Predictive Evaluation of Replication Studies

Samuel Pawel, Leonhard Held
Department of Biostatistics
Center for Reproducible Science

University of Zurich[UZH]

23.05.2019

Replication Studies

Direct Replication

- Tool to assess credibility of scientific discoveries
- Regulatory requirement

Replication Studies

Direct Replication

- Tool to assess credibility of scientific discoveries
- Regulatory requirement

Replication Crisis

- Replicability of science is low

$\rightarrow$ Increased interest in metascience

$\rightarrow$ Large-scale replication projects

Reproducibility Project Psychology

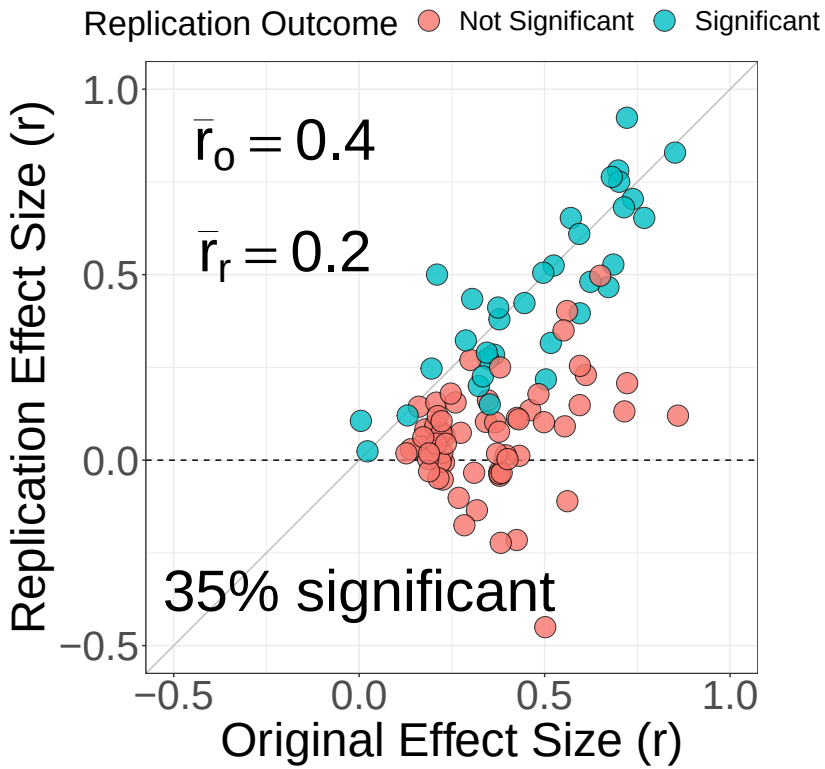
(N = 100)

Science

Estimating the reproducibility of psychological science

Open Science Collaboration

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



Experimental Economics Replication Project

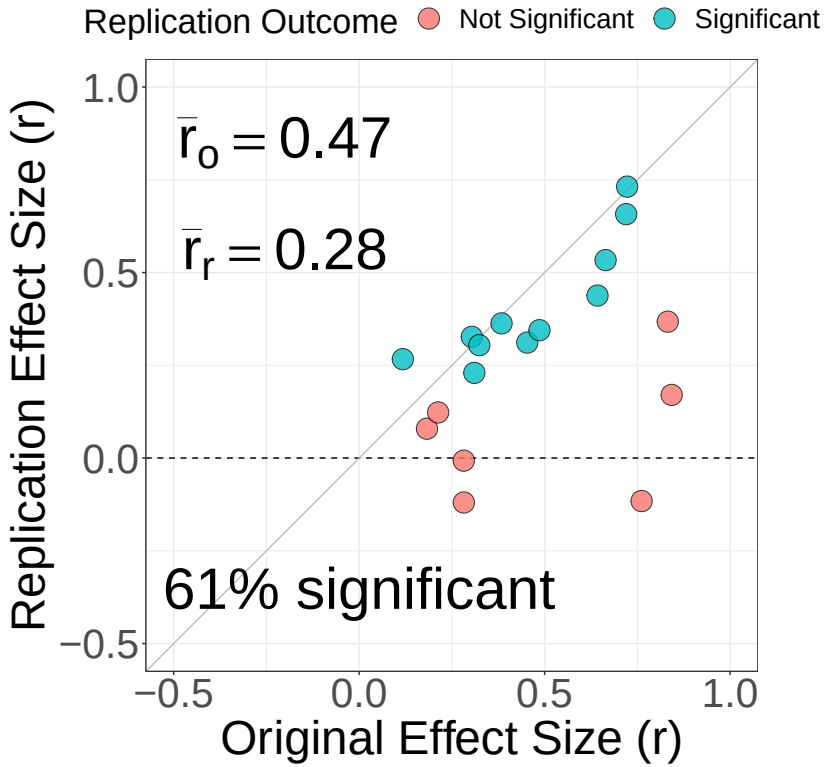
(N = 18)

Science REPORTS

Cite as: Camerer *et al.*, *Science* 10.1126/science.aaf0918 (2016).

Evaluating replicability of laboratory experiments in economics

Colin F. Camerer,[1*†] Anna Dreber,[2†] Eskil Forsell,[2†] Teck-Hua Ho,[3,4†] Jürgen Huber,[5†] Magnus Johannesson,[2†] Michael Kirchler,[5,6†] Johan Almenberg,[7] Adam Altmejd,[2] Taizan Chan,[8] Emma Heikensten,[2] Felix Holzmeister,[5] Taisuke Imai,[1] Siri Isaksson,[2] Gideon Nave,[1] Thomas Pfeiffer,[9,10] Michael Razen,[5] Hang Wu[4]



Social Sciences Replication Project

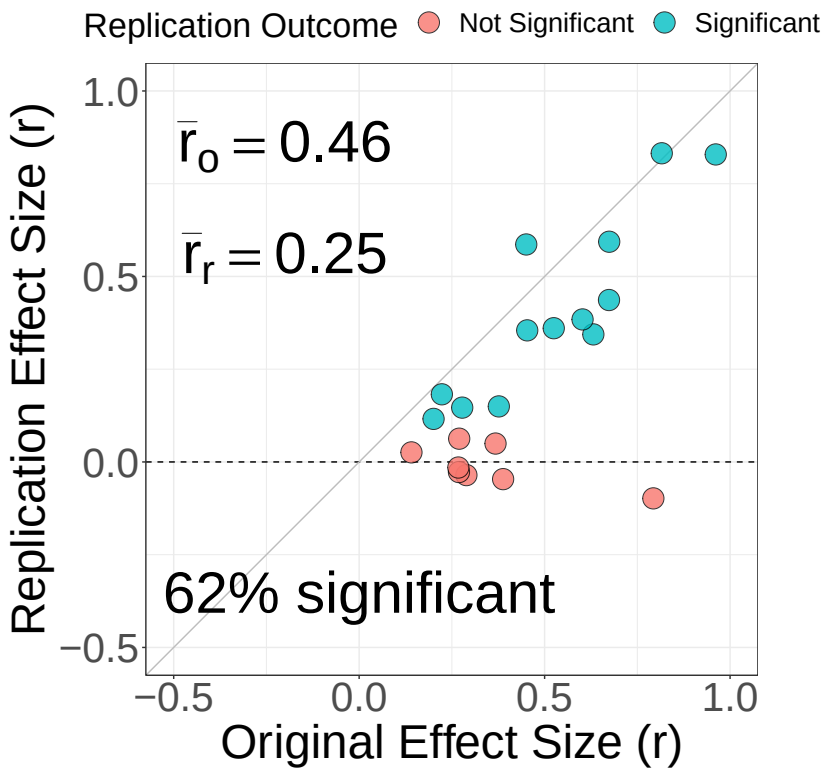
(N = 21)

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



Experimental Philosophy Replicability Project

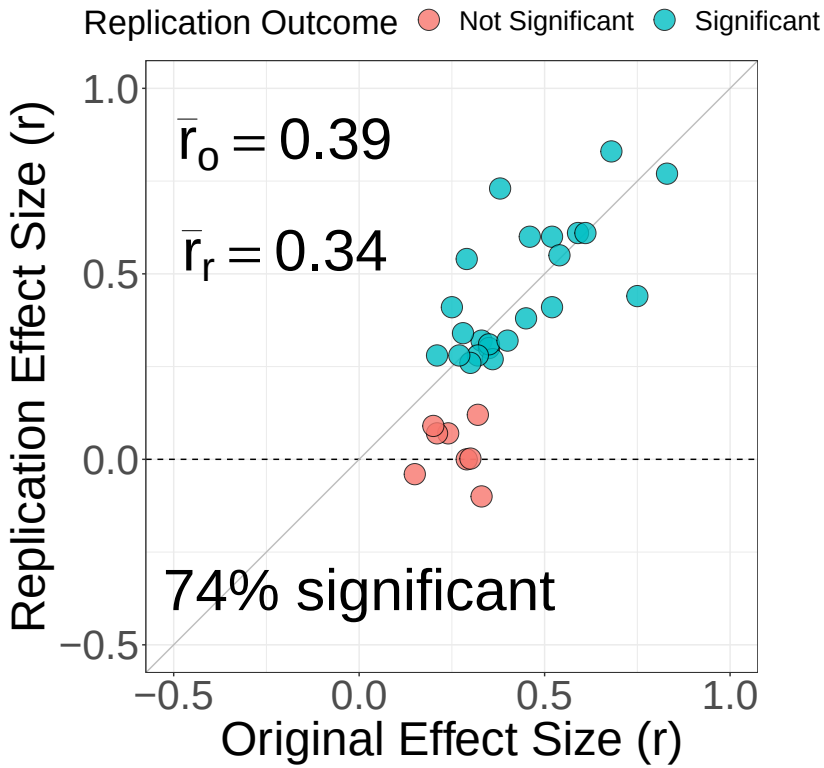
(N = 40)

Rev.Phil.Psych.
https://doi.org/10.1007/s13164-018-0400-9

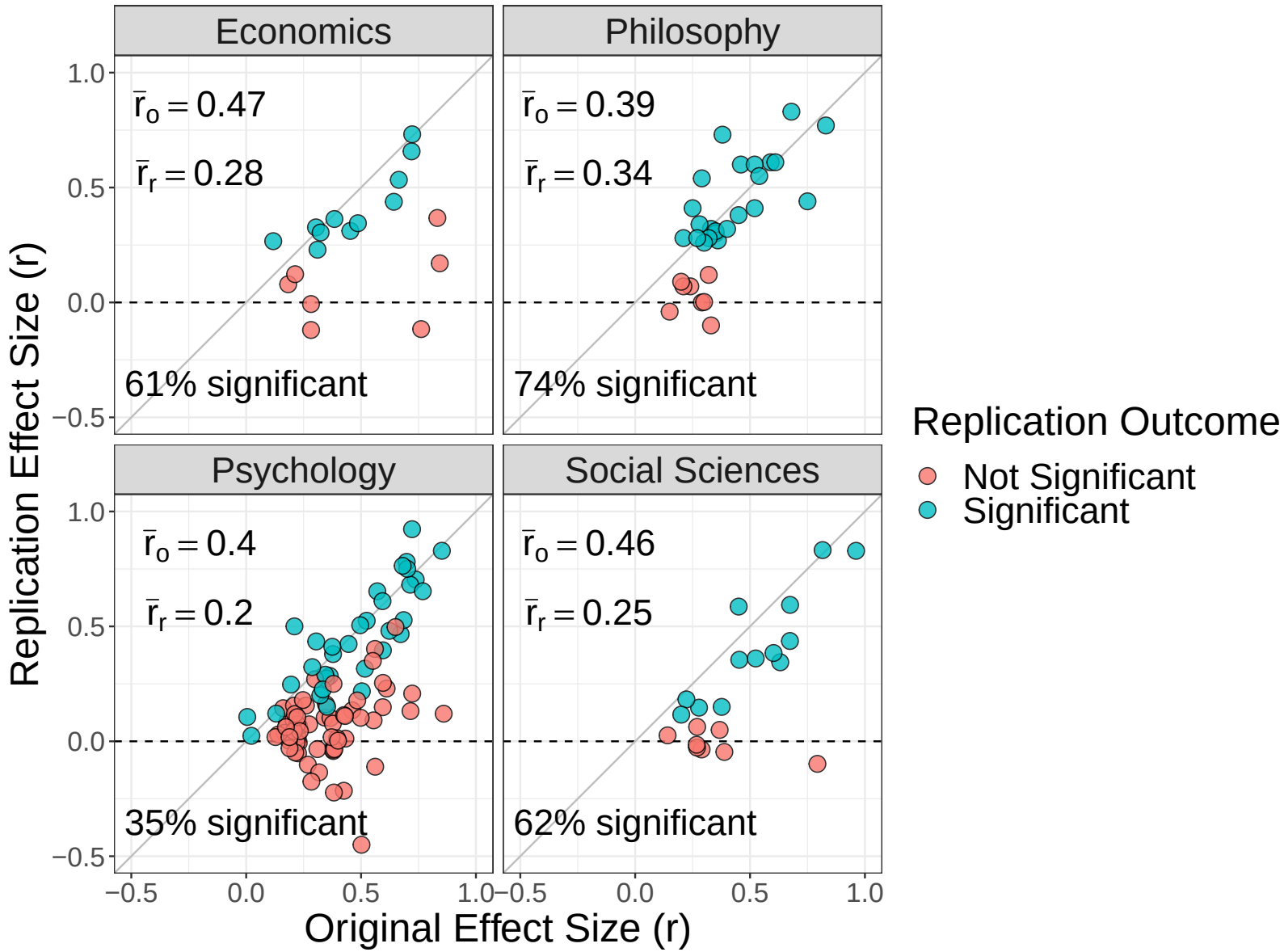
CrossMark

Estimating the Reproducibility of Experimental Philosophy

Florian Cova[1,2] · Brent Strickland[3,4] · Angela Abatista[5] · Aurélien Allard[6] · James Andow[7] · Mario Attie[8] · James Beebe[9] · Renatas Berniūnas[10] · Jordane Boudesseul[11] · Matteo Colombo[12] · Fiery Cushman[13] · Rodrigo Diaz[14] · Noah N'Djaye Nikolai van Dongen[15] · Vilius Dranseika[16] · Brian D. Earp[17] · Antonio Gaitán Torres[18] · Ivar Hannikainen[19] · José V. Hernández-Conde[20] · Wenjia Hu[21] · François Jaquet[1] · Kareem Khalifa[22] · Hanna Kim[23] · Markus Kneer[24] · Joshua Knobe[25] · Miklos Kurthy[26] · Anthony Lantian[27] · Shen-yi Liao[28] · Edouard Machery[29] · Tania Moerenhout[30] · Christian Mott[25] · Mark Phelan[21] · Jonathan Phillips[13] · Navin Rambharose[21] · Kevin Reuter[31] · Felipe Romero[15] · Paulo Sousa[32] · Jan Sprenger[33] · Emile Thalabard[34] · Kevin Tobia[25] · Hugo Viciana[35] · Daniel Wilkenfeld[29] · Xiang Zhou[36]



Original vs. Replication Effect Sizes



Can we predict the replication studies outcomes?

Prediction Methods

Objective

- Have $\hat{\theta}_o$, effect size of original study
- Want to predict $\hat{\theta}_r$, effect size of replication study

Prediction Methods

Objective

- Have $\hat{\theta}_o$, effect size of original study
- Want to predict $\hat{\theta}_r$, effect size of replication study

Assumptions

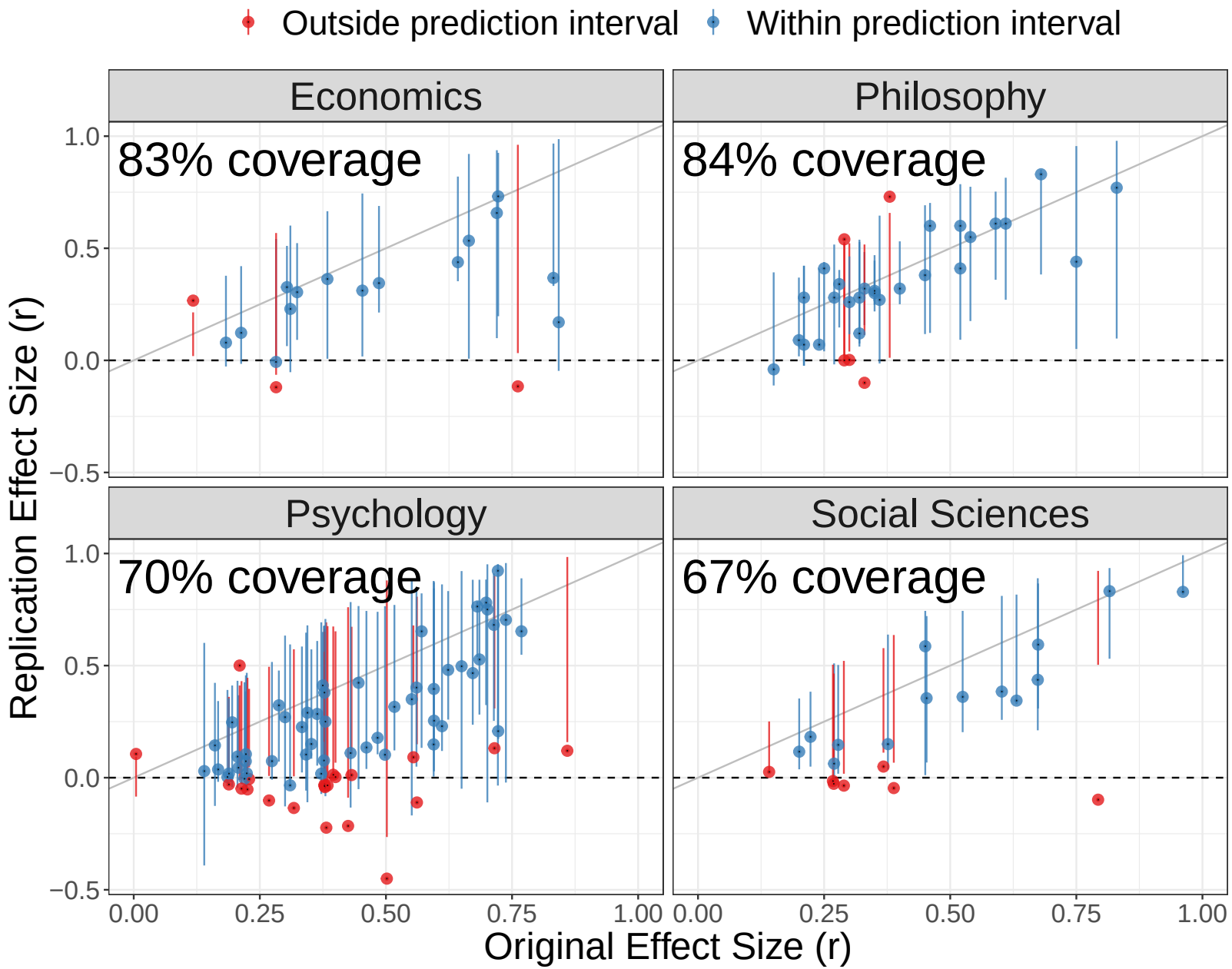
- Model $\hat{\theta}_o$ and $\hat{\theta}_r$ by normal distribution
- Standard errors σ_o and σ_r assumed to be known

Prediction Methods

Patil et al. (2016)

$$\hat{\theta}_r \mid \hat{\theta}_o \sim \mathsf{N}(\hat{\theta}_o, \sigma_o^2 + \sigma_r^2)$$

95% Prediction Intervals



Prediction Methods

Patil et al. (2016)

– Flat initial prior for θ

$$\hat{\theta}_r \mid \hat{\theta}_o \sim \mathsf{N}(\hat{\theta}_o, \sigma_o^2 + \sigma_r^2)$$

Prediction Methods

Patil et al. (2016)

- Flat initial prior for θ

$$\hat{\theta}_r \mid \hat{\theta}_o \sim \mathrm{N}(\hat{\theta}_o, \sigma_o^2 + \sigma_r^2)$$

How to obtain a better prediction?

- *Sceptical* prior: $\theta \sim \mathrm{N}(0, g \cdot \sigma_o^2),\ g \geq 0$

Prediction Methods

Patil et al. (2016)

- Flat initial prior for θ

$$\hat{\theta}_r \,|\, \hat{\theta}_o \sim \mathrm{N}(\hat{\theta}_o, \sigma_o^2 + \sigma_r^2)$$

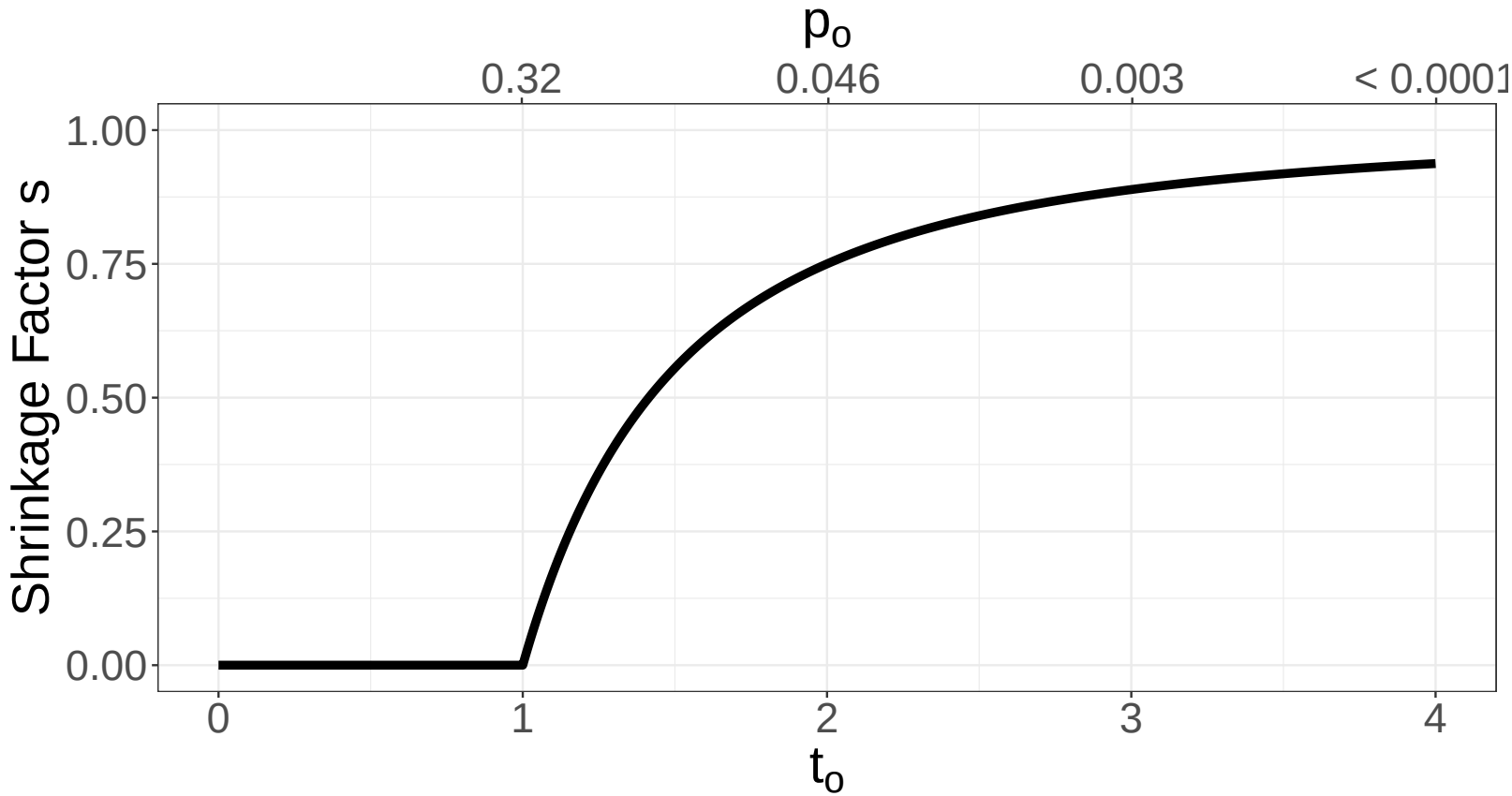
How to obtain a better prediction?

- *Sceptical* prior: $\theta \sim \mathrm{N}(0, g \cdot \sigma_o^2),\ g \geq 0$
- Estimate g by empirical Bayes

$\rightarrow$ *evidence-based shrinkage s*

$$\hat{\theta}_r \,|\, \hat{\theta}_o \sim \mathrm{N}(\boldsymbol{s} \cdot \hat{\theta}_o, \boldsymbol{s} \cdot \sigma_o^2 + \sigma_r^2)$$

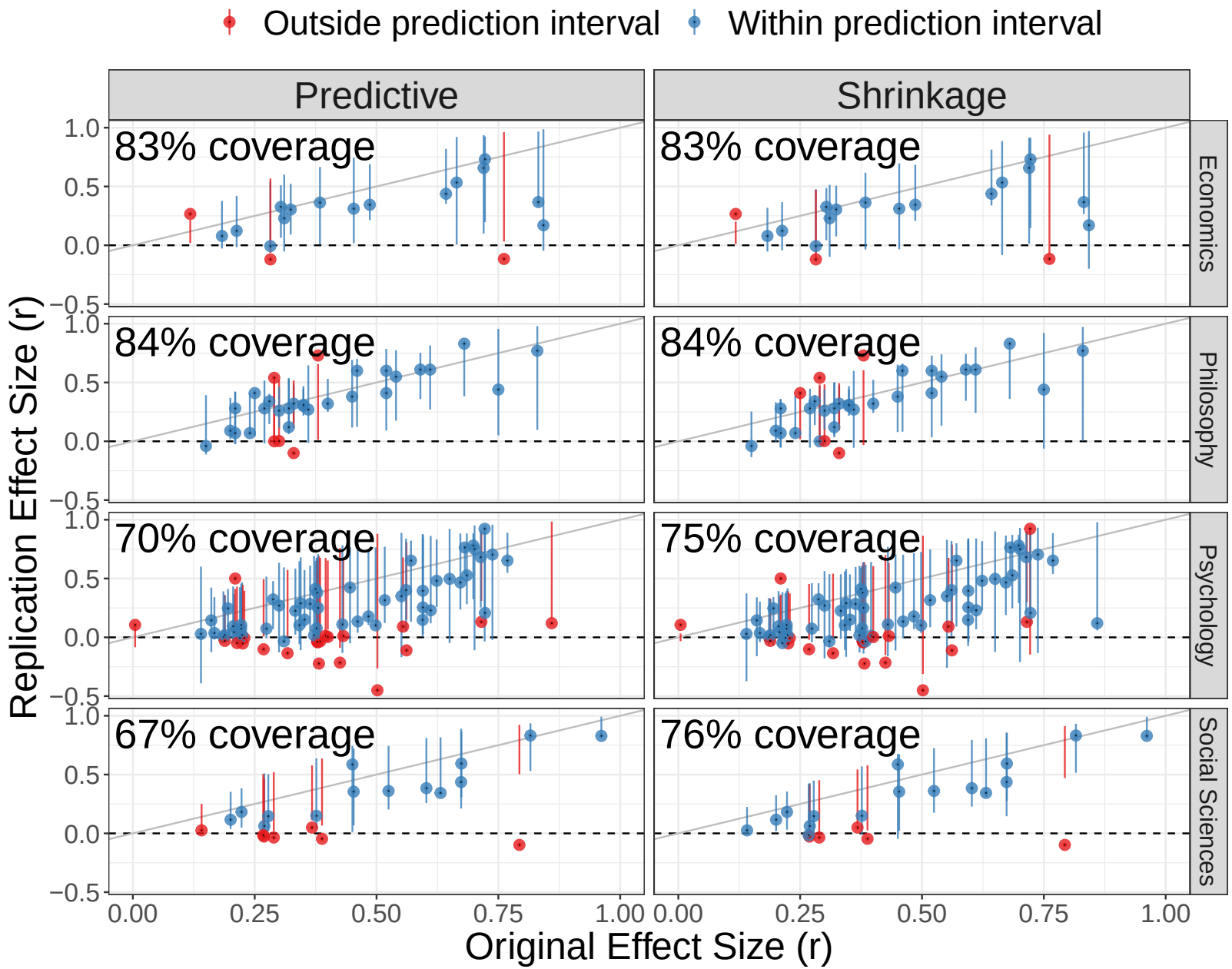
Evidence-Based Shrinkage



$$s = \max\left\{1 - 1/t_o^2, 0\right\}, \; t_o = \hat{\theta}_o/\sigma_o$$

same shrinkage factor as proposed by Copas (1997)

95% Prediction Intervals



Prediction Methods

Shrinkage predictive distribution

$$\hat{\theta}_r \mid \hat{\theta}_o \sim \mathsf{N}(s \cdot \hat{\theta}_o, s \cdot \sigma_o^2 + \sigma_r^2)$$

Prediction Methods

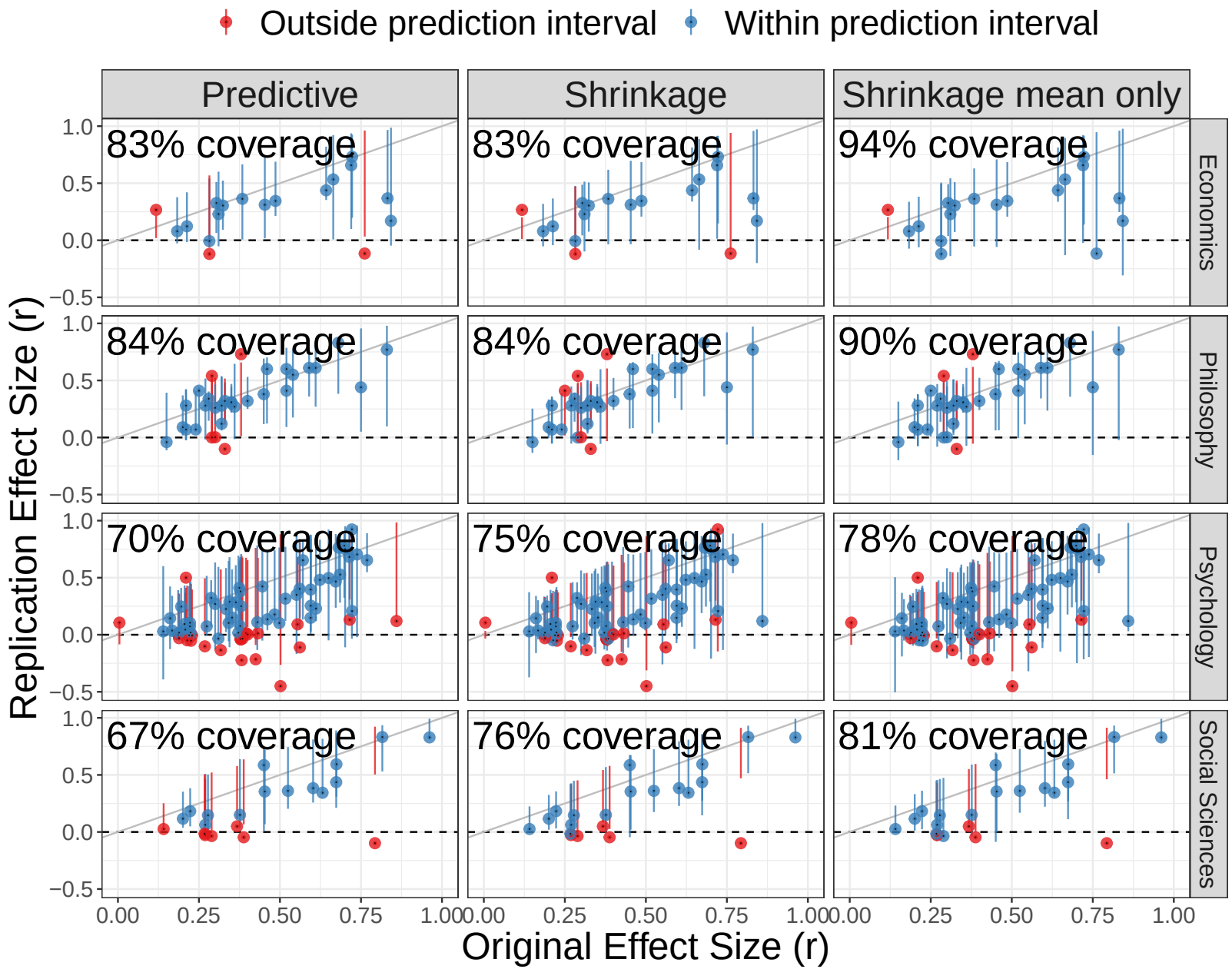
Shrinkage predictive distribution

$$\hat{\theta}_r \,|\, \hat{\theta}_o \sim \mathrm{N}(s \cdot \hat{\theta}_o, s \cdot \sigma_o^2 + \sigma_r^2)$$

Shrinkage *mean only* predictive distribution

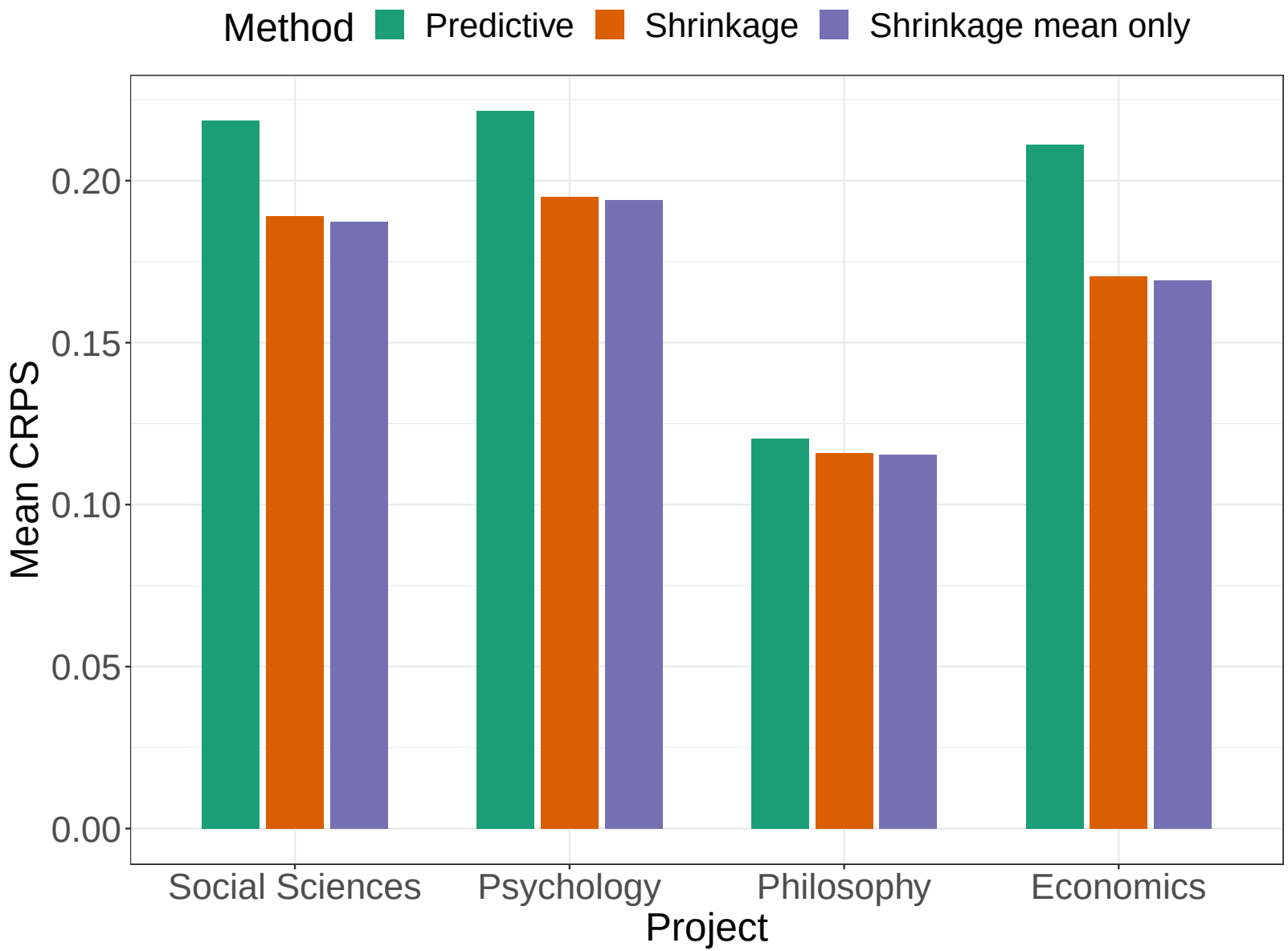
$$\hat{\theta}_r \,|\, \hat{\theta}_o \sim \mathrm{N}(s \cdot \hat{\theta}_o, \sigma_o^2 + \sigma_r^2)$$

95% Prediction Intervals



Continuous Ranked Probability Score

(CRPS negatively oriented)



Expected vs. Observed Significant Replications

Project	Method	N	Observed	Expected	*p*-value
Economics	Predictive	18	11	15.0	0.012
	Shrinkage	18	11	13.6	0.16
	Shrinkage mean only	18	11	13.4	0.19
Philosophy	Predictive	31	23	27.8	0.004
	Shrinkage	31	23	26.2	0.11
	Shrinkage mean only	31	23	26.0	0.14
Psychology	Predictive	73	24	55.4	< 0.0001
	Shrinkage	73	24	49.2	< 0.0001
	Shrinkage mean only	73	24	49.2	< 0.0001
Social Sciences	Predictive	21	13	19.9	< 0.0001
	Shrinkage	21	13	19.2	< 0.0001
	Shrinkage mean only	21	13	18.9	< 0.0001

Conclusions

Can we predict the outcome of the replications?

- Predictive performance depends on replication project

Conclusions

Can we predict the outcome of the replications?

- Predictive performance depends on replication project
- Evidence based shrinkage $\rightarrow$ Better prediction

Conclusions

Can we predict the outcome of the replications?

- Predictive performance depends on replication project
- Evidence based shrinkage $\rightarrow$ Better prediction

What can we learn from this?

- Non-significance likely from predictive perspective

Conclusions

Can we predict the outcome of the replications?

- Predictive performance depends on replication project
- Evidence based shrinkage $\rightarrow$ Better prediction

What can we learn from this?

- Non-significance likely from predictive perspective
- Evidence based shrinkage useful for other purposes

Conclusions

Limitations

- Approximation with normal distribution
- Small number of studies
- Data from academia

References

Camerer, C., Dreber, A., Holzmeister, F., Ho, T., Huber, J., Johannesson, M., Kirchler, M., Nave, G., Nosek, B., Pfeiffer, T., Altmejd, A., Buttrick, N., Chan, T., Chen, Y., Forsell, E., Gampa, A., Heikenstein, E., Hummer, L., Imai, T., Isaksson, S., Manfredi, D., Rose, J., Wagenmakers, E., and Wu, H. (2018). Evaluating the replicability of social science experiments in nature and science between 2010 and 2015. *Nature Human Behavior*, 2:637 – 644.

Camerer, C. F., Dreber, A., Forsell, E., Ho, T., 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*, 351:1433 – 1436.

Copas, J. B. (1997). Using regression models for prediction: shrinkage and regression to the mean. *Statistical Methods in Medical Research*, 6:167 – 183.

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.

Gneiting, T. and Raftery, E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the Amerian Statistical Association*, 102:359 – 377.

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science*, 349.

Patil, P., Peng, R. D., and Leek, J. T. (2016). What should researchers expect when they replicate studies? a statistical view of replicability in psychological science. *Perspectives on Psychological Science*, 11:539 – 544.

Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with *g*-prior distributions. *Bayesian Inference and Decision techniques*, 6:233 – 243.