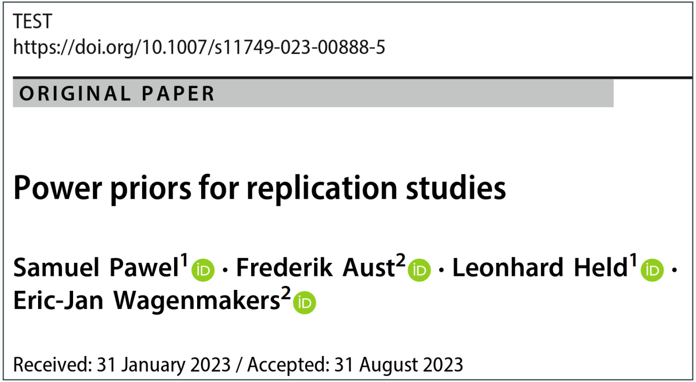
Power Priors for Replication Studies

Samuel Pawel[1], Frederik Aust[2], Leonhard Held[1], Eric-Jan Wagenmakers[2]

[1] Epidemiology, Biostatistics and Prevention Institute (EBPI), Center for Reproducible Science (CRS), University of Zurich
[2] Department of Psychological Methods, University of Amsterdam

October 26, 2023, Bayesian Biostatistics Conference, Utrecht

This talk



TEST
https://doi.org/10.1007/s11749-023-00888-5

ORIGINAL PAPER

Power priors for replication studies

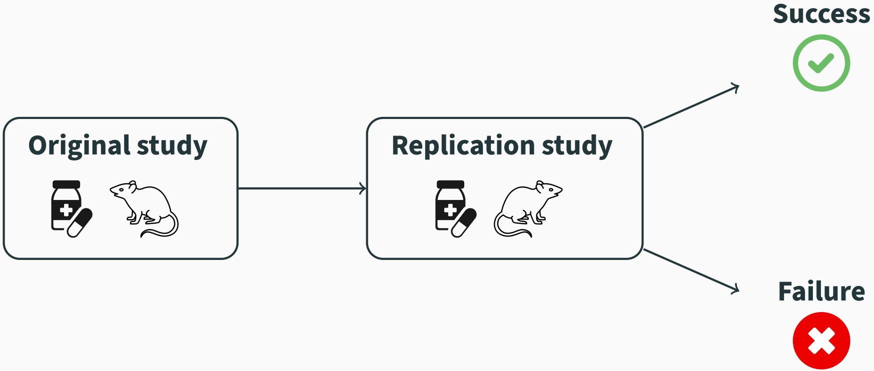
Samuel Pawel[1] · Frederik Aust[2] · Leonhard Held[1] · Eric-Jan Wagenmakers[2]

Received: 31 January 2023 / Accepted: 31 August 2023

Replication studies

Replicability

Obtaining similar results when repeating a study with new subjects



Replication studies in drug development

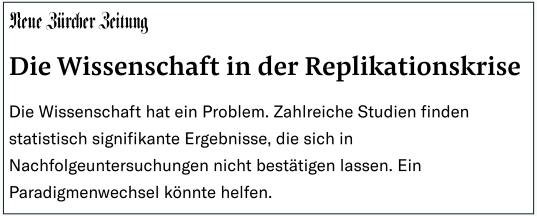
Two-trials rule

“*… at least two adequate and well-controlled studies, each convincing on its own, to establish effectiveness …*” (FDA, 1998)

FDA U.S. FOOD & DRUG ADMINISTRATION

EUROPEAN MEDICINES AGENCY
SCIENCE MEDICINES HEALTH

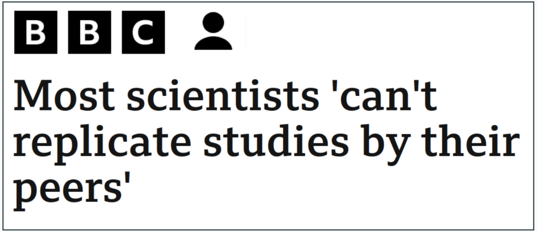
Replication crisis



Neue Zürcher Zeitung

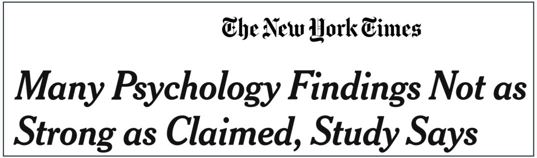
Die Wissenschaft in der Replikationskrise

Die Wissenschaft hat ein Problem. Zahlreiche Studien finden statistisch signifikante Ergebnisse, die sich in Nachfolgeuntersuchungen nicht bestätigen lassen. Ein Paradigmenwechsel könnte helfen.



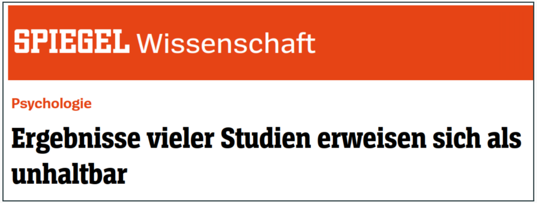
BBC

Most scientists 'can't replicate studies by their peers'



The New York Times

Many Psychology Findings Not as Strong as Claimed, Study Says



SPIEGEL Wissenschaft

Psychologie

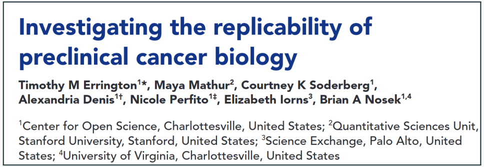
Ergebnisse vieler Studien erweisen sich als unhaltbar

Kovic (2016); Spiegel (2015); Carey (2015); Feilden (2017)

Replication crisis

Large-scale replication projects

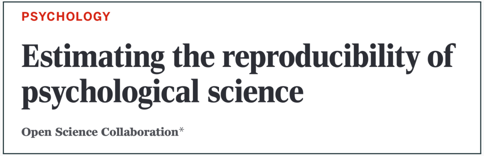
- Cancer biology: 42/97 = **43%** successful
- Psychology: 36/100 = **36%** successful
- Economics: 11/18 = **61%** successful
- Social sciences: 13/21 = **62%** successful



Investigating the replicability of preclinical cancer biology

Timothy M Errington[1*], Maya Mathur[2], Courtney K Soderberg[1], Alexandria Denis[1†], Nicole Perfito[1‡], Elizabeth Iorns[3], Brian A Nosek[1,4]

[1]Center for Open Science, Charlottesville, United States; [2]Quantitative Sciences Unit, Stanford University, Stanford, United States; [3]Science Exchange, Palo Alto, United States; [4]University of Virginia, Charlottesville, United States



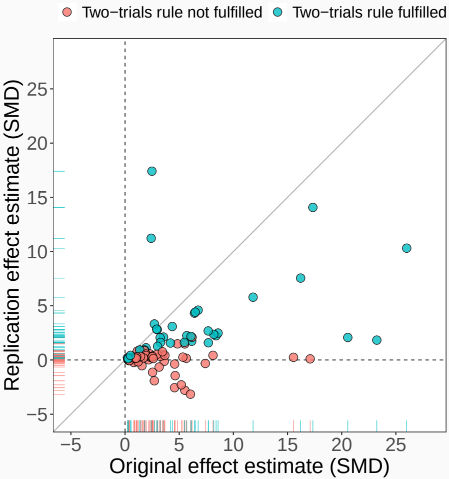
PSYCHOLOGY

Estimating the reproducibility of psychological science

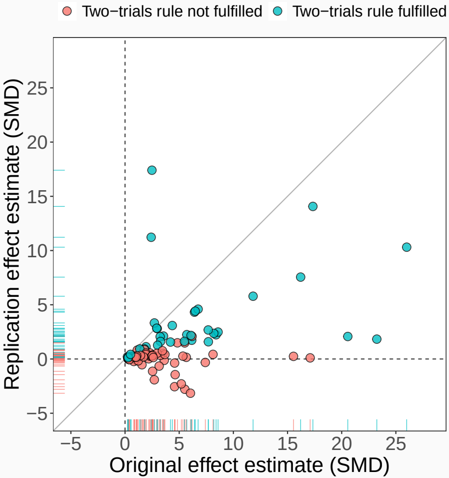
Open Science Collaboration*

Errington et al. (2021); Open Science Collaboration (2015); Camerer et al. (2016, 2018)

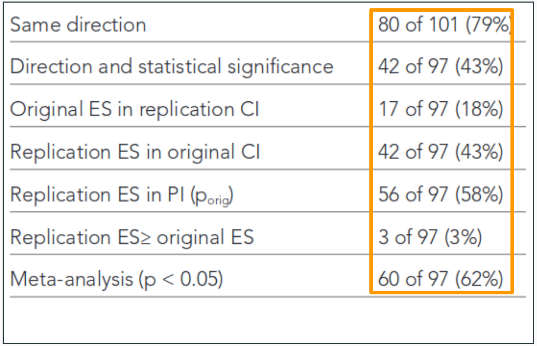
Reproducibility Project: Cancer Biology (Errington et al., 2021)



Reproducibility Project: Cancer Biology (Errington et al., 2021)



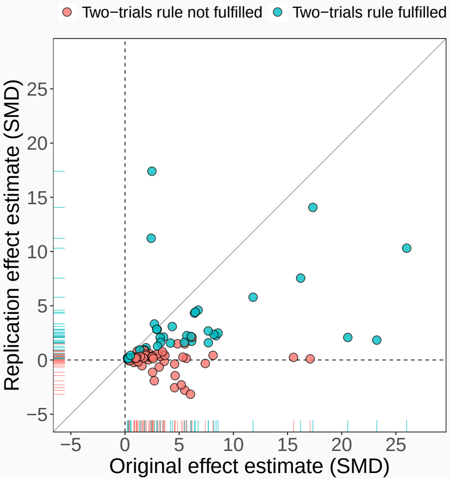
Other replicability criteria



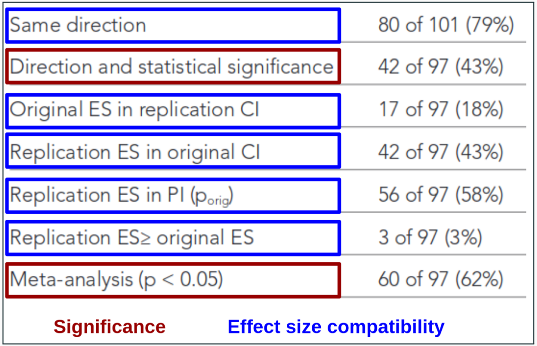
Same direction	80 of 101 (79%)
Direction and statistical significance	42 of 97 (43%)
Original ES in replication CI	17 of 97 (18%)
Replication ES in original CI	42 of 97 (43%)
Replication ES in PI (p_{orig})	56 of 97 (58%)
Replication ES≥ original ES	3 of 97 (3%)
Meta-analysis ($p < 0.05$)	60 of 97 (62%)

excerpt from Table 1 in Errington et al. (2021)

Reproducibility Project: Cancer Biology (Errington et al., 2021)



Other replicability criteria



Same direction	80 of 101 (79%)
Direction and statistical significance	42 of 97 (43%)
Original ES in replication CI	17 of 97 (18%)
Replication ES in original CI	42 of 97 (43%)
Replication ES in PI (p_{orig})	56 of 97 (58%)
Replication ES≥ original ES	3 of 97 (3%)
Meta-analysis ($p < 0.05$)	60 of 97 (62%)

Significance **Effect size compatibility**

excerpt from Table 1 in Errington et al. (2021)

→ Can different notions of replicability be assessed in a unified framework?

Power priors for replication studies

Setup

- Original and replication effect estimates $\hat{\theta}_o$ and $\hat{\theta}_r$ of unknown effect size θ
- Standard errors σ_o and σ_r

→ Normality assumption $\hat{\theta}_i \,|\, \theta \sim \mathrm{N}(\theta, \sigma_i^2)$ for $i \in \{o, r\}$

Power priors for replication studies

Setup

- Original and replication effect estimates $\hat{\theta}_o$ and $\hat{\theta}_r$ of unknown effect size θ
- Standard errors σ_o and σ_r

→ Normality assumption $\hat{\theta}_i \,|\, \theta \sim \mathrm{N}(\theta, \sigma_i^2)$ for $i \in \{o, r\}$

Normalized power prior (Duan et al., 2005; Neuenschwander et al., 2009)

- Prior for effect size

$$\pi(\theta \,|\, \hat{\theta}_o, \alpha) = \frac{f(\hat{\theta}_o \,|\, \theta)^\alpha}{\int f(\hat{\theta}_o \,|\, \theta)^\alpha \, \mathrm{d}\theta} = \mathrm{N}(\theta \,|\, \hat{\theta}_o, \sigma_o^2/\alpha)$$

- Prior for power parameter $\pi(\alpha) = \mathrm{Beta}(\alpha \,|\, x, y)$

→ Joint prior $\pi(\theta, \alpha \,|\, \hat{\theta}_o) = \mathrm{N}(\theta \,|\, \hat{\theta}_o, \sigma_o^2/\alpha) \times \mathrm{Beta}(\alpha \,|\, x, y)$

Replicability inferences based on power priors

Parameter estimation

- Marginal posterior of $\theta \rightarrow$ Effect estimation
- Marginal posterior of $\alpha \rightarrow$ Compatibility estimation

Replicability inferences based on power priors

Parameter estimation

- Marginal posterior of $\theta \rightarrow$ Effect estimation
- Marginal posterior of $\alpha \rightarrow$ Compatibility estimation

Hypothesis testing

- Bayes factor test related to $\theta \rightarrow$ Effect test
- Bayes factor test related to $\alpha \rightarrow$ Compatibility test

Replicability inferences based on power priors

Parameter estimation

- Marginal posterior of $\theta \rightarrow$ Effect estimation
- Marginal posterior of $\alpha \rightarrow$ Compatibility estimation

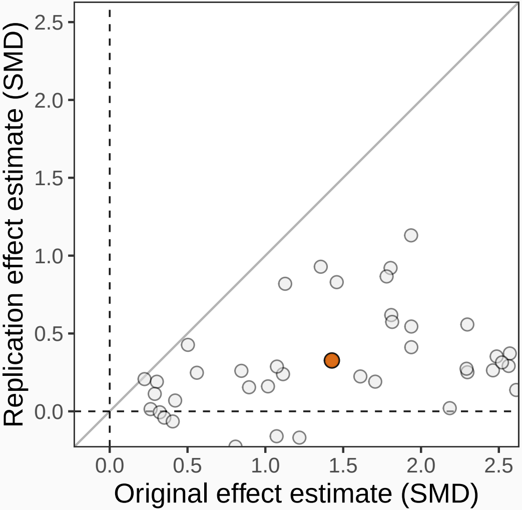
Hypothesis testing

- Bayes factor test related to $\theta \rightarrow$ Effect test
- Bayes factor test related to $\alpha \rightarrow$ Compatibility test

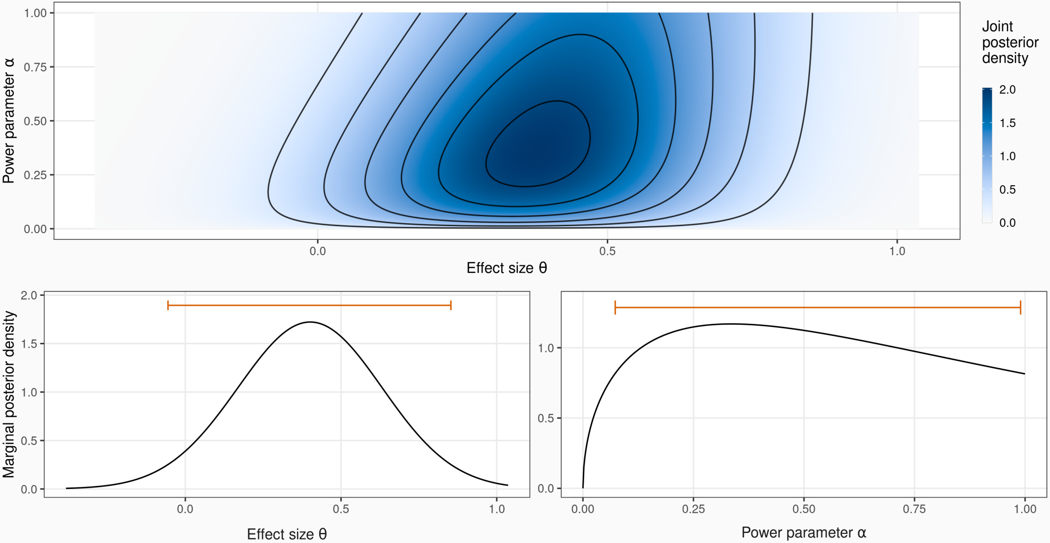
R implementation

- R package **ppRep** (`https://CRAN.R-project.org/package=ppRep`)
- Only numerical integration required, no (MC)MC needed

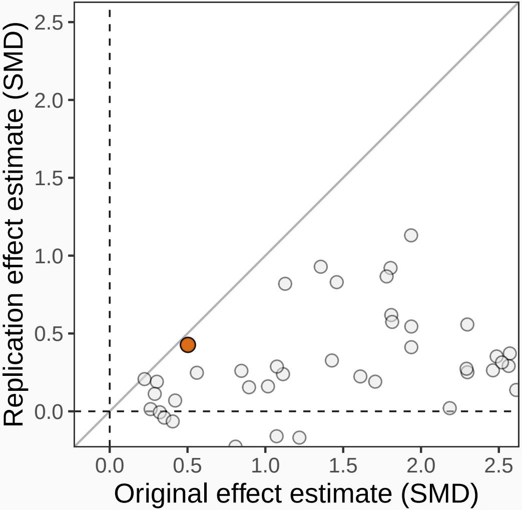
A replication failure $(\hat{\theta}_o = 1.43, \hat{\theta}_r = 0.33, \sigma_o = 0.62, \sigma_r = 0.24)$



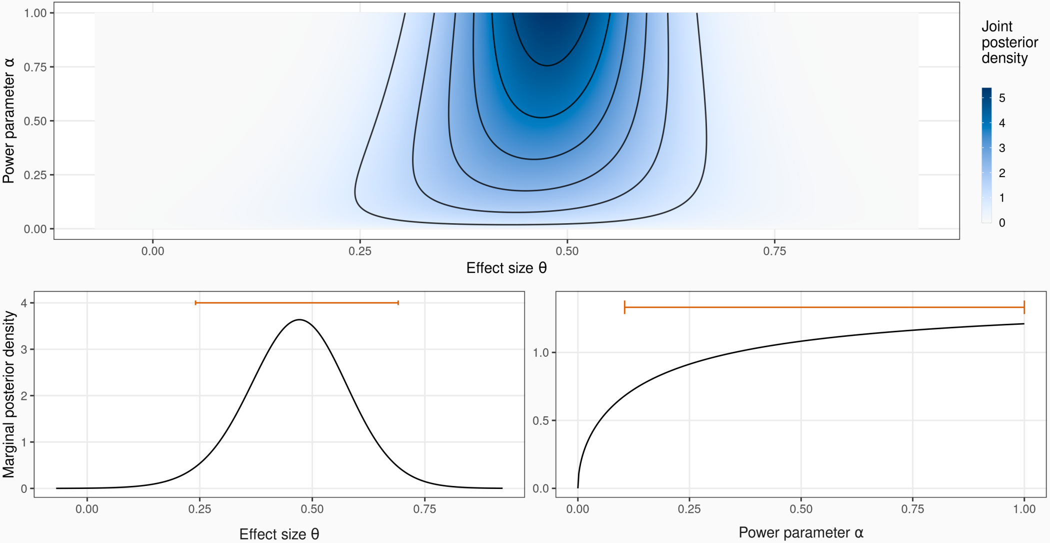
A replication failure ($\hat{\theta}_o = 1.43, \hat{\theta}_r = 0.33, \sigma_o = 0.62, \sigma_r = 0.24$)



An almost perfect replication ($\hat{\theta}_o = 0.5, \hat{\theta}_r = 0.43, \sigma_o = 0.11, \sigma_r = 0.17$)



An almost perfect replication ($\hat{\theta}_o = 0.5, \hat{\theta}_r = 0.43, \sigma_o = 0.11, \sigma_r = 0.17$)



Why does the posterior for α hardly change?

Received: 10 May 2023 | Accepted: 31 May 2023

DOI: 10.1002/sta4.591

Stat. 2023;12:e591.
https://doi.org/10.1002/sta4.591

ORIGINAL ARTICLE

Normalized power priors always discount historical data

Samuel Pawel[1] | Frederik Aust[2] | Leonhard Held[1] | Eric-Jan Wagenmakers[2]

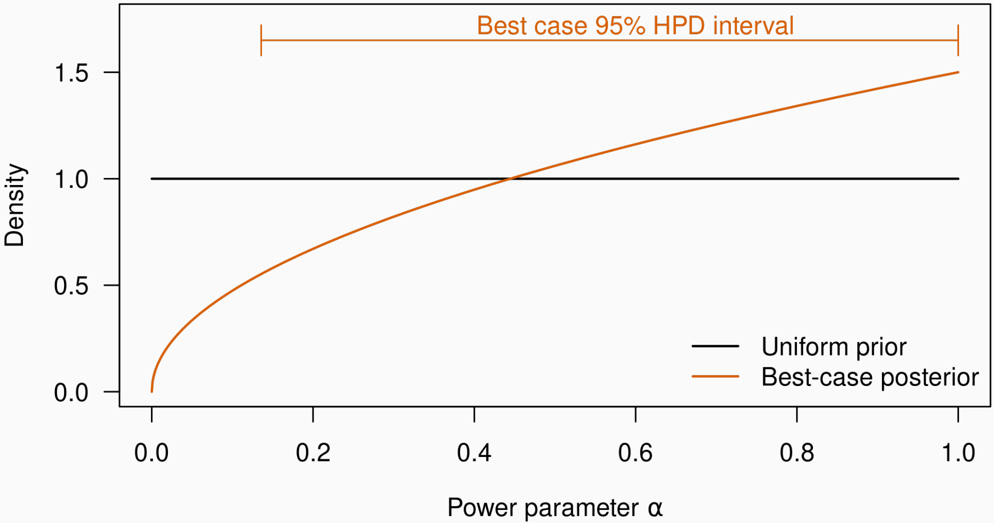
Limiting posterior of α

Replication data agree perfectly ($\hat{\theta}_o = \hat{\theta}_r$) and prior $\alpha \sim \text{Beta}(x, y)$

$\Rightarrow$ Posterior approaches $\alpha \sim \text{Beta}(x + 1/2, y)$ as replication standard error $\sigma_r \downarrow 0$

$\Rightarrow$ Complete pooling is impossible

Why does the posterior for α hardly change?



Is there an intuition in terms of hierarchical models?

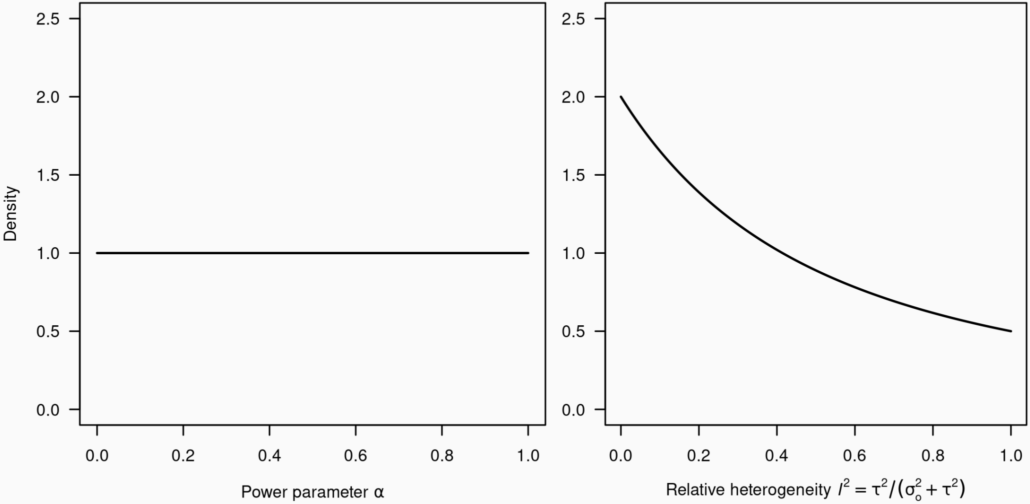
Normal-normal hierarchical model

- $\hat{\theta}_i \mid \theta_i \sim \mathrm{N}(\theta_i, \sigma_i^2)$ for $i \in \{o, r\}$
- $\theta_i \mid \tau^2 \sim \mathrm{N}(\mu, \tau^2)$
- $\pi(\mu) \propto 1$ and prior for τ^2 or $I^2 = \tau^2/(\sigma_o^2 + \tau^2)$

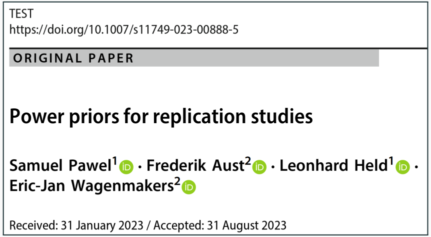
Matching effect size posterior between power prior and hierarchical model

For a beta prior $\alpha \sim \mathrm{Beta}(x, y)$ assigned to power parameter α
$\Rightarrow$ The posteriors for θ and θ_r match if generalized beta prior $I^2 \sim \mathrm{GenBeta}(y, x, 2)$ is assigned in the hierarchical model
$\Rightarrow$ α acts as a relative heterogeneity variance parameter similar to I^2

Is there an intuition in terms of hierarchical models?



Conclusions



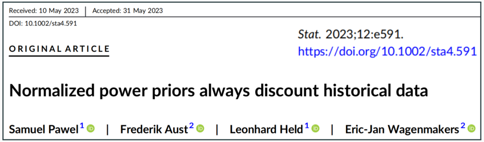
TEST
https://doi.org/10.1007/s11749-023-00888-5

ORIGINAL PAPER

Power priors for replication studies

Samuel Pawel[1] · Frederik Aust[2] · Leonhard Held[1] · Eric-Jan Wagenmakers[2]

Received: 31 January 2023 / Accepted: 31 August 2023



Received: 10 May 2023 | Accepted: 31 May 2023
DOI: 10.1002/sta4.591

ORIGINAL ARTICLE

Stat. 2023;12:e591.
https://doi.org/10.1002/sta4.591

Normalized power priors always discount historical data

Samuel Pawel[1] | Frederik Aust[2] | Leonhard Held[1] | Eric-Jan Wagenmakers[2]

→ Power prior framework gives suite of methods to assess replicability

→ Complete pooling of data sets impossible when beta prior assigned to α

→ Power parameter α similar role as relative heterogeneity I^2 in meta-analysis

→ R package **ppRep** (`https://CRAN.R-project.org/package=ppRep`)

References i

Camerer, C. F., Dreber, A., Forsell, E., Ho, T., Huber, J., Johannesson, M., Kirchler, M., Almenberg, J., Altmejd, A., et al. (2016). Evaluating replicability of laboratory experiments in economics. *Science*, 351(6280):1433–1436. doi:10.1126/science.aaf0918.

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T., Huber, J., Johannesson, M., Kirchler, M., Nave, G., Nosek, B., et al. (2018). Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behavior*, 2(9):637–644. doi:10.1038/s41562-018-0399-z.

Carey, B. (2015). Many psychology findings not as strong as claimed, study says. *The New York Times*. URL `https://www.nytimes.com/2015/08/28/science/many-social-science-findings-not-as-strong-as-claimed-study-says.html`.

Duan, Y., Ye, K., and Smith, E. P. (2005). Evaluating water quality using power priors to incorporate historical information. *Environmetrics*, 17(1):95–106. doi:10.1002/env.752.

Errington, T. M., Mathur, M., Soderberg, C. K., Denis, A., Perfito, N., Iorns, E., and Nosek, B. A. (2021). Investigating the replicability of preclinical cancer biology. *eLife*, 10:e71601. doi:10.7554/elife.71601.

References ii

FDA (1998). Providing clinical evidence of effectiveness for human drug and biological products. URL `www.fda.gov/regulatory-information/search-fda-guidance-documents/providing-clinical-evidence-effectiveness-human-drug-and-biological-products`.

Feilden, T. (2017). Most scientists 'can't replicate studies by their peers'. *BBC*. URL `https://www.bbc.com/news/science-environment-39054778`.

Kovic, M. (2016). Die Wissenschaft in der Replikationskrise. *Neue Zürcher Zeitung*. URL `https://www.nzz.ch/wissenschaft/physik/fallstricke-der-statistik-die-wissenschaft-in-der-replikationskrise-ld.86330`.

Neuenschwander, B., Branson, M., and Spiegelhalter, D. J. (2009). A note on the power prior. *Statistics in Medicine*, 28(28):3562–3566. doi:10.1002/sim.3722.

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251):aac4716. doi:10.1126/science.aac4716.

Pawel, S., Aust, F., Held, L., and Wagenmakers, E.-J. (2023a). Normalized power priors always discount historical data. *Stat*, 12(1):e591. doi:10.1002/sta4.591.

References iii

Pawel, S., Aust, F., Held, L., and Wagenmakers, E.-J. (2023b). Power priors for replication studies. *TEST*. doi:10.1007/s11749-023-00888-5.

Spiegel (2015). Ergebnisse vieler Studien erweisen sich als unhaltbar. URL `https://www.spiegel.de/wissenschaft/mensch/psychologie-ergebnisse-hunderter-studien-nicht-wiederholbar-a-1050202.html`.