Rethinking Replicability: a Unified Bayesian Framework for Multiple Facets

Ester Alongi
Joint work with Gianmarco Altoè & Giovanni Parmigiani

April 10, 2026

Psicostat

Introduction

ROYAL SOCIETY OPEN SCIENCE

royalsocietypublishing.org/journal/rsos

Review

Cite this article: Heyard R *et al*. 2025 A scoping review on metrics to quantify reproducibility: a multitude of questions leads to a multitude of metrics. *R. Soc. Open Sci.* **12**: 242076.
https://doi.org/10.1098/rsos.242076

A scoping review on metrics to quantify reproducibility: a multitude of questions leads to a multitude of metrics

Rachel Heyard[1], Samuel Pawel[2], Joris Frese[3], Bernhard Voelkl[4], Hanno Würbel[4], Sarah McCann[5], Leonhard Held[1,2], Kimberley E. Wever[6], Helena Hartmann[7], Louise Townsin[8] and Stephanie Zellers[9]

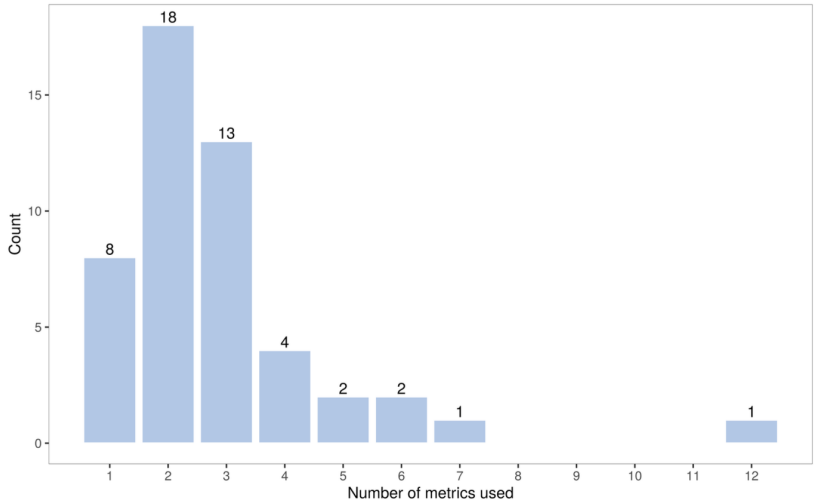


Figure 2. The total number of metrics used in the application papers to summarize reproducibility.

1 for a more detailed discussion of the terminology. The definition of reproducibility immediately asks for a specification of how to quantify the extent of agreement between a study and its replication. While there is no definition of reproducibility that is universally accepted across disciplines and research types, even less is known about the metric that best captures the reproducibility of a study or finding. However, selecting the most appropriate outcome for a reproducibility study[1] is crucial to ensure the accuracy and credibility of research into the reproducibility of science.

An increasing number of articles has discussed the relevance of various metrics to define 'successful replication' in the pairwise comparison of original-replication study pairs. Hereafter, we define a successful replication as 'a replication study for which the results agree with the corresponding original study'. 'Agreement of results' can mean different things: from an exact match of numeric values to matching conclusions. In a rapid review of replication studies in psychology published in 2013, Anderson & Maxwell [13] investigated the decision criteria for successful replication. They concluded that the majority of published replication studies (44 of the 50 included studies) classified the replication as successful when both studies came to the same conclusion based on statistical significance. Cobey *et al*. [14] conducted a scoping review of replication studies published in 2018 and 2019 in economics, education, psychology, health sciences and biomedicine to describe the epidemiological characteristics of this literature. They found large variability in how authors assessed reproducibility, although most of the included studies used a comparison of effect sizes to define success. Furthermore, large-scale reproducibility efforts, e.g. the replication projects in psychology [15], experimental economics [16] or cancer biology [17], all used a whole set of metrics based on statistical significance, effect sizes or methodology from meta-analysis to summarize the reproducibility of a research field. This list of traditional metrics for reproducibility includes the significance criterion, where a replication is considered successful if it finds a statistically significant effect in the same direction as the original study, and effect size comparisons, where success is determined by the similarity between the effect sizes of the replication and the original study. To in-

A noteworthy observation from our data extraction is that large-scale replication projects rarely provide a definition of reproducibility. Additionally, while these studies put a lot of effort into describing the design and methods used in the replication, they seldom outline the methods used to summarize reproducibility. Instead, they tend to only report the results in a descriptive manner in the results section. Therefore, we invite researchers to choose the metric(s) that align(s) with their research question and justify this choice. Sharing data and code could further allow for the assessment of the performance of other metrics or how they interact and complement each other in practice.

Motivating example

Scenario 1

- Two large studies with β_1 large, β_2 small.
- $P(\beta > \varepsilon|\text{data}) \approx 0.72$, with $\varepsilon > 0$ that is a pre-specified threshold.

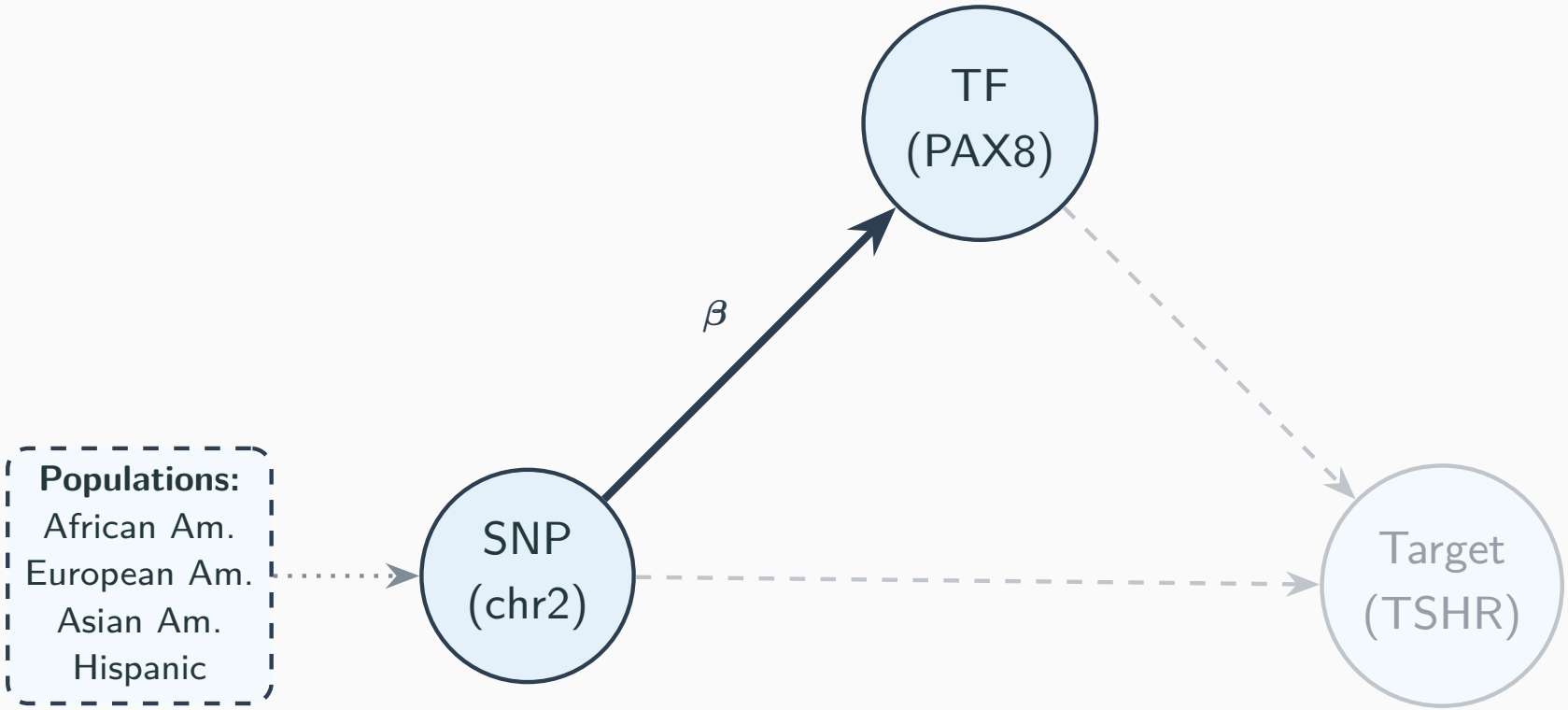
Scenario 2

- Two smaller studies with both β_3 and β_4 moderate and fairly consistent.
- $P(\beta > \varepsilon|\text{data}) \approx 0.72$.

Question

Do these scenarios with the same probability represent the **same state of knowledge?**

An introductive example



The biological replicability question

PAX8 is the master transcriptional regulator of the thyroid, driving the expression of the Thyroid Stimulating Hormone Receptor. Does the genetic variant regulating PAX8 have a **replicable effect** across different human populations?

Bayesian hierarchical model

The model

For study $s = 1, \ldots, S$ and individual $i = 1, \ldots, n_s$

$$Y_{si} \sim \mathcal{N}\left(\alpha_s + \beta_s X_{si},\ \sigma_s^2\right)$$

The study-specific parameters are drawn from **population-level** distributions

$$\begin{aligned} \beta_s \mid \beta, \tau_\beta &\sim \text{Student-}t(3, \beta, \tau_\beta) \\ \alpha_s \mid \alpha, \tau_\alpha &\sim \text{Student-}t(3, \alpha, \tau_\alpha) \\ \sigma_s &\sim \text{Student-t}_{[0,\infty)}(3, \mu_\sigma, \eta_\sigma) \end{aligned}$$

The generative parameters are assigned **data-driven informative priors** elicited from an Empirical Bayes approach

$$\begin{aligned} \beta &\sim \text{Student-}t(3, \mu_\beta, \sigma_\beta) & \alpha &\sim \text{Student-}t(3, \mu_\alpha, \sigma_\alpha) \\ \tau_\beta &\sim \text{Half-}t(3, 0, \sigma_{\tau_\beta}) & \tau_\alpha &\sim \text{Half-}t(3, 0, \sigma_{\tau_\alpha}). \end{aligned}$$

Prior and hyperprior specification

Empirical Bayes Approach

The hyperparameters are informed by meta-analytical estimates derived from a set of almost 200 independent biological triplets.

Triplet k	**SNP**	**TF**	**Gene**	**Pop 1 (African)**		**Pop 2 (European)**		**...**
				$\hat{\beta}_{1k}$	SE_{1k}	$\hat{\beta}_{2k}$	SE_{2k}	
$k = 1$	rs11249415	*TP53*	*CDKN1A*	0.42	0.05	0.38	0.04	...
$k = 2$	rs9282541	*MYC*	*CD44*	-0.15	0.08	-0.18	0.06	...
...	...	...	...	...	...	...	...	...

Empirical Bayes: hyperprior distributions

Path Coefficients and Intercepts

For each triplet k,

$$\mu_{\beta,k} = \frac{\sum_s w_{sk}\,\hat{\beta}_{sk}}{\sum_s w_{sk}} \qquad \tau_{\beta,k} = \sqrt{\frac{\sum_s w_{sk}(\hat{\beta}_{sk} - \mu_{\beta,k})^2}{\sum_s w_{sk}}},$$

where $w_{sk} = 1/SE_{sk}^2$.

Residual Standard Deviations

$$\mu_{\sigma,k} = \frac{\sum_s w_{sk}\,\hat{\sigma}_{sk}}{\sum_s w_{sk}},$$

where $w_{sk} = df_{sk}$.

Global Empirical Bayes Hyperparameters

To match the empirical variance to a Student-t distribution ($\nu = 3$), scale parameters are adjusted by $c = \sqrt{(\nu-2)/\nu}$

$$\mu_\beta = \text{Mean}(\mu_{\beta,k}) \qquad \sigma_\beta = c \cdot \text{SD}(\mu_{\beta,k}) \qquad \sigma_{\tau_\beta} = c \cdot \text{Mean}(\tau_{\beta,k})$$

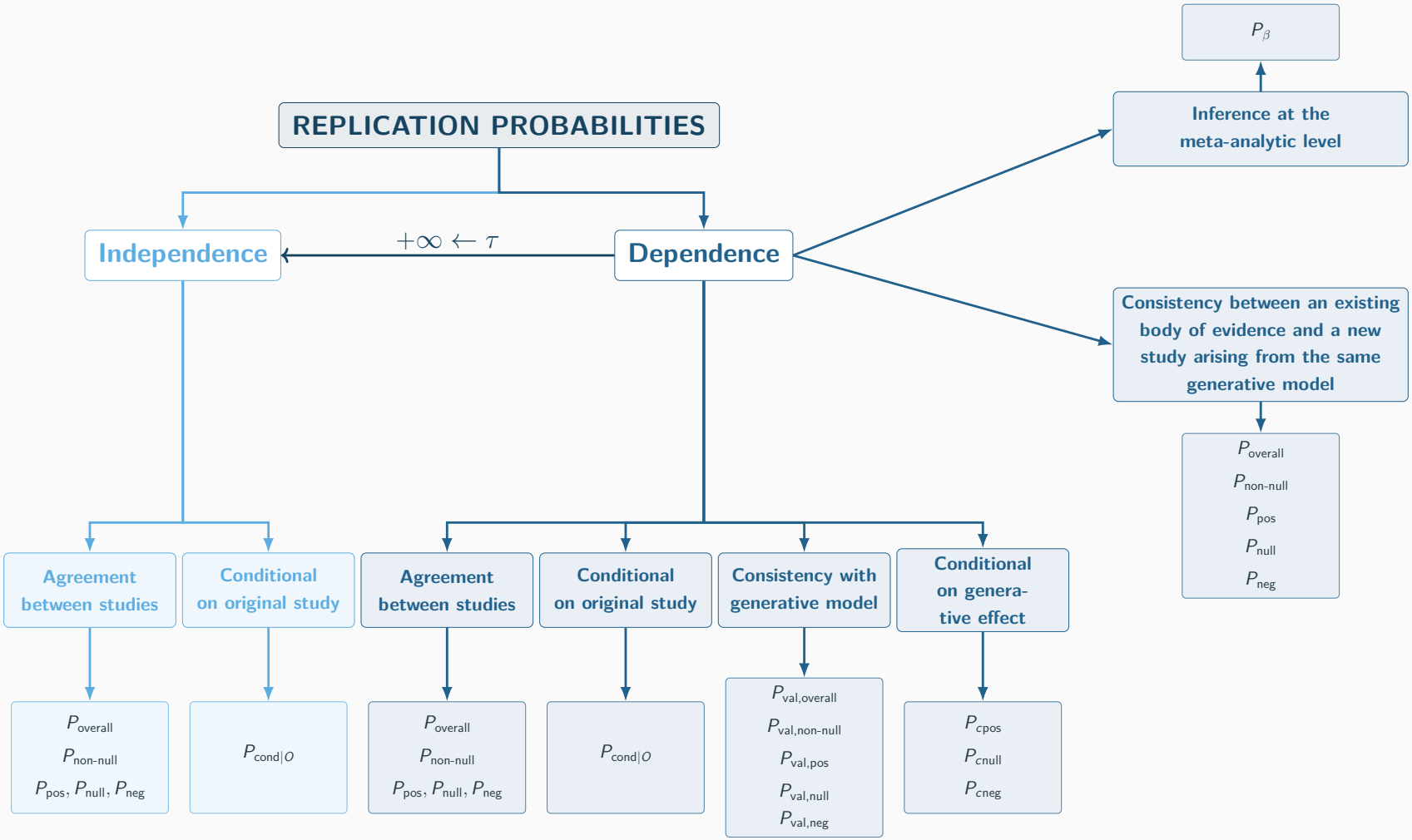
$$\mu_\sigma = \text{Mean}(\mu_{\sigma,k}) \qquad \eta_\sigma = c \cdot \text{SD}(\mu_{\sigma,k}).$$

Multiple facets of replicability

A single replicability analysis can address four complementary replicability questions

- **Replicability as traditionally defined in independent studies**: how consistent are the results of studies on the effect?
  -> *Independent analysis*
- **Inference at the meta-analytic level implied by a common generative model**: does the generative effect exceed a threshold?
  -> *Generative model*
- **Consistency between individual studies and the shared meta-analytic structure**: what is the probability that two studies that are consistent are also consistent with the generative model?
  -> *Hierarchical model*
- **Consistency between an existing body of evidence and a new study arising from the same generative model**: what is the probability that a new study, is consistent with the existing body of evidence?
  -> *Hierarchical model*

Taxonomy of replication metrics



Posterior draws & region of interest

- Let $t = 1, \ldots, T$ index the posterior samples generated by MCMC. For each study s, we obtain posterior draws

$$\beta_s^{(t)}, \qquad t = 1, \ldots, T.$$

- The **effect parameter subspace** (region of interest) $\mathcal{B} \subset \mathbb{R}$ represents values that are considered practically or theoretically meaningful

$$\mathcal{B} = \{\beta : |\beta| > \varepsilon\},$$

where $\varepsilon > 0$ is based on **theoretical knowledge of the phenomenon** of interest.

- For a given study s, the **posterior probability** that the effect lies in the region of interest is estimated as

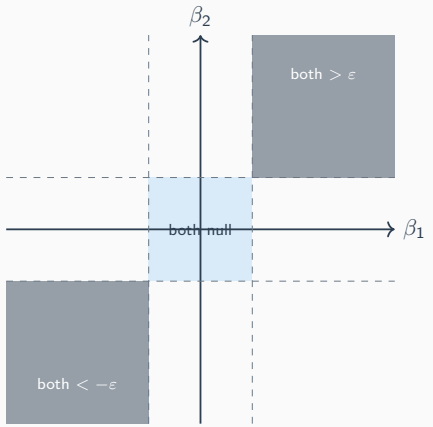
$$P\big(\beta_s \in \mathcal{B} \mid \text{Data}_s\big) \approx \frac{1}{T} \sum_{t=1}^{T} \mathbf{1}\Big(\beta_s^{(t)} \in \mathcal{B}\Big).$$

Replication probabilities for independent analyses

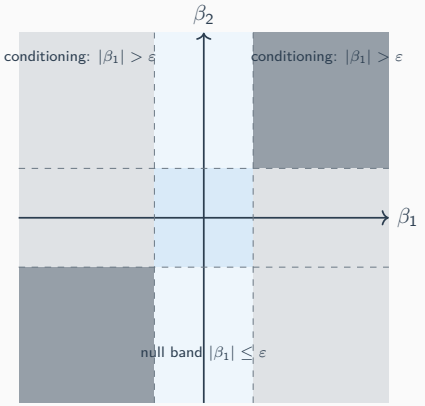
Independence

	Agreement between studies					Conditional probability on the original study
	Overall	Non-null	Signed-agree			Conditional
			pos	null	neg	
S	$P_{\text{overall},S}$	$P_{\text{non-null},S}$	$P_{\text{pos},S}$	$P_{\text{null},S}$	$P_{\text{neg},S}$	$P_{\text{cond}\|O,S}$
S_{-1}	$P_{\text{overall},S_{-1}}$	$P_{\text{non-null},S_{-1}}$	$P_{\text{pos},S_{-1}}$	$P_{\text{null},S_{-1}}$	$P_{\text{neg},S_{-1}}$	$P_{\text{cond}\|O,S_{-1}}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮
2	$P_{\text{overall},2}$	$P_{\text{non-null},2}$	$P_{\text{pos},2}$	$P_{\text{null},2}$	$P_{\text{neg},2}$	$P_{\text{cond}\|O,2}$

Agreement



Conditional probability



Overall agreement

For a threshold $\varepsilon > 0$, we define the count of studies showing positive, negative, or null effects

$$N_{pos}^{(t)} = \sum_{s=1}^{S} \mathbf{1}\big(\beta_s^{(t)} > \varepsilon\big), \quad N_{neg}^{(t)} = \sum_{s=1}^{S} \mathbf{1}\big(\beta_s^{(t)} < -\varepsilon\big), \quad N_{null}^{(t)} = \sum_{s=1}^{S} \mathbf{1}\big(|\beta_s^{(t)}| \leq \varepsilon\big).$$

Overall agreement (P_{overall})

The posterior probability that a consensus majority k (e.g., $k \geq 3$ out of $S = 4$) agree on the direction, or agree on the absence of a relevant effect

$$P_{\text{overall, k}} = \frac{1}{T} \sum_{t=1}^{T} \mathbf{1}\Big(N_{null}^{(t)} \geq k \ \vee \ N_{pos}^{(t)} \geq k \ \vee \ N_{neg}^{(t)} \geq k\Big)$$

Interpretation: How often do the populations qualitatively agree without contradicting each other?

Conditional replication probability

Conditional replication probability ($P_{\text{cond}|O}$)

The conditional posterior probability that the effect discovered in the original study O replicates in at least $k-1$ other studies (i.e., at least k studies in total):

$$P_{\text{cond}|O} = \frac{\sum_{t=1}^{T} \mathbf{1}\Big(\big(\beta_O^{(t)} > \varepsilon \wedge N_{pos}^{(t)} \geq k\big) \ \vee \ \big(\beta_O^{(t)} < -\varepsilon \wedge N_{neg}^{(t)} \geq k\big) \Big)}{\sum_{t=1}^{T} \mathbf{1}\big(|\beta_O^{(t)}| > \varepsilon\big)}$$

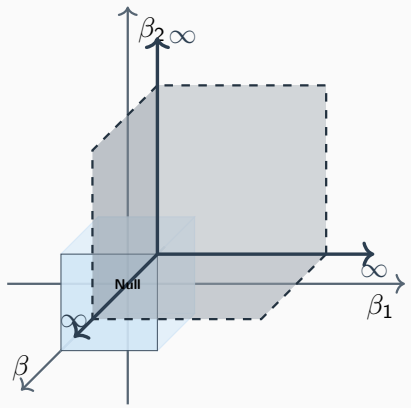
Interpretation: Assuming the effect is practically significant in the discovery cohort, what is the probability it exists in at least $k-1$ other populations?

Replication probabilities for dependent analyses

Dependence

	Agreement between studies					Conditional probability on the original study	Consistency between studies and the generative model					Conditional probability on the generative effect			
	Overall	Non-null	Signed-agree			Conditional	Overall	Non-null	Signed-agree			Conditional signed-agree			
			pos	null	neg				pos	null	neg	pos	null	neg	
S	$P_{\text{overall},S}$	$P_{\text{non-null},S}$	$P_{\text{pos},S}$	$P_{\text{null},S}$	$P_{\text{neg},S}$	$P_{\text{cond}\mid O,S}$	$P_{\text{val,overall},S}$	$P_{\text{val,non-null},S}$	$P_{\text{val,pos},S}$	$P_{\text{val,null},S}$	$P_{\text{val,neg},S}$	$P_{\text{cpos},S}$	$P_{\text{cnull},S}$	$P_{\text{cneg},S}$	
S_{-1}	$P_{\text{overall},S_{-1}}$	$P_{\text{non-null},S_{-1}}$	$P_{\text{pos},S_{-1}}$	$P_{\text{null},S_{-1}}$	$P_{\text{neg},S_{-1}}$	$P_{\text{cond}\mid O,S_{-1}}$	$P_{\text{val,overall},S_{-1}}$	$P_{\text{val,non-null},S_{-1}}$	$P_{\text{val,pos},S_{-1}}$	$P_{\text{val,null},S_{-1}}$	$P_{\text{val,neg},S_{-1}}$	$P_{\text{cpos},S_{-1}}$	$P_{\text{cnull},S_{-1}}$	$P_{\text{cneg},S_{-1}}$	
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	
2	$P_{\text{overall},2}$	$P_{\text{non-null},2}$	$P_{\text{pos},2}$	$P_{\text{null},2}$	$P_{\text{neg},2}$	$P_{\text{cond}\mid O,2}$	$P_{\text{val,overall},2}$	$P_{\text{val,non-null},2}$	$P_{\text{val,pos},2}$	$P_{\text{val,null},2}$	$P_{\text{val,neg},2}$	$P_{\text{cpos},2}$	$P_{\text{cnull},2}$	$P_{\text{cneg},2}$	

$P_{\text{val, pos, 2}}$



Consistency with the generative model

Overall consistency ($P_{\text{val, overall}}$)

How often do the majority of studies and the population-level effect simultaneously agree in sign (or null)?

$$P_{\text{val, overall, k}} = \frac{1}{T}\sum_{t=1}^{T}\mathbf{1}\begin{pmatrix}(N_{pos}^{(t)} \geq k \wedge \beta^{(t)} > \varepsilon) \vee \\ (N_{neg}^{(t)} \geq k \wedge \beta^{(t)} < -\varepsilon) \vee \\ (N_{null}^{(t)} \geq k \wedge |\beta^{(t)}| \leq \varepsilon)\end{pmatrix}$$

Signed consistency ($P_{\text{val, pos}}$, $P_{\text{val, neg}}$, $P_{\text{val, null}}$)

For a specific direction (e.g., positive)

$$P_{\text{val, pos, k}} = \frac{1}{T}\sum_{t=1}^{T}\mathbf{1}\left(N_{pos}^{(t)} \geq k \ \wedge \ \beta^{(t)} > \varepsilon\right)$$

Interpretation: Does the global generative effect β represent a robust biological consensus across most populations?

Conditional correctness of the generative model

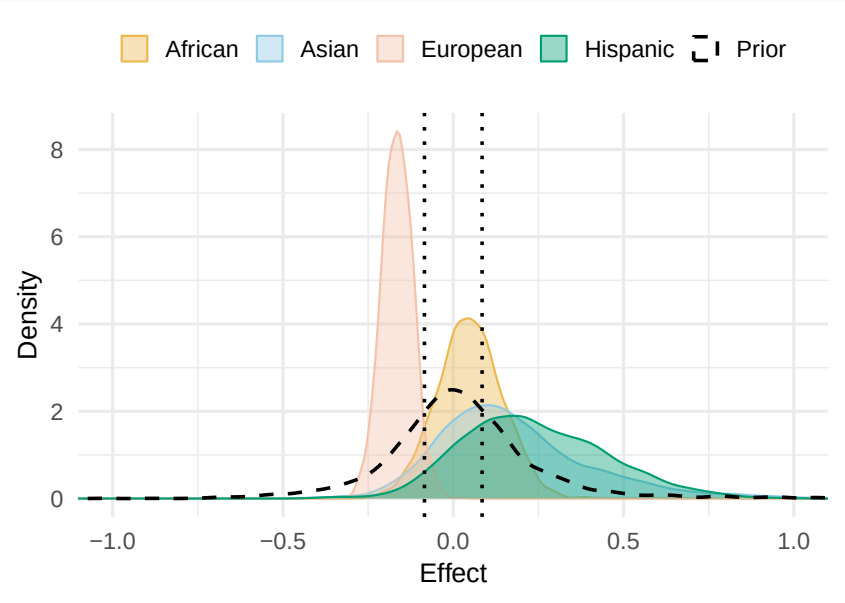
Conditional signed-agree ($P_{c\text{pos}}$, $P_{c\text{neg}}$, $P_{c\text{null}}$)

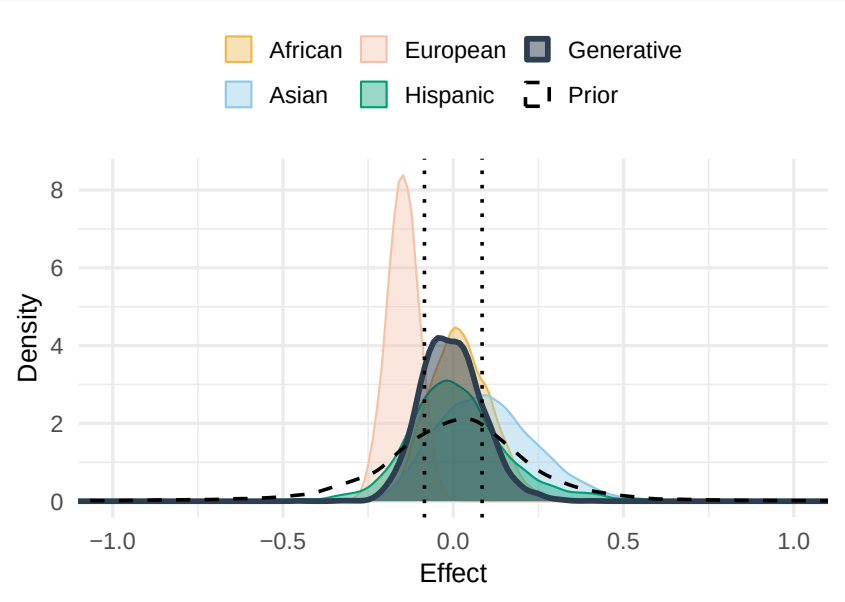
Conditional on the generative effect β being practically positive, the probability that this is robustly supported by a consensus of at least k studies is

$$P_{c\text{pos, k}} = \frac{\sum_{t=1}^{T} \mathbf{1}\left(N_{pos}^{(t)} \geq k \ \wedge \ \beta^{(t)} > \varepsilon\right)}{\sum_{t=1}^{T} \mathbf{1}\left(\beta^{(t)} > \varepsilon\right)}$$

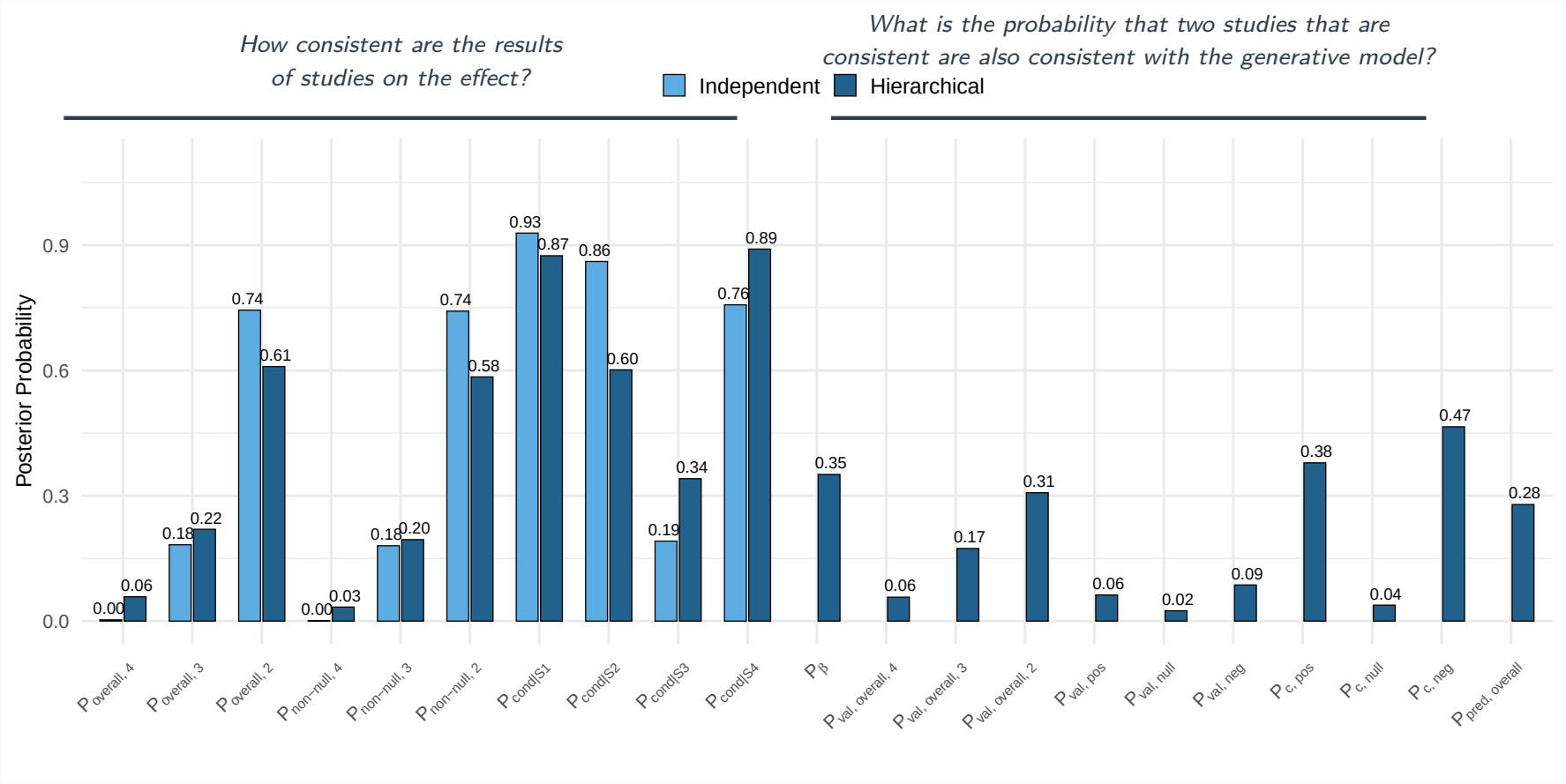
Interpretation: When the true global effect is positive, how often do the specific populations correctly identify and replicate a positive effect?

Results


African
Asian
European
Hispanic
Prior
Density
8
6
4
2
0
−1.0
−0.5
0.0
0.5
1.0
Effect


African
European
Generative
Asian
Hispanic
Prior
Density
8
6
4
2
0
−1.0
−0.5
0.0
0.5
1.0
Effect

Conclusion



Does the generative effect exceed a prespecified threshold?

What is the probability that a new study, arising from the same generative model, is consistent with the existing body of evidence?

References

- DuMouchel, W. (1994). *Hierarchical Bayes linear models for meta-analysis*. Technical Report 27, National Institute of Statistical Sciences.
- GTEx Consortium. (2020). The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science, 369(6509), 1318-1330.
- Heyard, R., Pawel, S., Frese, J., Voelkl, B., Würbel, H., McCann, S., ... Zellers, S. (2025). A scoping review on metrics to quantify reproducibility: a multitude of questions leads to a multitude of metrics. Royal Society Open Science, 12(7).
- Pawel, S., & Held, L. (2020). *Probabilistic forecasting of replication studies*. PLOS ONE, 15(4), e0231416.
- Verhagen, J., & Wagenmakers, E.-J. (2014). *Bayesian tests to quantify the result of a replication attempt*. Journal of Experimental Psychology: General, 143(4), 1457.