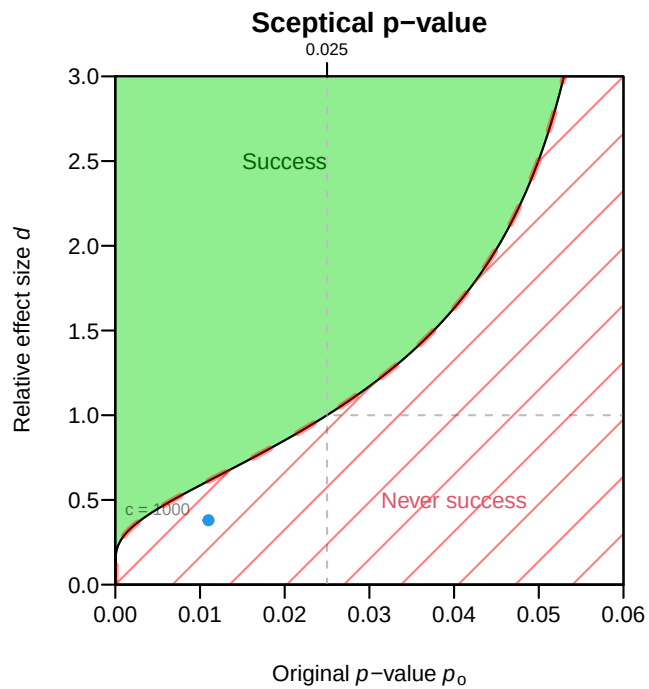
Comparison of success regions

Relative sample size $c = 1000$

Success if $d > d_{\text{min}}$







Shrinkage of effect estimates

Definition and example

Shrinkage

- represents how much the replication effect estimate is decreased as compared to the original effect estimate:

$$\boxed{s = 1 - d}$$

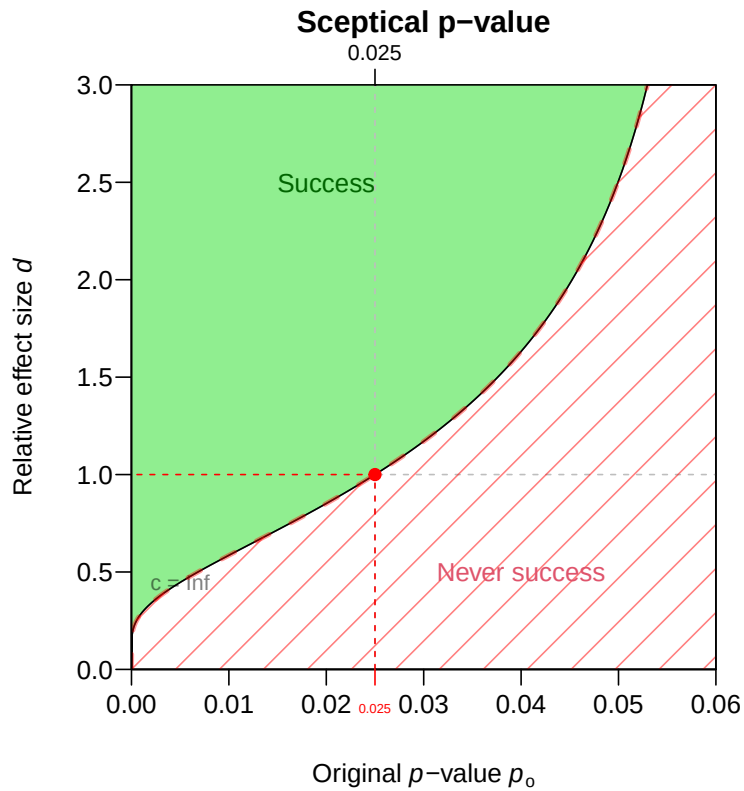
- Pyc and Rawson example: $s = 1 - 0.38 = 0.62$

  $\rightarrow$ replication effect is 62% smaller than original one

Shrinkage is penalized in the sceptical *p*-value approach

Calibration of the sceptical p-value

Bordeline significant original studies ($p_o = \alpha$) can only lead to success if there is no shrinkage ($d > 1$)



Exercise Session 1

Analysis of replication studies

Package ReplicationSuccess

- Installation

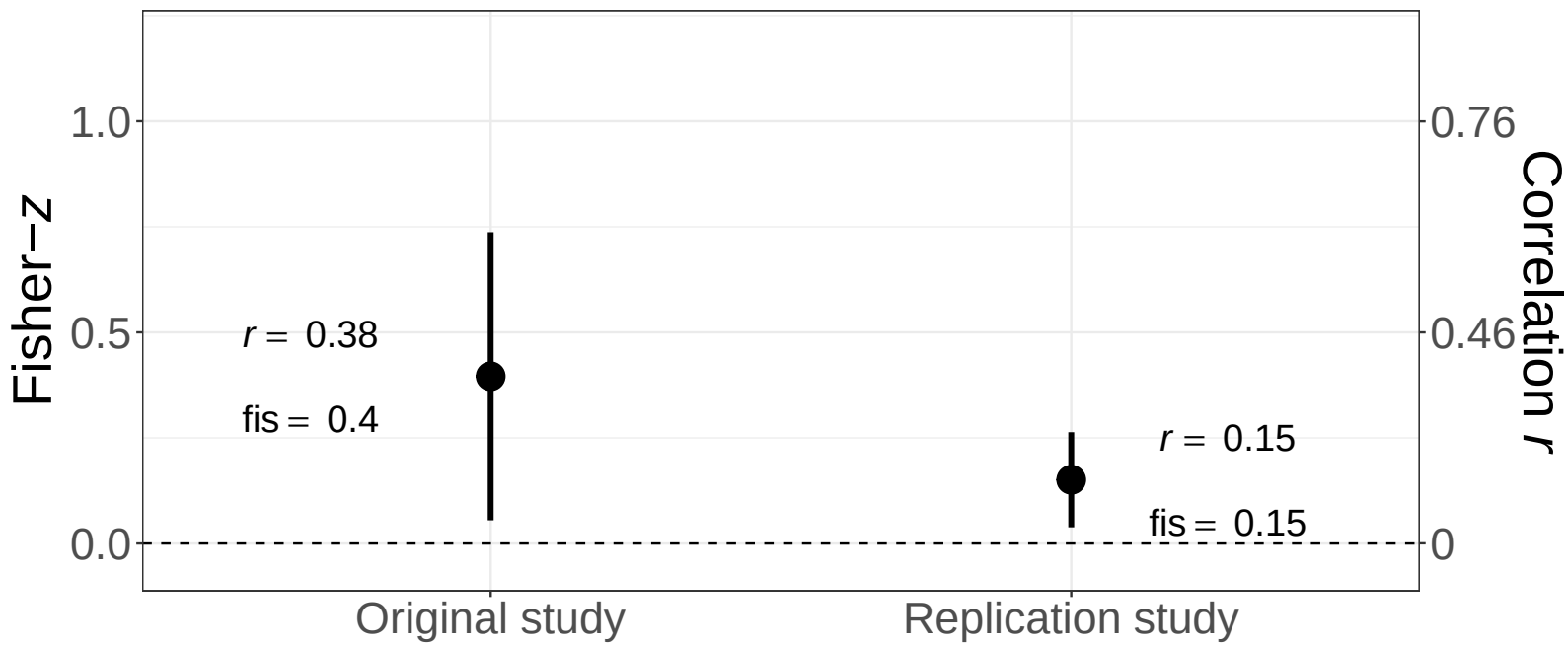
```
install.packages("ReplicationSuccess")
```

- Usage

```
library(ReplicationSuccess)
vignette("ReplicationSuccess")
?pSceptical # documentation
news(package = "ReplicationSuccess") # news page
```

Statistical framework

– Effect estimates are assumed to be normally distributed after suitable transformation
  → Fisher's z-transformation for correlation coefficients r with (effective) sample size $n - 3$



Data sets

```
?RProjects # Documentation
```

Most important variables

`project`	Replication project
`ro`	Original effect on correlation scale
`rr`	Replication effect on correlation scale
`fiso`	Original effect on Fisher-z scale
`fisr`	Replication effect on Fisher-z scale
`se_fiso`	Standard error of `fiso`
`se_fisr`	Standard error of `fisr`

Statistical framework of package

Key quantities

- z-value z_o or (one-sided) p-value p_o of original study

```
RProjects$zo <- RProjects$fiso/RProjects$se_fiso
RProjects$po1 <- z2p(RProjects$zo,
                     alternative = "one.sided")
```

- z-value z_r or (one-sided) p-value p_r of replication study

```
RProjects$zr <- RProjects$fisr/RProjects$se_fisr
RProjects$pr1 <- z2p(RProjects$zr,
                     alternative = "one.sided")
```

- relative sample size (or variance ratio)
  $c = \sigma_o^2/\sigma_r^2 = n_r/n_o$

```
RProjects$c <- RProjects$se_fiso^2/RProjects$se_fisr^2
```

Exercises

(Solutions: https://gitlab.uzh.ch/charlotte.micheloud/replicationstudies)

Load the package and the data sets with

```
library(ReplicationSuccess)
data("RProjects")
```

Compute the key quantities z_o, z_r, c, and the one-sided p-values p_o and p_r with

```
RProjects$zo <- RProjects$fiso/RProjects$se_fiso
RProjects$zr <- RProjects$fisr/RProjects$se_fisr
RProjects$c <- RProjects$se_fiso^2/RProjects$se_fisr^2
RProjects$po1 <- z2p(RProjects$zo,
                     alternative = "one.sided")
RProjects$pr1 <- z2p(RProjects$zr,
                     alternative = "one.sided")
```

Exercises

(Solutions: https://gitlab.uzh.ch/charlotte.micheloud/replicationstudies)

For all studies from the replication projects investigate

Exercise 1.1

How many study pairs fulfill the **two-trials rule** criterion for replication success? Use a threshold of $\alpha = 0.025$ for the one-sided *p*-values.

Exercise 1.2

For how many study pairs do you find evidence for **incompatible** effect estimates (on Fisher *z*-scale)? Use the function `Qtest()` and a threshold of $\alpha = 0.05$ for the resulting *p*-value.

Exercises

(Solutions: https://gitlab.uzh.ch/charlotte.micheloud/replicationstudies)

For all studies from the replication projects investigate

Exercise 1.3

Compute the one-sided **sceptical *p*-value**. How many replication studies are successful at 0.025? Use the function `pSceptical()`

Exercise 1.4

Look closer at the studies which show **discrepancies** in terms of replication success based on the two-trials rule and the sceptical *p*-value. How do their effect estimates and sample sizes compare?

Exercises

Exercise 1.5 (if time permits)

Calculate the **relative effect size** $d = \hat{\theta}_r / \hat{\theta}_o$ for the discrepant studies, **as well as the minimum relative effect size** d_{min} with the two approaches (two-trials rule and sceptical p-value).

Use the functions `effectSizeSignificance` and `effectSizeReplicationSuccess`.

Design of replication studies

Design based on the two-trials rule

Design of replication studies

Sample size of replication study

- Direct replication $\rightarrow$ procedures of replication study as closely matched as possible to original study
- But same sample size as in original study can lead to a very low power (Goodman, 1992)

  $\rightarrow$ proper sample size calculation is essential

STATISTICS IN MEDICINE, VOL. 11, 875–879 (1992)

A COMMENT ON REPLICATION, *P*-VALUES AND EVIDENCE

STEVEN N. GOODMAN

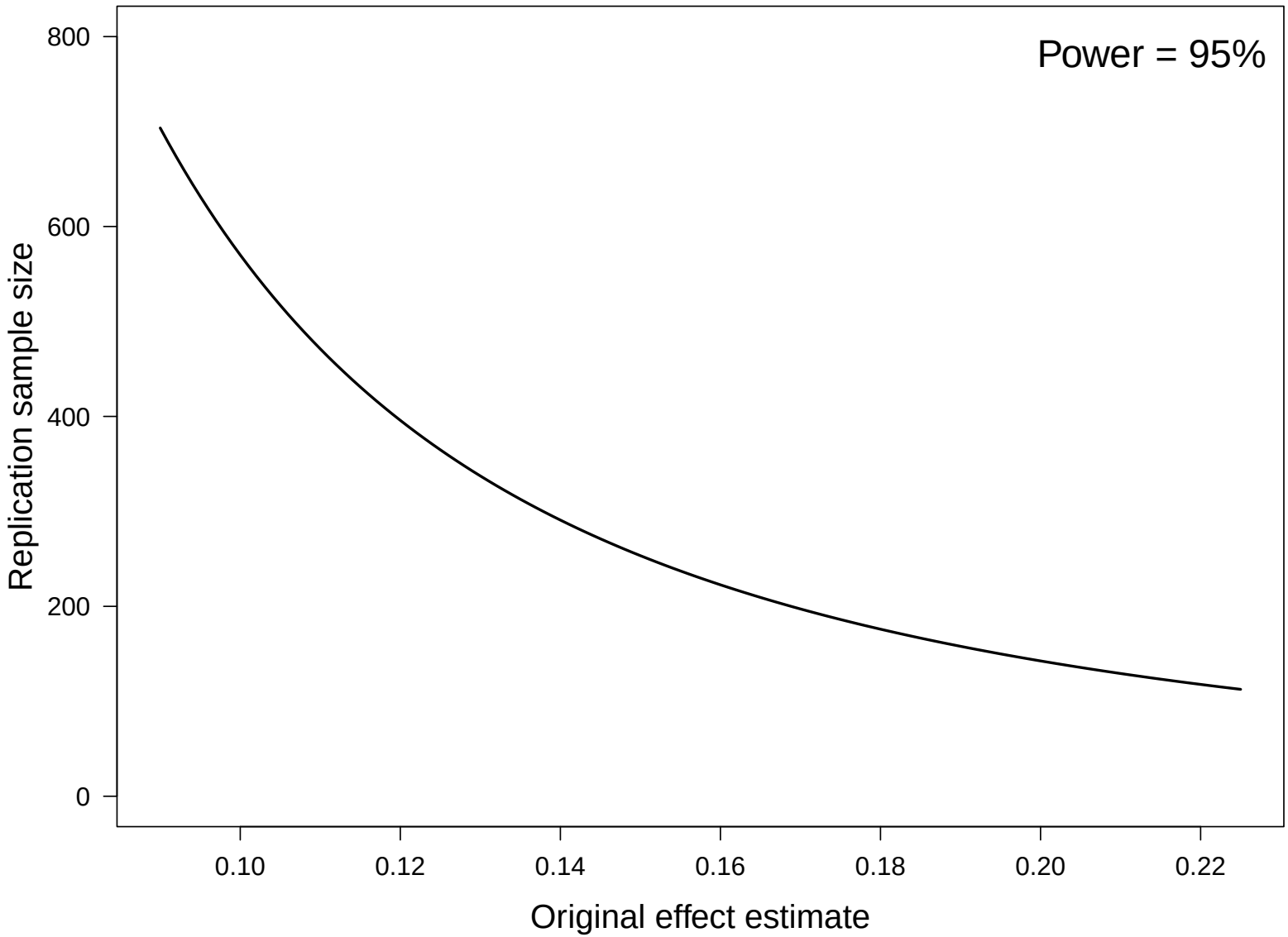
Johns Hopkins University School of Medicine, Department of Oncology, Division of Biostatistics, 550 N. Broadway, Suite 1103, Baltimore MD 21205, U.S.A.

What is used in practice

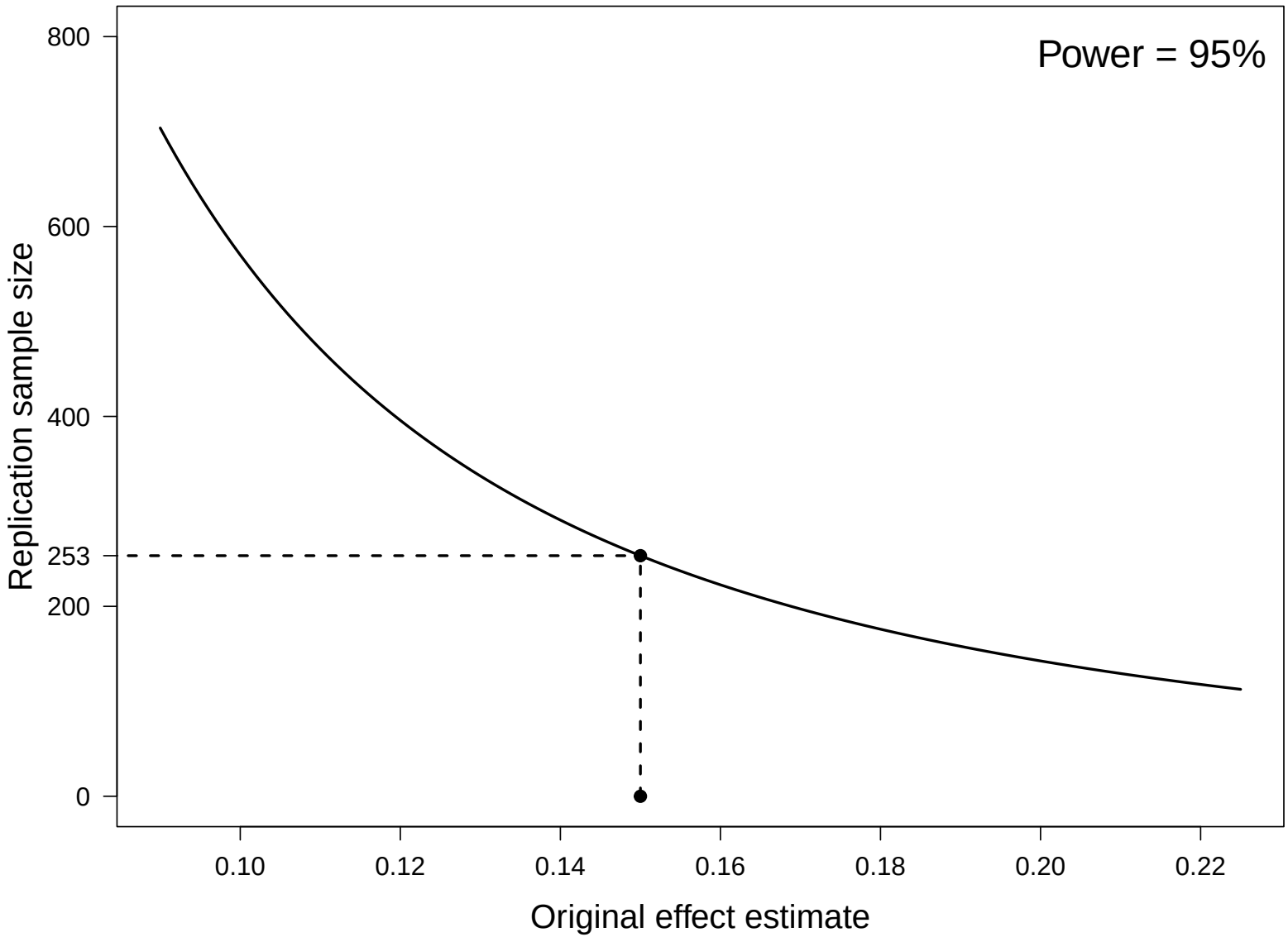
Standard sample size calculation

- Goal is to have between 80% and 95% power in the replication study to detect the effect estimate from the original study.
- Original effect estimate is sometimes shrunken by a factor of 50%.
- Uncertainty of original effect estimate is ignored

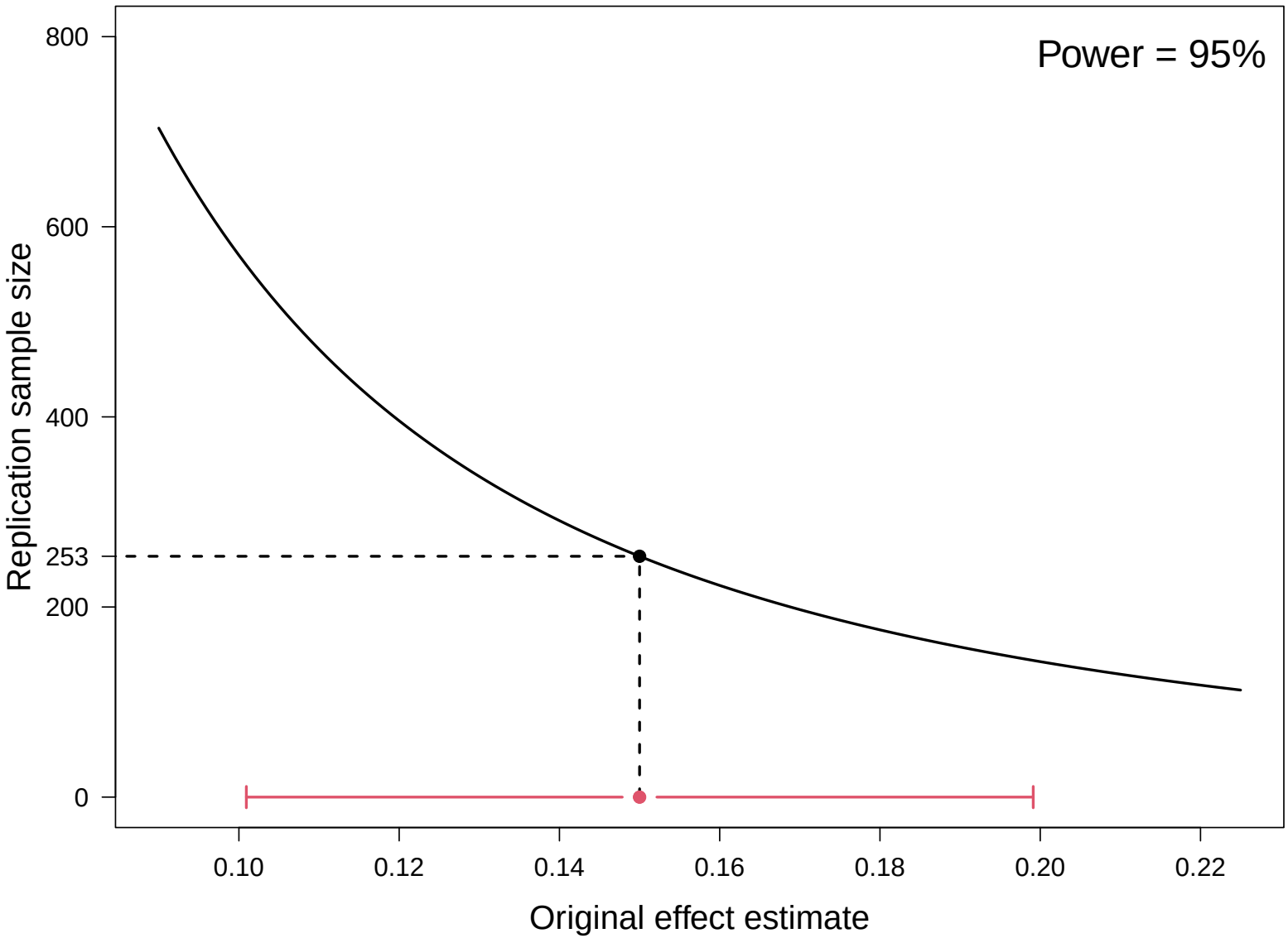
Standard sample size calculation



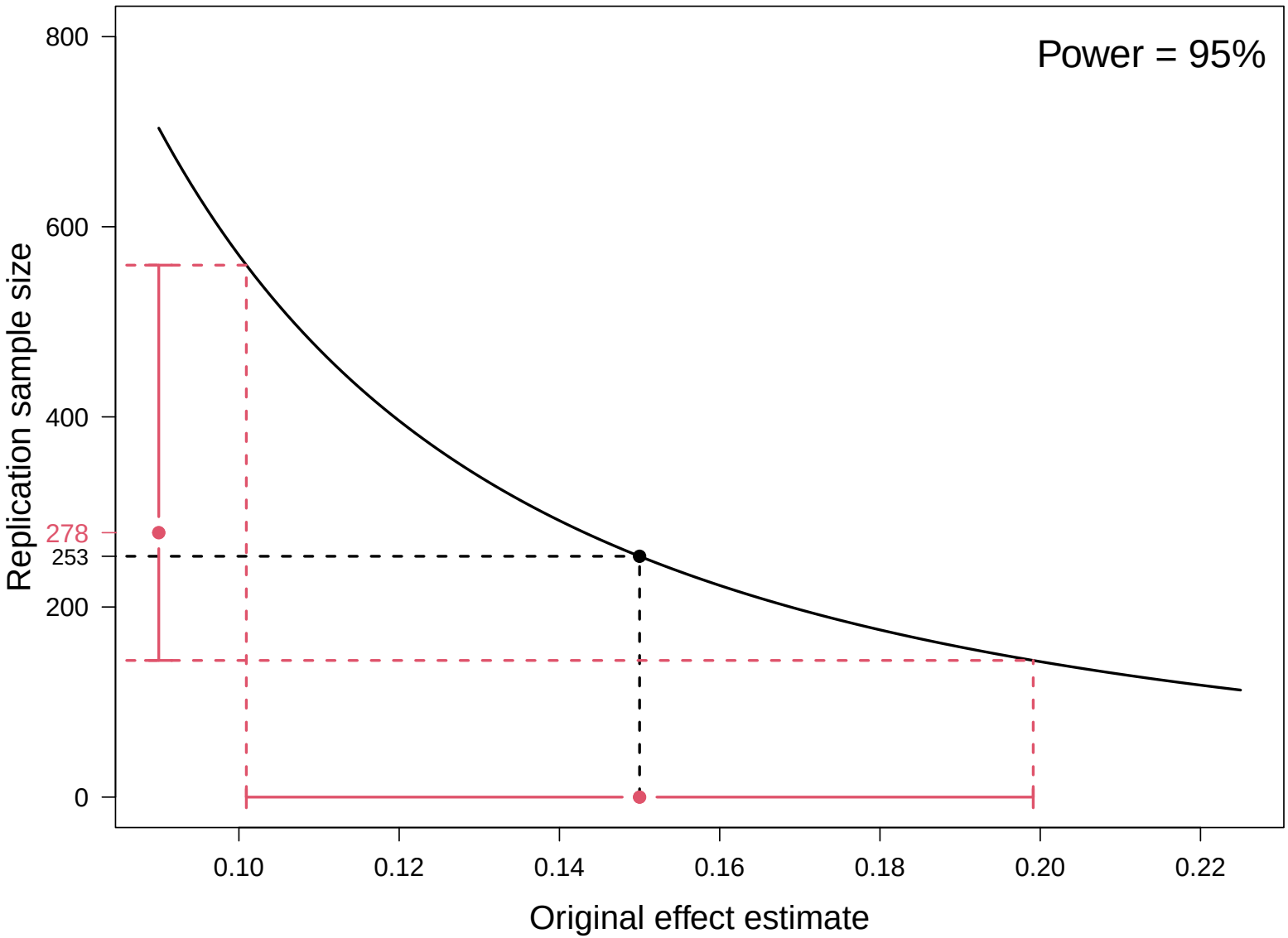
Standard sample size calculation



Incorporation of uncertainty



Incorporation of uncertainty



Incorporation of uncertainty

Design prior

- *Conditional*: ignores uncertainty of original study
- *Predictive*: reflects that there is uncertainty about the true effect after the original experiment

Power of the two-trials rule

in absolute terms

Conditional design prior

$$\text{Power} = \Phi\left(\frac{\hat{\theta}_o\sqrt{n_r}}{\sigma} - z_{1-\alpha}\right)$$

Predictive design prior

$$\text{Power} = \Phi\left(\sqrt{\frac{n_o}{n_o + n_r}}\left(\frac{\hat{\theta}_o\sqrt{n_r}}{\sigma} - z_{1-\alpha}\right)\right)$$

Power of the two-trials rule

in relative terms

Conditional design prior

$$\text{Power} = \Phi\left(z_o\sqrt{c} - z_{1-\alpha}\right)$$

Predictive design prior

$$\text{Power} = \Phi\left(\frac{1}{\sqrt{c+1}}\left(z_o\sqrt{c} - z_{1-\alpha}\right)\right)$$

Power of the two-trials rule

with shrinkage

Conditional design prior

$$\text{Power} = \Phi\left((1 - s)z_o\sqrt{c} - z_{1-\alpha}\right)$$

Predictive design prior

$$\text{Power} = \Phi\left(\frac{1}{\sqrt{c+1}}\left((1 - s)z_o\sqrt{c} - z_{1-\alpha}\right)\right)$$

Exercise session 2

Design based on the two-trials rule

Design based on the two-trials rule

Two functions:

- `powerSignificance()` and `sampleSizeSignificance()`

Main arguments (`default`):

- `zo`
- `c` (1)
- `power`
- `designPrior` (`"conditional"`)
- `shrinkage` (0)
- `level` (0.025)
- `alternative` (`"one.sided"`)

Example from Pyc and Rawson (2010)

No shrinkage

- p-value $p_o = 0.011$
- relative sample size $c = 9.2$

```
# power calculation
powerSignificance(zo = p2z(0.011, alternative = "one.sided"),
                  c = 9.2,
                  designPrior = "conditional")
```

```
## [1] 0.9999997
```

Example from Pyc and Rawson (2010)

With 50% shrinkage

- p-value $p_o = 0.011$
- relative sample size $c = 9.2$

```
# power calculation
powerSignificance(zo = p2z(0.011, alternative = "one.sided"),
                  c = 9.2,
                  shrinkage = 0.5,
                  designPrior = "conditional")
```

```
## [1] 0.9349301
```

Exercises

(Solutions: https://gitlab.uzh.ch/charlotte.micheloud/replicationstudies)

Exercise 2.1

We have five original studies that we want to replicate. The one-sided p-values are 0.0001, 0.001, 0.005, 0.01, and 0.025, respectively. We decide to use the same sample size as in the original study ($c = 1$).

- Compute and plot the conditional and predictive power of the five replication studies. Use the function `powerSignificance()`
- Shrink the original effect estimate by a factor of 25% and use a conditional design prior. How does the power compare to the conditional power without shrinkage?

Exercises

(Solutions: https://gitlab.uzh.ch/charlotte.micheloud/replicationstudies)

Exercise 2.2

- Compute and plot the relative sample sizes of the five studies to achieve a power of 80% with the conditional and the predictive design prior. Use the function `sampleSizeSignificance()`.
- Shrink the original effect estimate by a factor of 25% and use a conditional design prior. How does the required relative sample size change compared to not shrinking the estimate?

Design based on replication success (the sceptical p-value)

Power for replication success

Conditional design prior

$$\text{Power} = \Phi\left(z_o\sqrt{c}(1 - d_{\text{min}})\right)$$

Predictive design prior

$$\text{Power} = \Phi\left(\frac{1}{\sqrt{c+1}}\left(z_o\sqrt{c}(1 - d_{\text{min}})\right)\right)$$

Exercise Session 3

Design based on replication success (sceptical *p*-value)

Design based on replication success

Two functions:

- `powerReplicationSuccess()` and `sampleSizeReplicationSuccess()`

Main arguments (`default`):

- `zo`
- `c` (1)
- `power`
- `designPrior` (`"conditional"`)
- `level` (`0.025`)
- `alternative` (`"one.sided"`)
- `type` (`"golden"`)

Example from Pyc and Rawson (2010)

- p-value $p_o = 0.011$
- relative sample size $c = 9.2$

```
# power calculation
powerReplicationSuccess(zo = p2z(0.011, alternative = "one.sided"),
                        c = 9.2,
                        designPrior = "conditional")
```

```
## [1] 0.9923838
```

Example from Pyc and Rawson (2010)

With 50% shrinkage

- p-value $p_o = 0.011$
- relative sample size $c = 9.2$

```
# power calculation
powerReplicationSuccess(zo = p2z(0.011, alternative = "one.sided"),
                        c = 9.2,
                        shrinkage = 0.5,
                        designPrior = "conditional")
```

```
## [1] 0.1476171
```

Exercises

(Solutions: https://gitlab.uzh.ch/charlotte.micheloud/replicationstudies)

Exercise 3.1

- Compute and plot the conditional and predictive power for replication success. Use the function `powerReplicationSuccess()` with $c = 1$ and $p_o = 0.0001, 0.001, 0.005, 0.01$ and 0.025.
- Compare conditional power for replication success with conditional power for significance (exercise 2.1).

More advanced topics

Meta-analysis and multiple replication studies

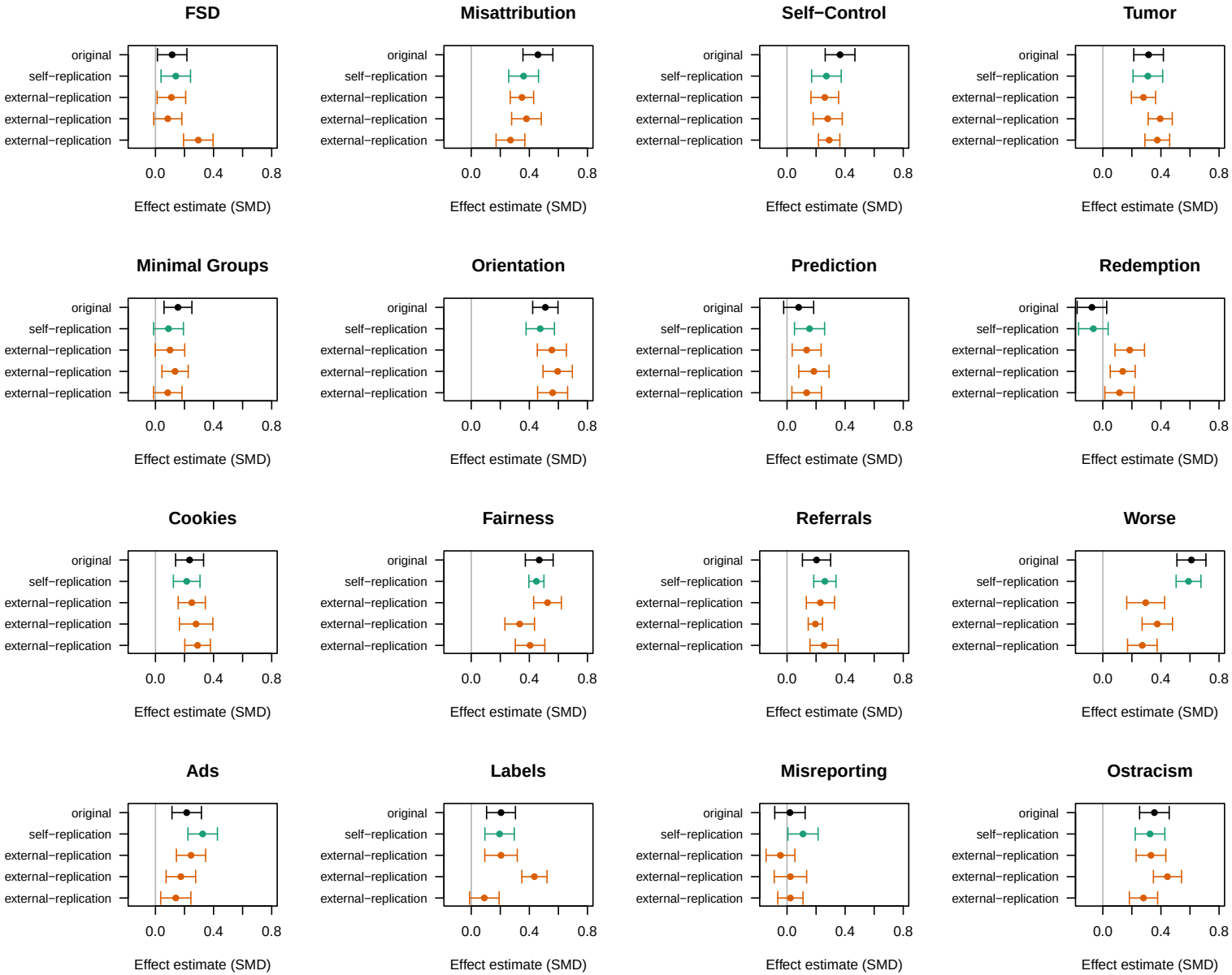
Multiple replications


```
data("protzko2020")
head(protzko2020, 10)
```

```
##         experiment                type lab    smd         se    n
## 1              FSD            original   1 0.1150 0.05176026 1500
## 2              FSD     self-replication   1 0.1400 0.05176026 1501
## 3              FSD external-replication   3 0.2950 0.05181392 1586
## 4              FSD external-replication   2 0.0850 0.04939294 1648
## 5              FSD external-replication   4 0.1100 0.05006561 1606
## 6  Misattribution            original   1 0.4590 0.05253345 1497
## 7  Misattribution external-replication   2 0.2700 0.05051459 1597
## 8  Misattribution     self-replication   1 0.3605 0.05229597 1500
## 9  Misattribution external-replication   4 0.3800 0.05205281 1519
## 10 Misattribution external-replication   3 0.3490 0.04110517 2421
```

Forest plots

16 experiments



Effect measures

- For continuous outcomes
  - Standardized mean difference θ
- For binary/event outcomes relative treatment effects are preferred:
  - Relative risk RR
  - Odds ratio OR
  - Hazard ratio HR
- These are usually considered on a log-scale:
  $\theta = \log(\text{RR})$, $\theta = \log(\text{OR})$, $\theta = \log(\text{HR})$

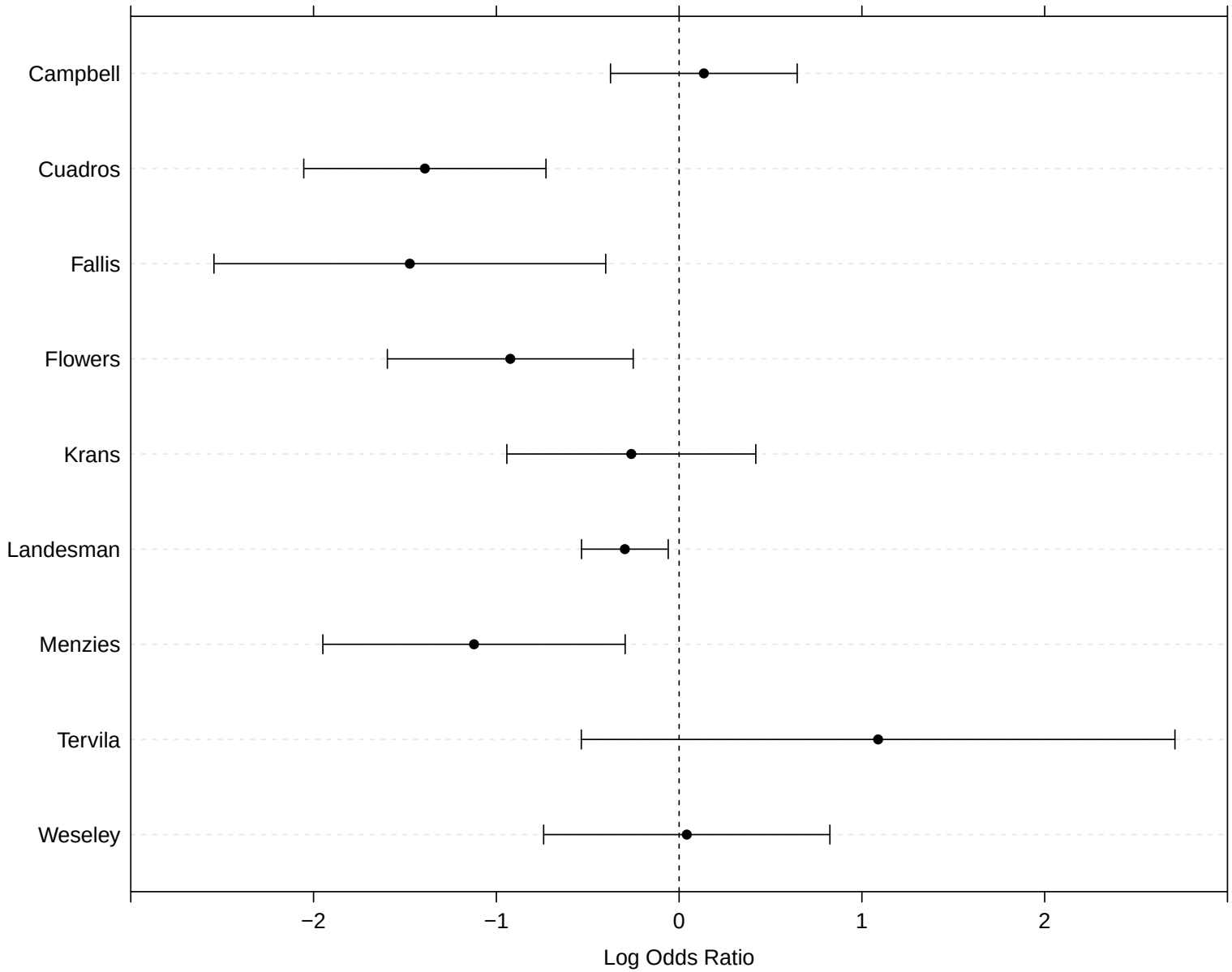
Incidence of preeclampsia

Effect measure: Odds ratio for diuretic vs. control

Study	Diuretic		Control		OR
Weseley	11%	(14/131)	10%	(14/136)	1.04
Flowers	5%	(21/385)	13%	(17/134)	0.40
Menzies	25%	(14/57)	50%	(24/48)	0.33
Fallis	16%	(6/38)	45%	(18/40)	0.23
Cuadros	1%	(12/1011)	5%	(35/760)	0.25
Landesman	10%	(138/1370)	13%	(175/1336)	0.74
Krans	3%	(15/506)	4%	(20/524)	0.77
Tervila	6%	(6/108)	2%	(2/103)	2.97
Campbell	42%	(65/153)	39%	(40/102)	1.14

Forest plot

Effect estimates with 95% CIs



Notation and overall effect

Fixed effect model

- Notation:
  - $i = 1, \ldots, n$ trials
  - θ_i, $\hat{\theta}_i$: true and estimated study-specific treatment effect
  - standard error $\sigma_i = \text{se}(\hat{\theta}_i)$ of $\hat{\theta}_i$
  - variance $v_i = \sigma_i^2$
- Homogeneity assumption: $\theta_i = \theta$ for all i
- The estimate $\hat{\theta}$ of the overall treatment effect θ is then a weighted average of study-specific estimates $\hat{\theta}_i$ with inverse variance weights $w_i = 1/v_i$:

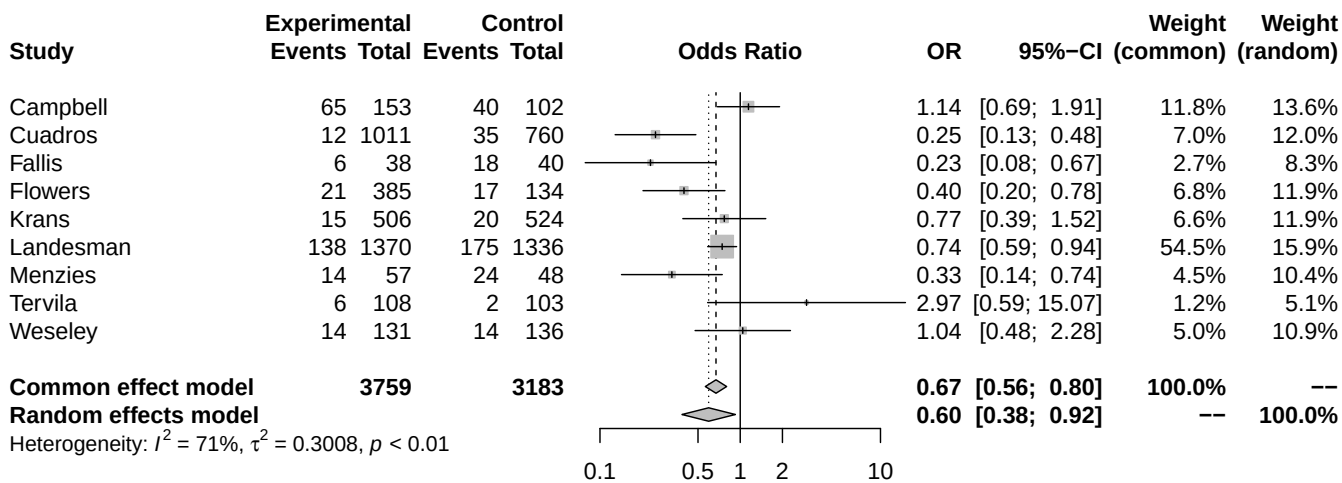
$$\hat{\theta} = \frac{\sum w_i \hat{\theta}_i}{\sum w_i} \text{ with } \text{se}(\hat{\theta}) = 1/\sqrt{\sum w_i}.$$

Presentation in forest plot

Fixed effect model

```
library(meta)
meta1 <- metabin(event.e = treatedPre,
                 n.e = treatedTot,
                 event.c = controlPre,
                 n.c = controlTot,
                 sm = "OR", method = "Inverse", studlab = study)

forest(meta1, comb.fixed = TRUE, comb.random = FALSE)
```



Test for heterogeneity

- Under the homogeneity assumption, we have

$$Q = \sum w_i(\hat{\theta}_i - \hat{\theta})^2 \overset{a}{\sim} \chi^2_{n-1}$$

- This can be used to calculate the p-value of Cochran's test for heterogeneity.
- For the preeclampsia data, this test yields $Q = 27.3$ at $n - 1 = 8$ degrees of freedom ($p = 0.0006$).

$\rightarrow$ Strong evidence for heterogeneity between studies.

$\rightarrow$ Fixed effect model questionable.

Random effects model

- Assumes that the θ_i's come from a normal distribution with mean θ^* and heterogeneity variance τ^2:

$$\hat{\theta}_i \,|\, \theta_i \sim \mathrm{N}(\theta_i, v_i) \quad \text{and} \quad \theta_i \sim \mathrm{N}(\theta^*, \tau^2),$$

  so marginally

$$\hat{\theta}_i \sim \mathrm{N}(\theta^*, v_i + \tau^2).$$

- The overall effect estimate $\hat{\theta}^*$ is now based on the weights

$$w_i^* = 1/(v_i + \tau^2).$$

- Compared to the fixed effect model,
  - the confidence intervals for the overall treatment effect will become wider,
  - large studies will obtain less weight.

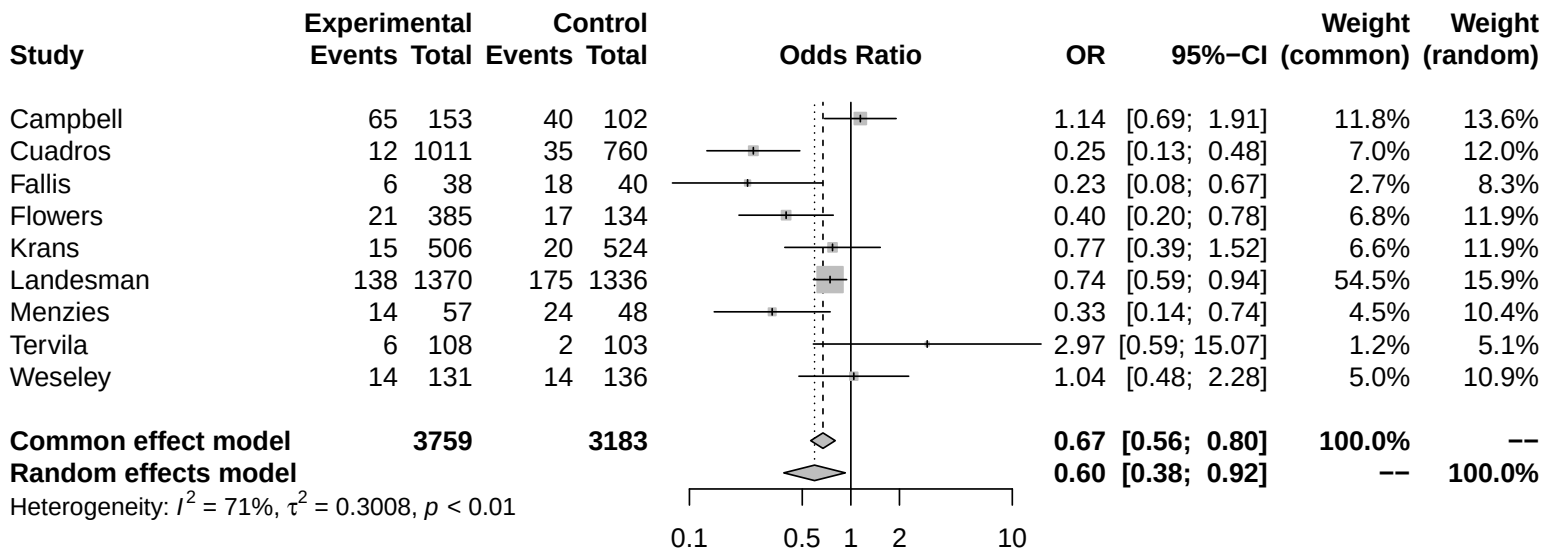
Estimate of heterogeneity variance and Higgins' I^2

- Moment estimator of heterogeneity variance τ^2 compares Q-statistic to its expectation $n-1$ under homogeneity.
- Truncation at zero and appropriate scaling gives estimate $\hat{\tau}^2$.
- Standard DerSimonian-Laird approach plugs $\hat{\tau}^2$ in $w_i^* = 1/(v_i + \tau^2)$.
- Easier to interpret is Higgins' I^2: Percentage of variance that is attributable to study heterogeneity.

Presentation in forest plot

Random effects model

```
library(meta)
forest(meta1, comb.fixed=TRUE, comb.random=TRUE)
```



Assessing replication success

How to assess replication success with multiple replication studies?

One option:

1. Perform a meta-analysis with the replication studies
2. Retrieve/compute:
   - the meta-analytic combined effect $\hat{\theta}_r$
   - the standard error of the combined effect $\sigma_r = \text{se}(\hat{\theta}_r)$
   - the meta-analytic z-value $z_r = \hat{\theta}_r / \sigma_r$
   - the variance ratio (relative sample size) $c = \sigma_o^2 / \sigma_r^2$
3. Apply the same methods as in a 1-1 scenario

Example for experiment FSD

Meta-analysis of the external replications

```
FSD <-  subset(protzko2020, experiment == "FSD")
(meta_FSD <- metagen(TE = smd, seTE = se, studlab = lab,
                     exclude = (type == "original" | type == "self-replication"),
                     data = FSD, sm = "SMD"))
```

```
## Number of studies combined: k = 3
##
##                          SMD           95%-CI    z  p-value
## Common effect model  0.1597 [0.1026; 0.2167] 5.49 < 0.0001
## Random effects model 0.1626 [0.0334; 0.2918] 2.47   0.0136
##
## Quantifying heterogeneity:
##  tau^2 = 0.0105 [0.0009; 0.5171]; tau = 0.1024 [0.0308; 0.7191]
##  I^2 = 80.2% [37.5%; 93.7%]; H = 2.25 [1.26; 3.99]
##
## Test of heterogeneity:
##      Q d.f. p-value
##  10.09    2  0.0064
##
## Details on meta-analytical method:
## - Inverse variance method
## - Restricted maximum-likelihood estimator for tau^2
## - Q-profile method for confidence interval of tau^2 and tau
```

Retrieving results

```r
(thetar_FSD <- meta_FSD$TE.fixed) # meta-analytic estimate
```

```
## [1] 0.1596578
```

```r
(se_thetar_FSD <- meta_FSD$seTE.fixed) # standard error of estimate
```

```
## [1] 0.02909475
```

```r
(zr_FSD <- meta_FSD$zval.fixed) # meta-analytic z-value
```

```
## [1] 5.487511
```

```r
FSD_ori <- subset(FSD, type == "original")
(c_FSD <- FSD_ori$se^2/se_thetar_FSD^2) # variance ratio (relative sample size)
```

```
## [1] 3.164927
```

Assessing replication success

Two-trials rule

```
zo_FSD <- FSD_ori$smd/FSD_ori$se
(po_FSD <- z2p(zo_FSD, alternative = "one.sided"))

## [1] 0.01314904

(pr_FSD <- z2p(zr_FSD, alternative = "one.sided"))

## [1] 2.038184e-08
```

Replication success with the two-trials rule

Sceptical *p*-value

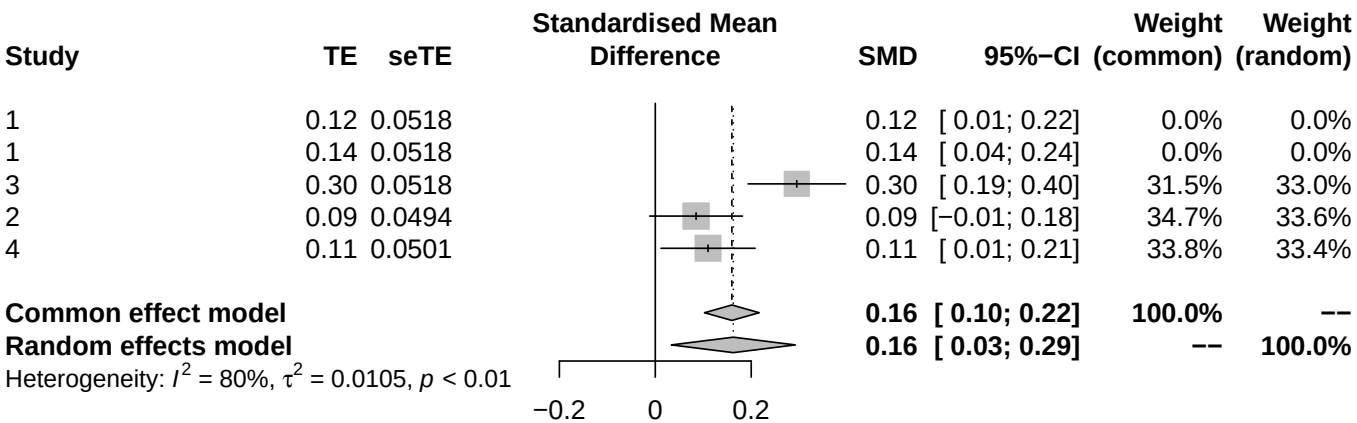
```
(pS_FSD <- pSceptical(zo = zo_FSD, zr = zr_FSD, c = c_FSD,
                      alternative = "one.sided", type = "golden"))

## [1] 0.008750058
```

Replication success with the sceptical *p*-value

Forest plot of experiment FSD

```
forest(meta_FSD)
```



Exercise Session 4

Multiple replication studies

Exercises

(Solutions: https://gitlab.uzh.ch/charlotte.micheloud/replicationstudies)

For each experiment in the Protzko dataset

Exercise 4.1

Compute a fixed and random-effects meta-analysis for the external replications.
Extract the meta-analytic combined estimate ($\hat{\theta}_r$) and compute the relative effect size ($d = \hat{\theta}_r / \hat{\theta}_o$).
Hint: Use the function `metagen` *from the R package* `meta`.

Exercise 4.2

Compute Higgins' I^2.

Exercises

(Solutions: https://gitlab.uzh.ch/charlotte.micheloud/replicationstudies)

Exercise 4.3

Using the results from Exercise 4.1, assess replication success in the 16 experiments with the two methods (two-trials rule and replication success)

Hint:

1. *Compute the original z-value*
2. *Extract the meta-analytic z-value*
3. *Compute the variance ratio (using the standard error of the meta-analytic estimate)*
4. *Using these quantities, calculate the original, replication and sceptical p-values*

Replication studies with a sequential design

Social Sciences Replication Project

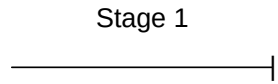
nature human behaviour

Letter | Published: 27 August 2018

Evaluating the replicability of social science experiments in *Nature* and *Science* between 2010 and 2015

Colin F. Camerer, Anna Dreber, Felix Holzmeister, Teck-Hua Ho, Jürgen Huber, Magnus Johannesson, Michael Kirchler, Gideon Nave, Brian A. Nosek, Thomas Pfeiffer, Adam Altmejd, Nick Buttrick, Taizan Chan, Yiling Chen, Eskil Forsell, Anup Gampa, Emma Heikensten, Lily Hummer, Taisuke Imai, Siri Isaksson, Dylan Manfredi, Julia Rose, Eric-Jan Wagenmakers & Hang Wu

- 21 studies replicated
- 11 stopped at stage 1



Significant *p*-value and effect in the same direction: stop

Otherwise: stage 2

Social Sciences Replication Project

nature human behaviour

Letter | Published: 27 August 2018

Evaluating the replicability of social science experiments in *Nature* and *Science* between 2010 and 2015

Colin F. Camerer, Anna Dreber, Felix Holzmeister, Teck-Hua Ho, Jürgen Huber, Magnus Johannesson, Michael Kirchler, Gideon Nave, Brian A. Nosek, Thomas Pfeiffer, Adam Altmejd, Nick Buttrick, Taizan Chan, Yiling Chen, Eskil Forsell, Anup Gampa, Emma Heikensten, Lily Hummer, Taisuke Imai, Siri Isaksson, Dylan Manfredi, Julia Rose, Eric-Jan Wagenmakers & Hang Wu

- 21 studies replicated
- 11 stopped at stage 1
- 10 continued in stage 2

Significant *p*-value and effect in the same direction: stop

Otherwise: stage 2

Interim analyses

Stopping a study at an interim analysis

- popular in clinical trials
- saves time and money
- Two situations

Efficacy stopping

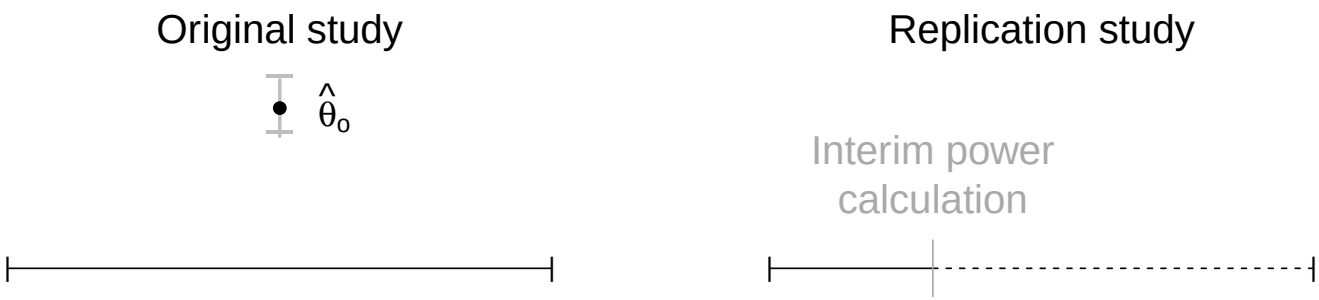
Futility stopping

Interim power

- Power of a replication study taking into account **data from an interim analysis**
- *Conditional* vs *predictive*
- Depends on p_o, c, but also
  - p_i: the p-value at interim
  - f: the fraction of the replication study already completed

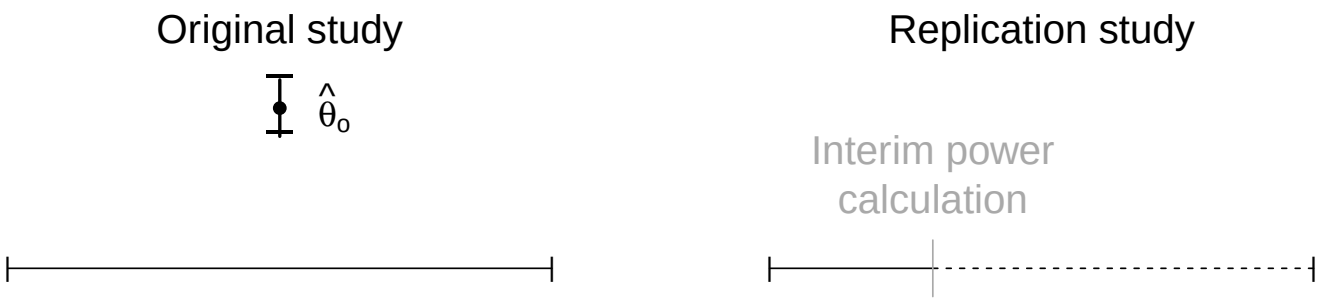
Conditional power at interim

Uncertainty of original result is ignored



Predictive power at interim

Uncertainty of original result is not ignored



Results

	Original		Interim		Interim power (in %)		Replication	
Study	p_o	$\hat{\theta}_o$	p_i	$\hat{\theta}_i$	Cond.	Pred.	p_r	$\hat{\theta}_r$
Duncan	0.002	0.82	0.14	0.18	100	74.6	< 0.0001	0.47
Pyc and Rawson	0.011	0.40	0.045	0.15	100	85.3	0.004	0.15
Sparrow	0.0009	0.39	0.13	0.11	99.7	74.1	0.23	0.05
Rand	0.004	0.14	0.18	0.03	99.8	51.9	0.12	0.03
Gervais and Norenzayan	0.014	0.30	0.79	-0.05	97.5	1.9	0.79	-0.04
Kidd and Castano	0.006	0.28	0.87	-0.07	98.9	1.6	0.77	-0.03
Lee and Schwarz	0.006	0.41	0.77	-0.07	97.7	3.1	0.78	-0.05
Ramirez and Beilock	< 0.0001	1.08	0.64	-0.08	100	61.4	0.80	-0.10
Shah	0.023	0.27	0.93	-0.09	87	0.1	0.65	-0.02

Results

	Original		Interim		Interim power (in %)		Replication	
Study	p_o	$\hat{\theta}_o$	p_i	$\hat{\theta}_i$	Cond.	Pred.	p_r	$\hat{\theta}_r$
Duncan	0.002	0.82	0.14	0.18	100	74.6	< 0.0001	0.47
Pyc and Rawson	0.011	0.40	0.045	0.15	100	85.3	0.004	0.15
Sparrow	0.0009	0.39	0.13	0.11	99.7	74.1	0.23	0.05
Rand	0.004	0.14	0.18	0.03	99.8	51.9	0.12	0.03
Gervais and Norenzayan	0.014	0.30	0.79	-0.05	97.5	1.9	0.79	-0.04
Kidd and Castano	0.006	0.28	0.87	-0.07	98.9	1.6	0.77	-0.03
Lee and Schwarz	0.006	0.41	0.77	-0.07	97.7	3.1	0.78	-0.05
Ramirez and Beilock	< 0.0001	1.08	0.64	-0.08	100	61.4	0.80	-0.10
Shah	0.023	0.27	0.93	-0.09	87	0.1	0.65	-0.02

Results

futility boundary: 10%

$\rightarrow$ no study stopped for futility with conditional power at interim

	Original		Interim		Interim power (in %)		Replication	
Study	p_o	$\hat{\theta}_o$	p_i	$\hat{\theta}_i$	Cond.	Pred.	p_r	$\hat{\theta}_r$
Duncan	0.002	0.82	0.14	0.18	100	74.6	< 0.0001	0.47
Pyc and Rawson	0.011	0.40	0.045	0.15	100	85.3	0.004	0.15
Sparrow	0.0009	0.39	0.13	0.11	99.7	74.1	0.23	0.05
Rand	0.004	0.14	0.18	0.03	99.8	51.9	0.12	0.03
Gervais and Norenzayan	0.014	0.30	0.79	-0.05	97.5	1.9	0.79	-0.04
Kidd and Castano	0.006	0.28	0.87	-0.07	98.9	1.6	0.77	-0.03
Lee and Schwarz	0.006	0.41	0.77	-0.07	97.7	3.1	0.78	-0.05
Ramirez and Beilock	< 0.0001	1.08	0.64	-0.08	100	61.4	0.80	-0.10
Shah	0.023	0.27	0.93	-0.09	87	0.1	0.65	-0.02

Results

futility boundary: 10%

$\rightarrow$ four studies stopped for futility with predictive power at interim

	Original		Interim		Interim power (in %)		Replication	
Study	p_o	$\hat{\theta}_o$	p_i	$\hat{\theta}_i$	Cond.	Pred.	p_r	$\hat{\theta}_r$
Duncan	0.002	0.82	0.14	0.18	100	74.6	< 0.0001	0.47
Pyc and Rawson	0.011	0.40	0.045	0.15	100	85.3	0.004	0.15
Sparrow	0.0009	0.39	0.13	0.11	99.7	74.1	0.23	0.05
Rand	0.004	0.14	0.18	0.03	99.8	51.9	0.12	0.03
Gervais and Norenzayan	0.014	0.30	0.79	-0.05	97.5	1.9	0.79	-0.04
Kidd and Castano	0.006	0.28	0.87	-0.07	98.9	1.6	0.77	-0.03
Lee and Schwarz	0.006	0.41	0.77	-0.07	97.7	3.1	0.78	-0.05
Ramirez and Beilock	< 0.0001	1.08	0.64	-0.08	100	61.4	0.80	-0.10
Shah	0.023	0.27	0.93	-0.09	87	0.1	0.65	-0.02

Results

futility boundary: 10%

$\rightarrow$ four studies stopped for futility with predictive power at interim

	Original		Interim		Interim power (in %)		Replication	
Study	p_o	$\hat{\theta}_o$	p_i	$\hat{\theta}_i$	Cond.	Pred.	p_r	$\hat{\theta}_r$
Duncan	0.002	0.82	0.14	0.18	100	74.6	< 0.0001	0.47
Pyc and Rawson	0.011	0.40	0.045	0.15	100	85.3	0.004	0.15
Sparrow	0.0009	0.39	0.13	0.11	99.7	74.1	0.23	0.05
Rand	0.004	0.14	0.18	0.03	99.8	51.9	0.12	0.03
Gervais and Norenzayan	0.014	0.30	0.79	-0.05	97.5	1.9	0.79	-0.04
Kidd and Castano	0.006	0.28	0.87	-0.07	98.9	1.6	0.77	-0.03
Lee and Schwarz	0.006	0.41	0.77	-0.07	97.7	3.1	0.78	-0.05
Ramirez and Beilock	< 0.0001	1.08	0.64	-0.08	100	61.4	0.80	-0.10
Shah	0.023	0.27	0.93	-0.09	87	0.1	0.65	-0.02

Conclusion/Outlook

- Incorporation of smallest effect size of interest
- Replication of non-inferiority and equivalence studies
- Bayes factors to assess replication success

Statisticians, roll up your sleeves! There's a crisis to be solved

Statisticians play a key role in almost all scientific research. As such, they may be key to solving the reproducibility crisis. **Heidi Seibold**, **Alethea Charlton, Anne-Laure Boulesteix** and **Sabine Hoffmann** urge statisticians to take an active role in promoting more credible science

https://doi.org/10.1111/1740-9713.01554

Acknowledgments

UZH Center for Reproducible Science

RN CH Swiss Reproducibility Network

Swiss National Science Foundation

http://p3.snf.ch/Project-189295

Samuel Pawel

References I

Camerer, C. F., Dreber, A., Forsell, E., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Almenberg, J., Altmejd, A., Chan, T., Heikensten, E., Holzmeister, F., Imai, T., Isaksson, S., Nave, G., Pfeiffer, T., Razen, M., and Wu, H. (2016). Evaluating replicability of laboratory experiments in economics. *Science*, 351(6280):1433–1436. https://doi.org/10.1126/science.aaf0918.

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Nave, G., Nosek, B. A., Pfeiffer, T., Altmejd, A., Buttrick, N., Chan, T., Chen, Y., Forsell, E., Gampa, A., Heikensten, E., Hummer, L., Imai, T., Isaksson, S., Manfredi, D., Rose, J., Wagenmakers, E.-J., and Wu, H. (2018). Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour*, 2(9):637–644. https://doi.org/10.1038/s41562-018-0399-z.

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. (2018). Estimating the reproducibility of experimental philosophy. *Review of Philosophy and Psychology*.

Errington, T. M., Denis, A., Perfito, N., Iorns, E., and Nosek, B. A. (2021). Challenges for assessing replicability in preclinical cancer biology. *eLife*, 10.

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

Held, L. (2020). A new standard for the analysis and design of replication studies (with discussion). *Journal of the Royal Statistical Society, Series A*, 183:431–469. https://doi.org/10.1111/rssa.12493.

Held, L., Micheloud, C., and Pawel, S. (2021). The assessment of replication success based on relative effect size. *The Annals of Applied Statistics*. To appear, preprint available at https://arxiv.org/abs/2009.07782.

References II

Micheloud, C. and Held, L. (2021). Power calculations for replication studies. Technical report. https://arxiv.org/abs/2004.10814.

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science*, 349(517):aac4716. https://doi.org/10.1126/science.aac4716.

Pawel, S. and Held, L. (2020). Probabilistic forecasting of replication studies. *PLOS ONE*, 15(4):e0231416. https://doi.org/10.1371/journal.pone.0231416.

Protzko, J., Krosnick, J., Nelson, L. D., Nosek, B. A., Axt, J., Berent, M., Buttrick, N., DeBell, M., Ebersole, C. R., Lundmark, S., et al. (2020). High replicability of newly-discovered social-behavioral findings is achievable.

Pyc, M. A. and Rawson, K. A. (2010). Why testing improves memory: Mediator effectiveness hypothesis. *Science*, 330(6002):335–335.