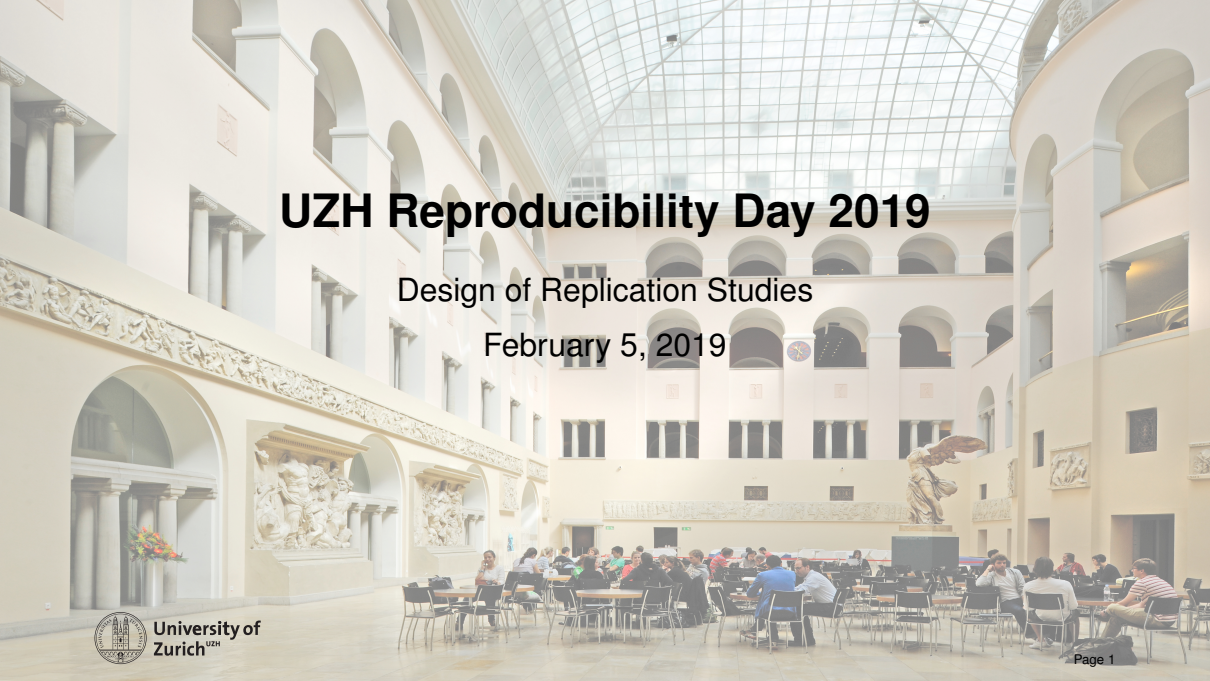


UZH Reproducibility Day 2019

Design of Replication Studies

February 5, 2019

University of Zurich[UZH]

UZH Reproducibility Day 2019

Design of Replication Studies

February 5, 2019

Leonhard Held

The Reproducibility of Psychological Science

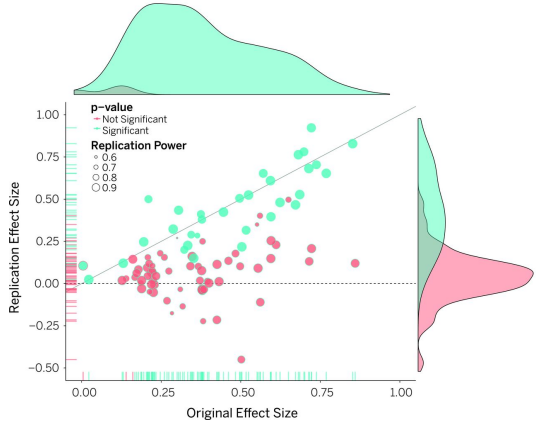
***Science* (2015)**

RESEARCH ARTICLE SUMMARY

PSYCHOLOGY

Estimating the reproducibility of psychological science

Open Science Collaboration*

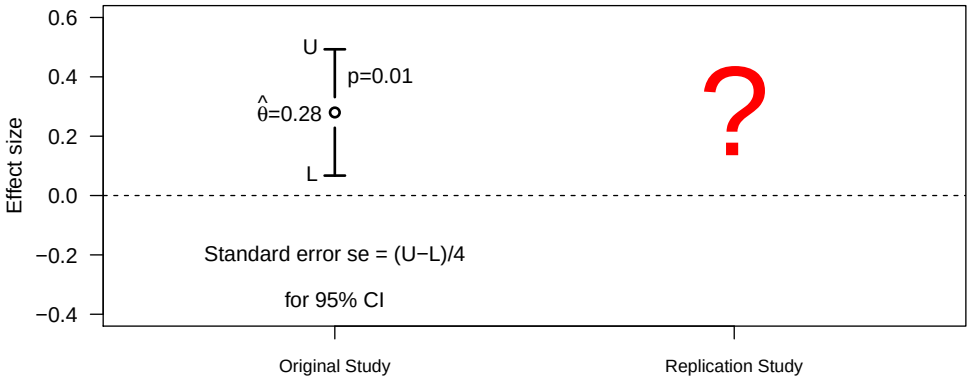


4 Studies from OSC Paper

	Original study			Replication study			Relative sample size
Study	n_o	$\hat{\theta}$	p-value	n_r	$\hat{\theta}$	p-value	n_r/n_o
1	92	0.28	0.01	139	-0.1	0.23	1.5
2	43	0.31	0.05	43	0.27	0.08	1.0
3	28	0.39	0.051	66	0.017	0.89	2.4
4	154	0.38	< 0.0001	50	0.28	0.045	0.3

Effect size θ: correlation coefficient

Original and Replication Study



P-values and Replication

- A significant *p*-value is often interpreted as if the observed effect is “real”.
- But does a replication of the very same experiment produce a significant result with high probability?
- This replication probability turns out to be much lower than expected, even in the absence of publication and other biases.

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.

Replication Probability

What is the probability that a replication study is significant at $\alpha = 5\%$?

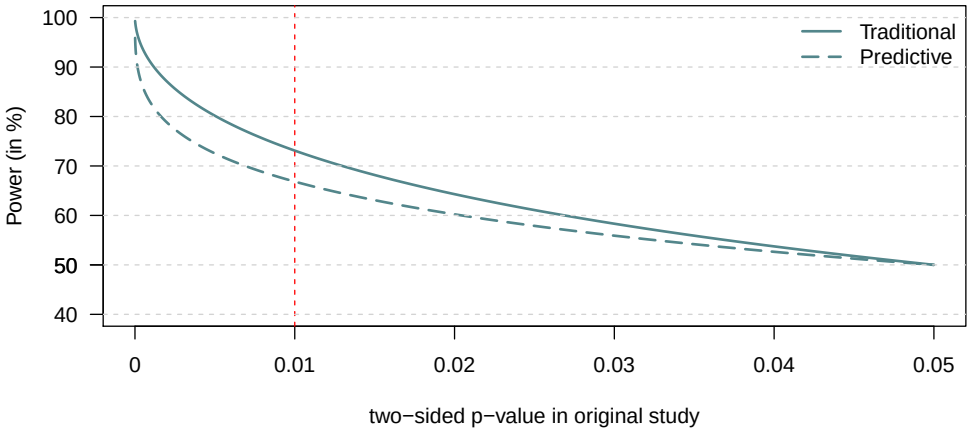
Goodman (1992)

Answer depends on

- p-value of original study
- Relative sample size n_r / n_o
- Assumptions about effect θ:
  - A1 Traditional power to detect the observed effect $\theta = \hat{\theta}$ from the original study
  - A2 Predictive power to detect $\theta \sim \mathrm{N}(\hat{\theta}, \mathrm{se}^2)$ – reflects the fact that there is still uncertainty about the true effect after the original experiment.

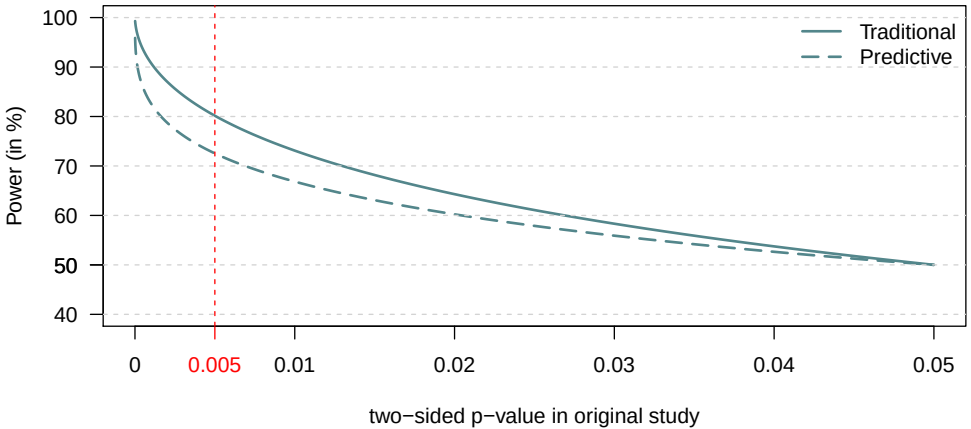
Replication Power

Equal sample sizes $n_r = n_o$



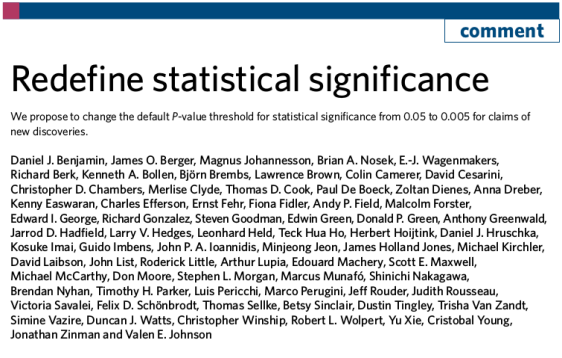
Replication Power

Equal sample sizes $n_r = n_o$



Redefine Statistical Significance for Claims of New Discoveries

***Nature Human Behaviour* (2018)**



comment

Redefine statistical significance

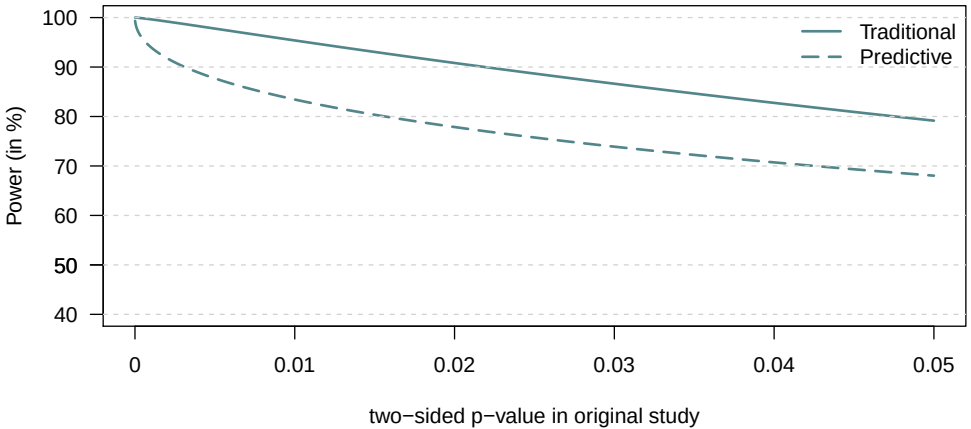
We propose to change the default *P*-value threshold for statistical significance from 0.05 to 0.005 for claims of new discoveries.

Daniel J. Benjamin, James O. Berger, Magnus Johannesson, Brian A. Nosek, E.-J. Wagenmakers, Richard Berk, Kenneth A. Bollen, Björn Brembs, Lawrence Brown, Colin Camerer, David Cesarini, Christopher D. Chambers, Merlise Clyde, Thomas D. Cook, Paul De Boeck, Zoltan Dienes, Anna Dreber, Kenny Easwaran, Charles Efferson, Ernst Fehr, Fiona Fidler, Andy P. Field, Malcolm Forster, Edward I. George, Richard Gonzalez, Steven Goodman, Edwin Green, Donald P. Green, Anthony Greenwald, Jarrod D. Hadfield, Larry V. Hedges, Leonhard Held, Teck Hua Ho, Herbert Hoijtink, Daniel J. Hruschka, Kosuke Imai, Guido Imbens, John P. A. Ioannidis, Minjeong Jeon, James Holland Jones, Michael Kirchler, David Laibson, John List, Roderick Little, Arthur Lupia, Edouard Machery, Scott E. Maxwell, Michael McCarthy, Don Moore, Stephen L. Morgan, Marcus Munafó, Shinichi Nakagawa, Brendan Nyhan, Timothy H. Parker, Luis Pericchi, Marco Perugini, Jeff Rouder, Judith Rousseau, Victoria Savalei, Felix D. Schönbrodt, Thomas Sellke, Betsy Sinclair, Dustin Tingley, Trisha Van Zandt, Simine Vazire, Duncan J. Watts, Christopher Winship, Robert L. Wolpert, Yu Xie, Cristobal Young, Jonathan Zinman and Valen E. Johnson

"This simple step would immediately improve the reproducibility of scientific research in many fields."

Replication Power

Double sample size $n_r = 2 \times n_o$



Replication Sample Size

As before we assume that the result from the original study is known.

What is the required sample size of the replication study?

Answer depends on

- *p*-value of original study
- Power to detect an effect θ, *e. g.* 80% or 90%
- Assumptions about effect θ:
  - A1 Traditional power to detect the observed effect $\theta = \hat{\theta}$ from the original study
  - A2 Predictive power to detect $\theta \sim \mathrm{N}(\hat{\theta}, \mathrm{se}^2)$ – reflects the fact that there is still uncertainty about the true effect after the original experiment.

$\rightarrow$ Shiny App

Exercises

http://tiny.uzh.ch/TD
https://crsuzh.shinyapps.io/replication/

1. Determine the replication power for the studies shown. Compare the results for traditional and predictive power.
2. Determine the relative sample size to achieve a power of 80%. Compare the results for traditional and predictive power with the observed relative sample size.

	Original study			Replication study			Relative sample size
Study	n_o	$\hat{\theta}$	p-value	n_r	$\hat{\theta}$	p-value	n_r/n_o
1	92	0.28	0.01	139	-0.1	0.23	1.5
2	43	0.31	0.05	43	0.27	0.08	1.0
3	28	0.39	0.051	66	0.017	0.89	2.4
4	154	0.38	< 0.0001	50	0.28	0.045	0.3

University of Zurich[UZH]