Design and Analysis of Replication Studies

Leonhard Held, Charlotte Micheloud

University of Zurich

UZH Center for Reproducible Science

Swiss National Science Foundation

ReproducibiliTea Journal Club, Geneva

Workshop

Analysis of replication studies

Solutions and slides available at

https://gitlab.uzh.ch/charlotte.micheloud/replicationstudies

Package ReplicationSuccess

– Installation

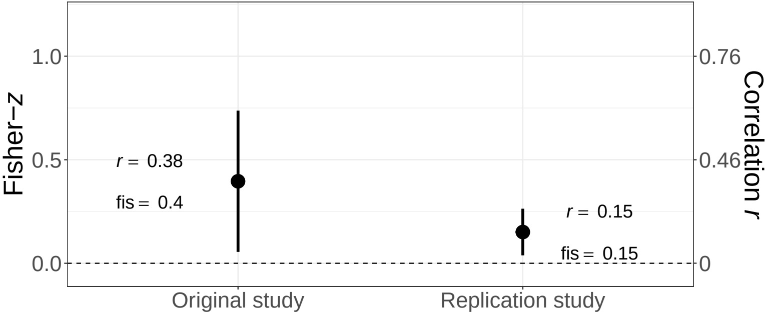
```r
install.packages("ReplicationSuccess")
```

– Usage

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

Statistical framework

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



Data sets

```
data("RProjects")
?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")
```

Statistical framework of package

Key quantities

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

```r
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

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

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

```r
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

```r
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)

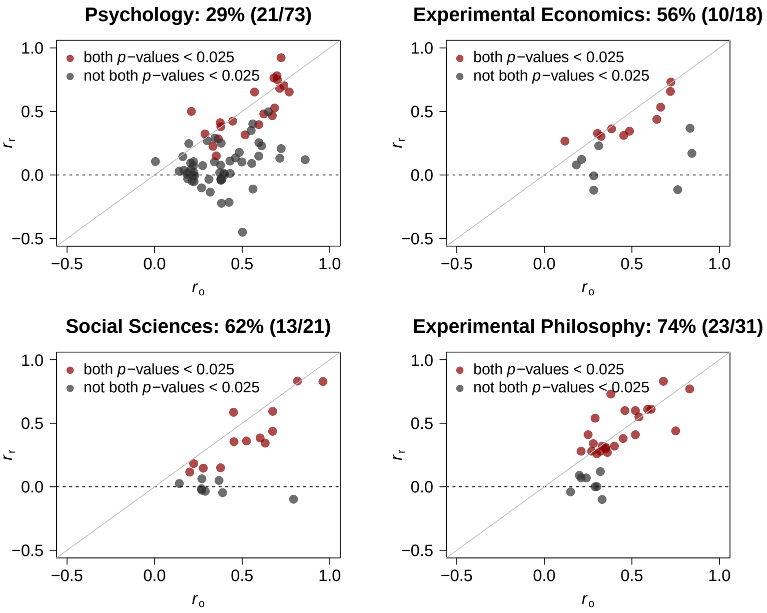
Calculate the **relative effect size** $d = \hat{\theta}_r / \hat{\theta}_o$ for the discrepant studies, **as well as the minimum relative effect size** $d_{\min}$ with the two approaches (two-trials rule and sceptical *p*-value).

Use the functions `effectSizeSignificance` and `effectSizeReplicationSuccess`.

Solution: Exercise 1.1

Project	Both *p*-values < 0.025
Psychology	29% (21/73)
Experimental Economics	56% (10/18)
Social Sciences	62% (13/21)
Experimental Philosophy	74% (23/31)
all	47% (67/143)

Solution: Exercise 1.1



Psychology: 29% (21/73)

Experimental Economics: 56% (10/18)

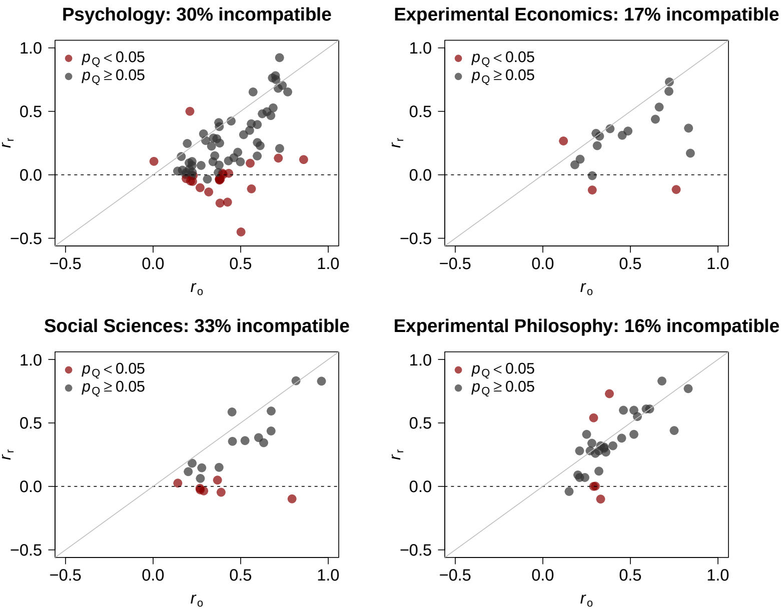
Social Sciences: 62% (13/21)

Experimental Philosophy: 74% (23/31)

Solution: Exercise 1.2

Project	Incompatible estimates ($p_Q < 0.05$)
Psychology	30% (22/73)
Experimental Economics	17% (3/18)
Social Sciences	33% (7/21)
Experimental Philosophy	16% (5/31)
all	26% (37/143)

Solution: Exercise 1.2



Psychology: 30% incompatible

Experimental Economics: 17% incompatible

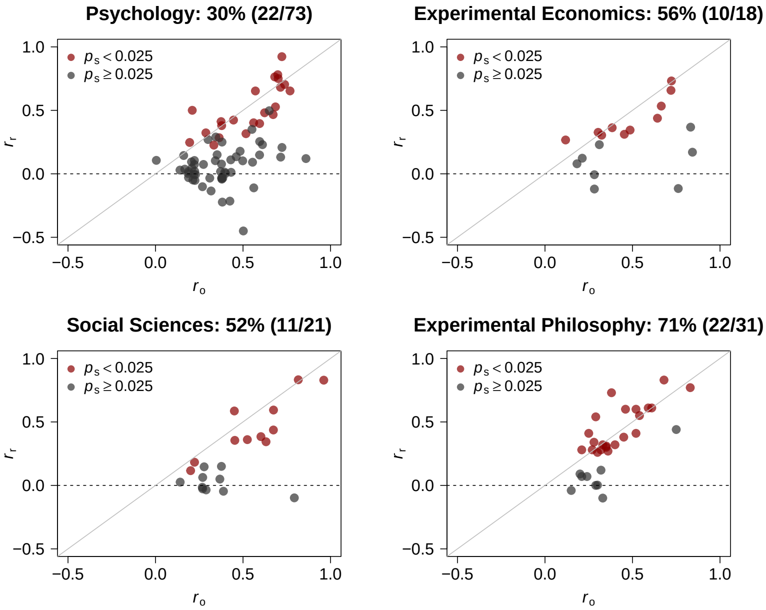
Social Sciences: 33% incompatible

Experimental Philosophy: 16% incompatible

Solution: Exercise 1.3

Project	sceptical p-value < 0.025
Psychology	30% (22/73)
Experimental Economics	56% (10/18)
Social Sciences	52% (11/21)
Experimental Philosophy	71% (22/31)
all	45% (65/143)

Solution: Exercise 1.3



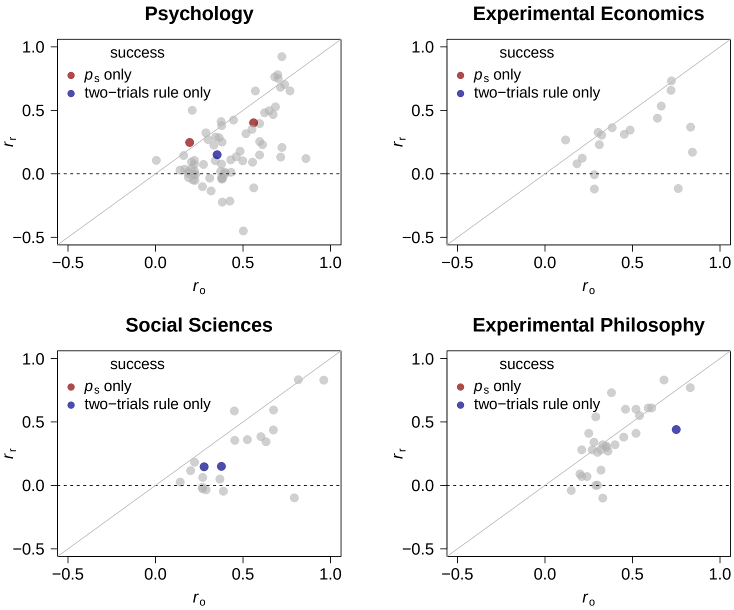
Psychology: 30% (22/73)

Experimental Economics: 56% (10/18)

Social Sciences: 52% (11/21)

Experimental Philosophy: 71% (22/31)

Solution: Exercise 1.4



Solution: Exercise 1.4

Study	n_r/n_o	r_o	r_r	p_o	p_r	p_s
Schmidt and Besner (2008)	2.6	0.2	0.25	0.028	< 0.0001	0.024
Oberauer (2008)	0.6	0.56	0.4	0.0003	0.035	0.017
Payne, Burkley, and Stokes (2008)	2.7	0.35	0.15	0.001	0.023	0.031
Balafoutas and Sutter (2012)	3.5	0.28	0.15	0.009	0.011	0.04
Pyc and Rawson (2010)	9.2	0.38	0.15	0.011	0.004	0.061
Nichols (2006)	9.4	0.75	0.44	0.015	0.0006	0.049

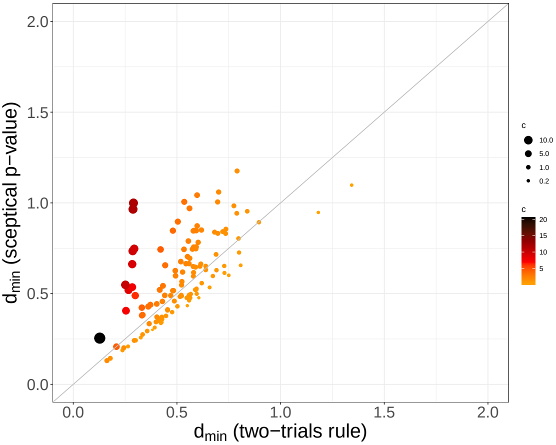
Solution: Exercise 1.5

Study	n_r/n_o	p_o	d	$d_{min}(2TR)$	$d_{min}(p_S)$
Schmidt and Besner (2008)	2.6	0.028	1.28	0.64	1.22
Oberauer (2008)	0.6	0.0003	0.67	0.73	0.61
Payne, Burkley, and Stokes (2008)	2.7	0.001	0.41	0.4	0.44
Balafoutas and Sutter (2012)	3.5	0.009	0.52	0.44	0.66
Pyc and Rawson (2010)	9.2	0.011	0.38	0.28	0.66
Nichols (2006)	9.4	0.015	0.49	0.29	0.75

Solution: Exercise 1.5 (extended)

Significant original studies only

Minimum relative effect size d_{min} with the two-trials rule vs the sceptical *p*-value



Outlook

Design of replication studies

- So far, focus on the analysis of replication studies

→ Design is also of interest

Outlook

Design of replication studies

- So far, focus on the analysis of replication studies

→ Design is also of interest

  - What is the power of the replication study with a certain sample size n_r?

    `powerSignificance(), powerReplicationSuccess()`

Outlook

Design of replication studies

- So far, focus on the analysis of replication studies

→ Design is also of interest

  - What is the power of the replication study with a certain sample size n_r?

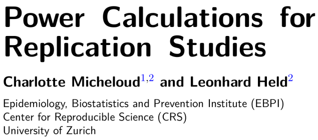
    `powerSignificance(), powerReplicationSuccess()`

  - Which sample size is required to reach a certain level of power?

    `sampleSizeSignificance(), sampleSizeReplicationSuccess()`

Design of replication studies

Literature



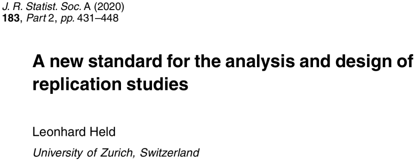
Power Calculations for Replication Studies

Charlotte Micheloud[1,2] and Leonhard Held[2]

Epidemiology, Biostatistics and Prevention Institute (EBPI)
Center for Reproducible Science (CRS)
University of Zurich

to appear in *Statistical Science* (2022)

https://arxiv.org/abs/2004.10814



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

published in *JRSSA* (2020)

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

Interested to participate?

Swiss Reproducibility Network Academy

- Aim: connect early-career researchers interested in reproducibility, open science, good research practices, etc.
- More info: https://www.swissrn.org/academy/
- Contact: swissrnacademy@gmail.com

Next Event

- **What** – Reproducibility Hackathon
- **When** – 20th May 2022, from 10am to 5pm
- **Where** – University of Bern
- **Target group** – young researchers interested in reproducibility
- **More info** – https://www.reprohack.org/event/16/
- **Train tickets** – the SwissRN can reimburse you travel expenses to Bern if needed

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.

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.

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

References II

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.