

The Sceptical P -Value

Leonhard Held



University of
Zurich^{UZH}

Winter School 2022, Les Diablerets

A New Approach to Define Replication Success



J. R. Statist. Soc. A (2020)

A new standard for the analysis and design of replication studies

Leonhard Held

University of Zurich, Switzerland

[Read before The Royal Statistical Society at a meeting on 'Signs and sizes: understanding and replicating statistical findings' at the Society's 2019 annual conference in Belfast on Wednesday, September 4th, 2019, the President, Professor D. Ashby, in the Chair]

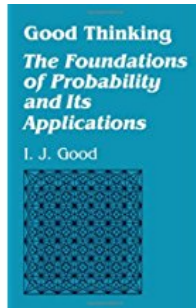
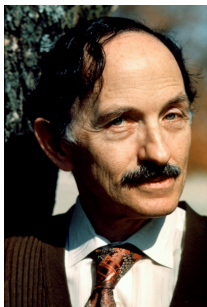
- A **Bayes/non-Bayes** compromise based on
 1. Reverse-Bayes analysis
 2. Prior criticism

→ The **sceptical p -value** p_S quantifies degree of **replication success**

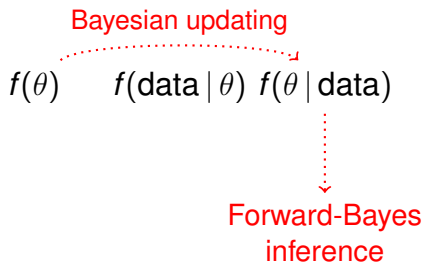
Step 1: Reverse-Bayes Analysis

Jack Good (1916-2009)

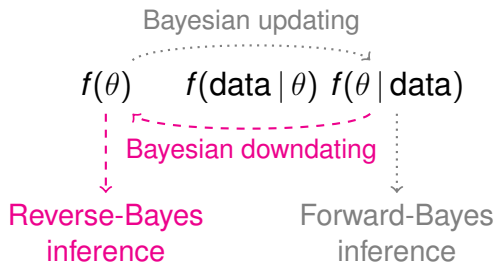
*“We can make judgments of initial probabilities and infer final ones, or we can equally make judgments of final ones and infer initial ones by **Bayes’s theorem in reverse.**”*



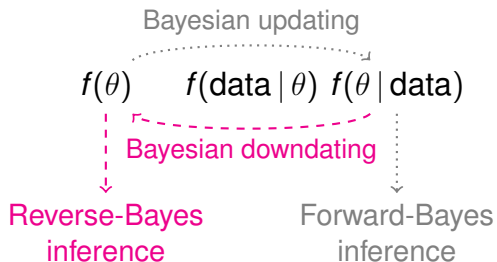
Forward- and Reverse-Bayes



Forward- and Reverse-Bayes



Forward- and Reverse-Bayes



Reverse-Bayes methods for evidence assessment and research synthesis

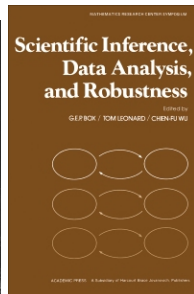
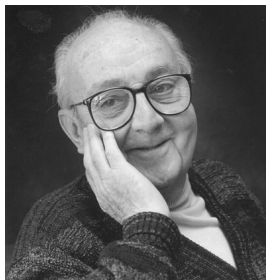
Leonhard Held ^{*,§}, Robert Matthews ^{†,§}, Manuela Ott ^{*,†}, and Samuel Pawel ^{*,§}

<https://arxiv.org/abs/2102.13443>

Step 2: Prior-Data Conflict

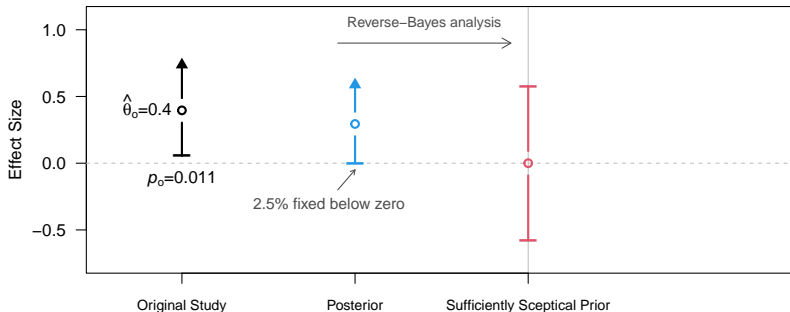
George Box (1919-2013)

“The process of scientific investigation involves not one but two kinds of inference: estimation and criticism, used iteratively and in alternation.”



The Proposed Approach: Step 1

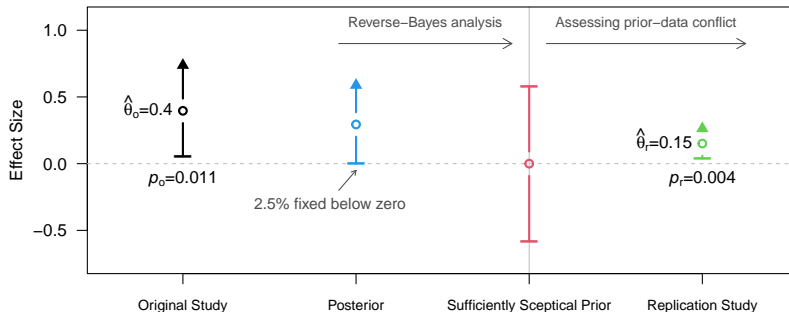
One-sided $\alpha = 2.5\%$



- Determine the variance τ^2 of a **sceptical prior** $N(0, \tau^2)$ that makes the original result no longer convincing.

The Proposed Approach: Step 2

One-sided $\alpha = 2.5\%$

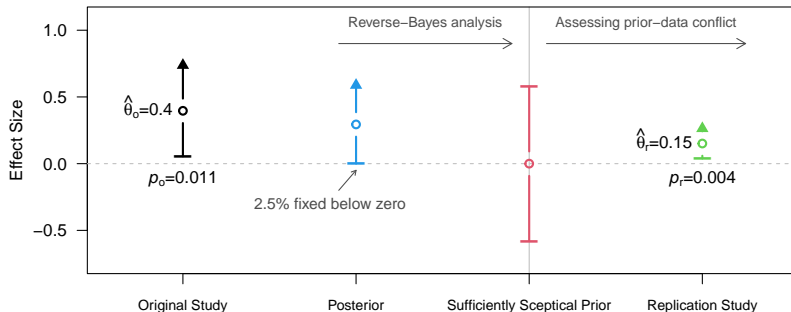


- **Prior-data conflict** is quantified based on the **prior-predictive distribution**:

$$p_{\text{Box}} = \Pr\{N(0, \tau^2 + \sigma_r^2) \geq \hat{\theta}_r\}.$$

The Proposed Approach: Step 2

One-sided $\alpha = 2.5\%$



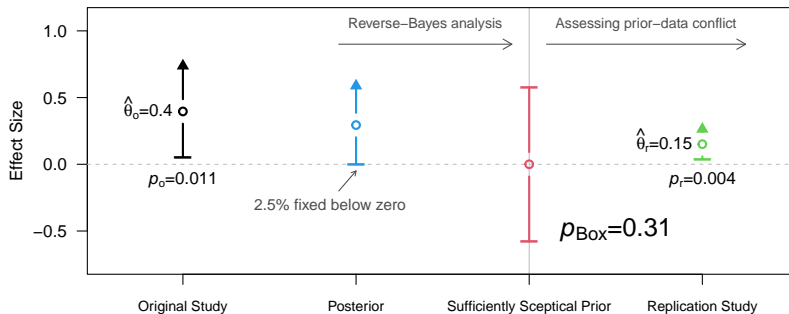
- **Prior-data conflict** is quantified based on the **prior-predictive distribution**:

$$p_{\text{Box}} = \Pr\{N(0, \tau^2 + \sigma_r^2) \geq \hat{\theta}_r\}.$$

- **Replication success** is achieved if $p_{\text{Box}} \leq \alpha$.

The Proposed Approach

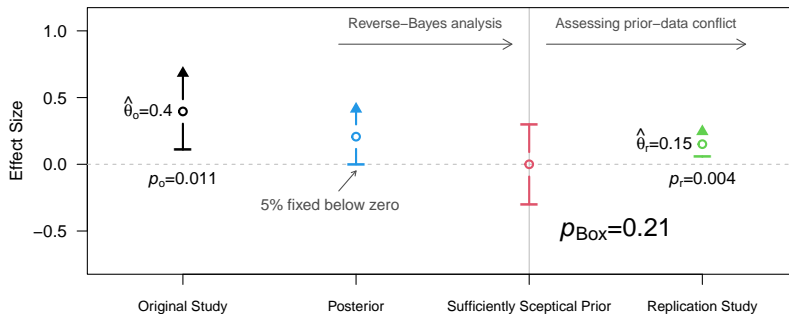
One-sided $\alpha = 2.5\%$



No replication success at level $\alpha = 2.5\%$

The Proposed Approach

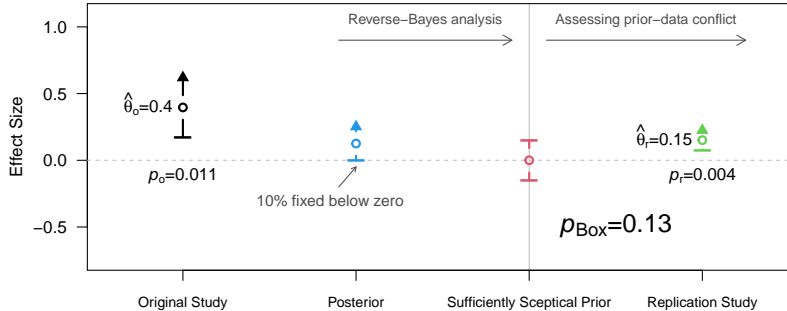
One-sided $\alpha = 5\%$



No replication success at level $\alpha = 5\%$

The Proposed Approach

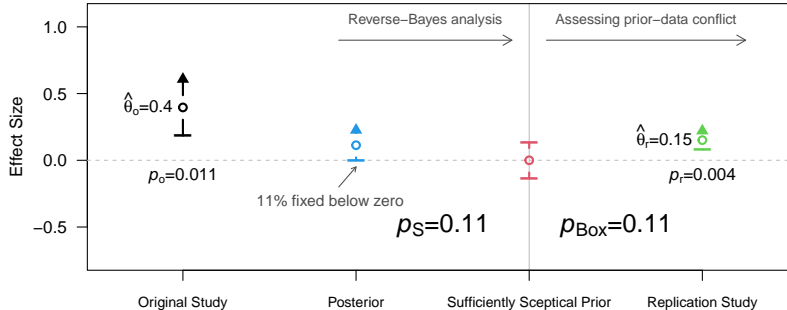
One-sided $\alpha = 10\%$



No replication success at level $\alpha = 10\%$

The Proposed Approach

One-sided $\alpha = 11\%$



Replication success at level $\alpha = 11\%$

The Sceptical p -Value

The **sceptical p -value** p_S is the smallest level α where replication success is achieved.

The Sceptical p -Value

The **sceptical p -value** p_S is the smallest level α where replication success is achieved.

- always exists, fulfills $p_S > \max\{p_o, p_r\}$

The Sceptical p -Value

The **sceptical p -value** p_S is the smallest level α where replication success is achieved.

- always exists, fulfills $p_S > \max\{p_o, p_r\}$
- can be computed analytically under standard normality assumptions

The Sceptical p -Value

The **sceptical p -value** p_S is the smallest level α where replication success is achieved.

- always exists, fulfills $p_S > \max\{p_o, p_r\}$
- can be computed analytically under standard normality assumptions
- depends on both p -values p_o and p_r and the **relative sample size** c

The Sceptical p -Value

The **sceptical p -value** p_S is the smallest level α where replication success is achieved.

- always exists, fulfills $p_S > \max\{p_o, p_r\}$
- can be computed analytically under standard normality assumptions
- depends on both p -values p_o and p_r and the **relative sample size** c
- has a particularly simple form for $c = 1$ with known null distribution

ROYAL
STATISTICAL
SOCIETY
DATA | EVIDENCE | DECISIONS



Journal of the Royal Statistical Society
Applied Statistics
Series C

Appl. Statist. (2020)

The harmonic mean χ^2 -test to substantiate scientific findings

Leonhard Held

University of Zurich, Switzerland

Replication Success in Terms of Relative Effect Size

Goal: Comparison of

- **sceptical p -value**
- **two-trials rule**
- **meta-analysis**

Replication Success in Terms of Relative Effect Size

Goal: Comparison of

- **sceptical p -value**
- **two-trials rule**
- **meta-analysis**

Key: Formulation in terms of

1. **Original p -value** p_o
2. **Relative effect size** $d = \hat{\theta}_r / \hat{\theta}_o$
3. **Relative sample size** $c = n_r / n_o$

Replication Success in Terms of Relative Effect Size

Goal: Comparison of

- **sceptical p -value**
- **two-trials rule**
- **meta-analysis**

Key: Formulation in terms of

1. **Original p -value p_o**
2. **Relative effect size $d = \hat{\theta}_r / \hat{\theta}_o$**
3. **Relative sample size $c = n_r / n_o$**

THE ASSESSMENT OF REPLICATION SUCCESS BASED ON RELATIVE EFFECT SIZE

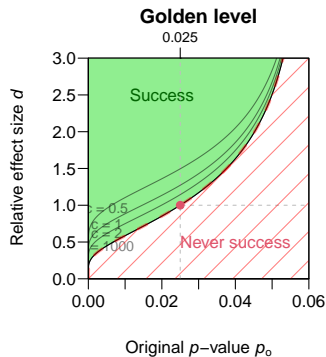
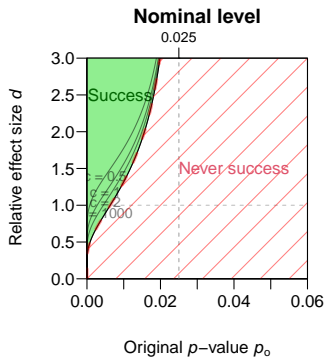
BY LEONHARD HELD, CHARLOTTE MICHELOUD AND SAMUEL PAWEL

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

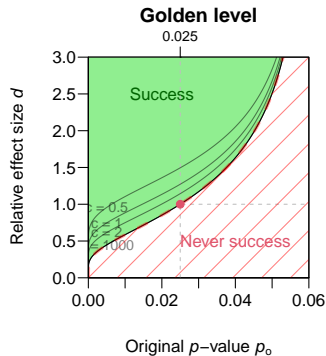
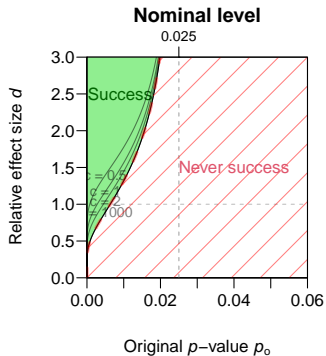
AOAS, to appear

<https://arxiv.org/abs/2009.07782>

Recalibration of the Sceptical p -Value



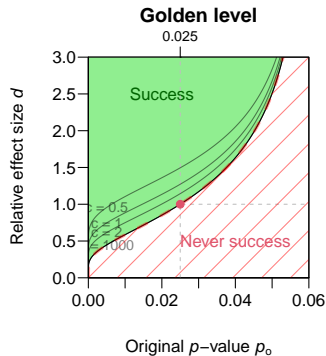
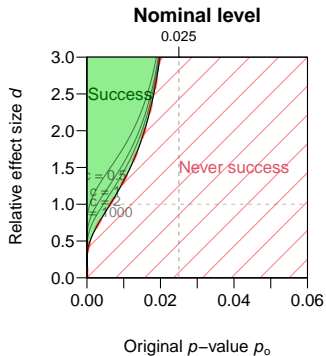
Recalibration of the Sceptical p -Value



For a borderline convincing original result ($p_o = 0.025$), replication success

- is impossible at the **nominal level**.

Recalibration of the Sceptical p -Value



For a borderline convincing original result ($p_o = 0.025$), replication success

- is impossible at the **nominal level**.
- is possible at the **golden level** if the relative effect size is larger than one.

The Golden Level

Replication success:

$$p_S < \alpha_S$$

$$\alpha_S = 1 - \Phi(z_\alpha / \sqrt{\varphi})$$

$$\text{with golden ratio } \varphi = (\sqrt{5} + 1)/2 \approx 1.62$$

The Golden Level

Replication success:

$$p_S < \alpha_S$$

$$\alpha_S = 1 - \Phi(z_\alpha / \sqrt{\varphi})$$

$$\text{with golden ratio } \varphi = (\sqrt{5} + 1)/2 \approx 1.62$$

- For example, for $\alpha = 0.025$ we obtain $\alpha_S = 0.062$.

The Golden Level

Replication success:

$$p_S < \alpha_S$$

$$\alpha_S = 1 - \Phi(z_\alpha / \sqrt{\varphi})$$

$$\text{with golden ratio } \varphi = (\sqrt{5} + 1)/2 \approx 1.62$$

- For example, for $\alpha = 0.025$ we obtain $\alpha_S = 0.062$.
- Equivalently, the sceptical p -value can be **recalibrated**:

$$\tilde{p}_S = 1 - \Phi(z_S \sqrt{\varphi})$$

(default in the R-package `ReplicationSuccess`)

The Golden Level

Replication success:

$$p_S < \alpha_S$$

$$\alpha_S = 1 - \Phi(z_\alpha / \sqrt{\varphi})$$

$$\text{with golden ratio } \varphi = (\sqrt{5} + 1)/2 \approx 1.62$$

- For example, for $\alpha = 0.025$ we obtain $\alpha_S = 0.062$.
- Equivalently, the sceptical p -value can be **recalibrated**:

$$\tilde{p}_S = 1 - \Phi(z_S \sqrt{\varphi})$$

(default in the R-package `ReplicationSuccess`)

- For example, for $p_S = 0.11$ we obtain $\tilde{p}_S = 0.061$

Type-I Error Rate and Project Power

