# A Coherent Framework for Many Replicability Questions

Ester Alongi
Joint work with Gianmarco Altoè & Giovanni Parmigiani

# Original vs. replication effect sizes (Nosek & Errington, Fig. 3)

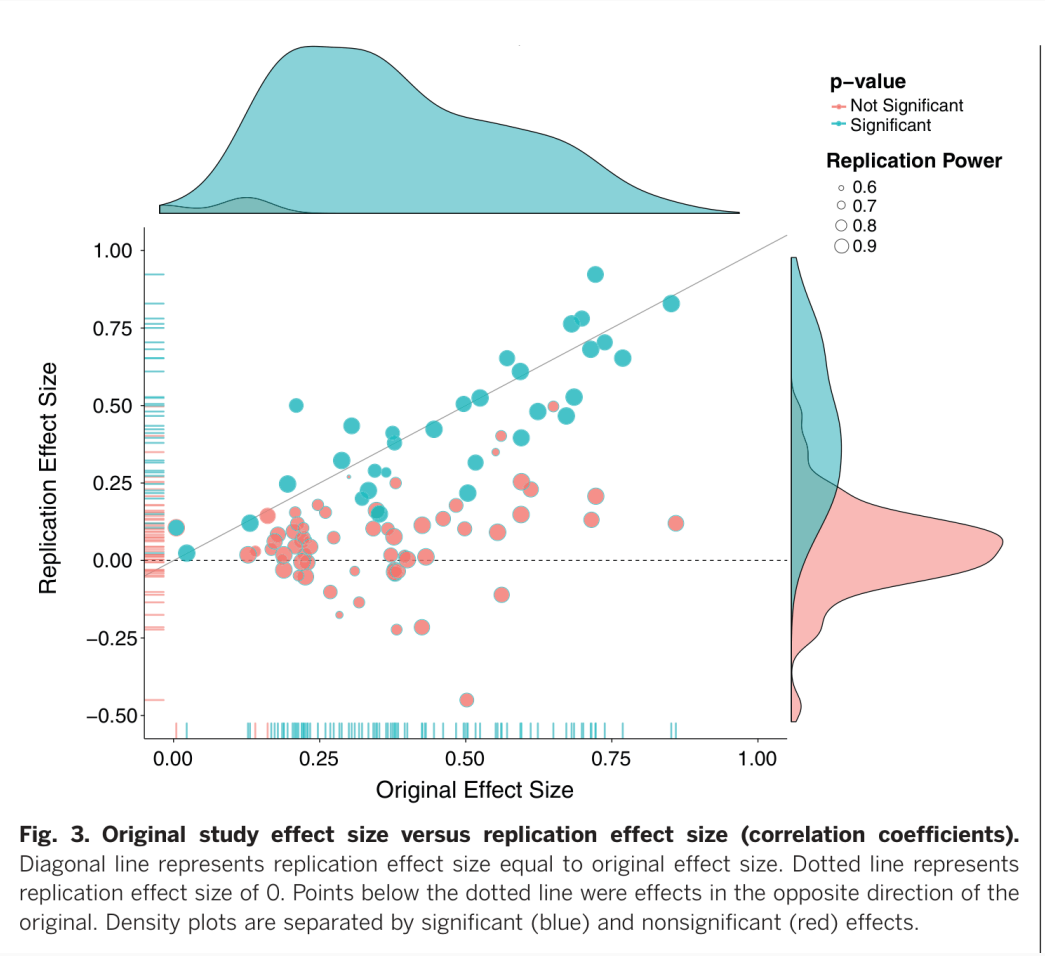



**Fig. 3. Original study effect size versus replication effect size (correlation coefficients).** Diagonal line represents replication effect size equal to original effect size. Dotted line represents replication effect size of 0. Points below the dotted line were effects in the opposite direction of the original. Density plots are separated by significant (blue) and nonsignificant (red) effects.

# Five success criteria for replication (Nosek & Errington, Fig. 3)

| # | Indicator | Question it answers |
|---|---|---|
| 1 | Replication $p < .05$, same direction | Did the replication itself reject the null? |
| 2 | Original effect inside replication 95% CI | Is the original estimate compatible with the replication? |
| 3 | Replication effect size vs. original | How much smaller (or larger) is the new estimate? |
| 4 | Meta-analytic combination of the two | What is the pooled evidence? |
| 5 | Subjective "did it replicate?" rating | Holistic judgement by the replicating team |

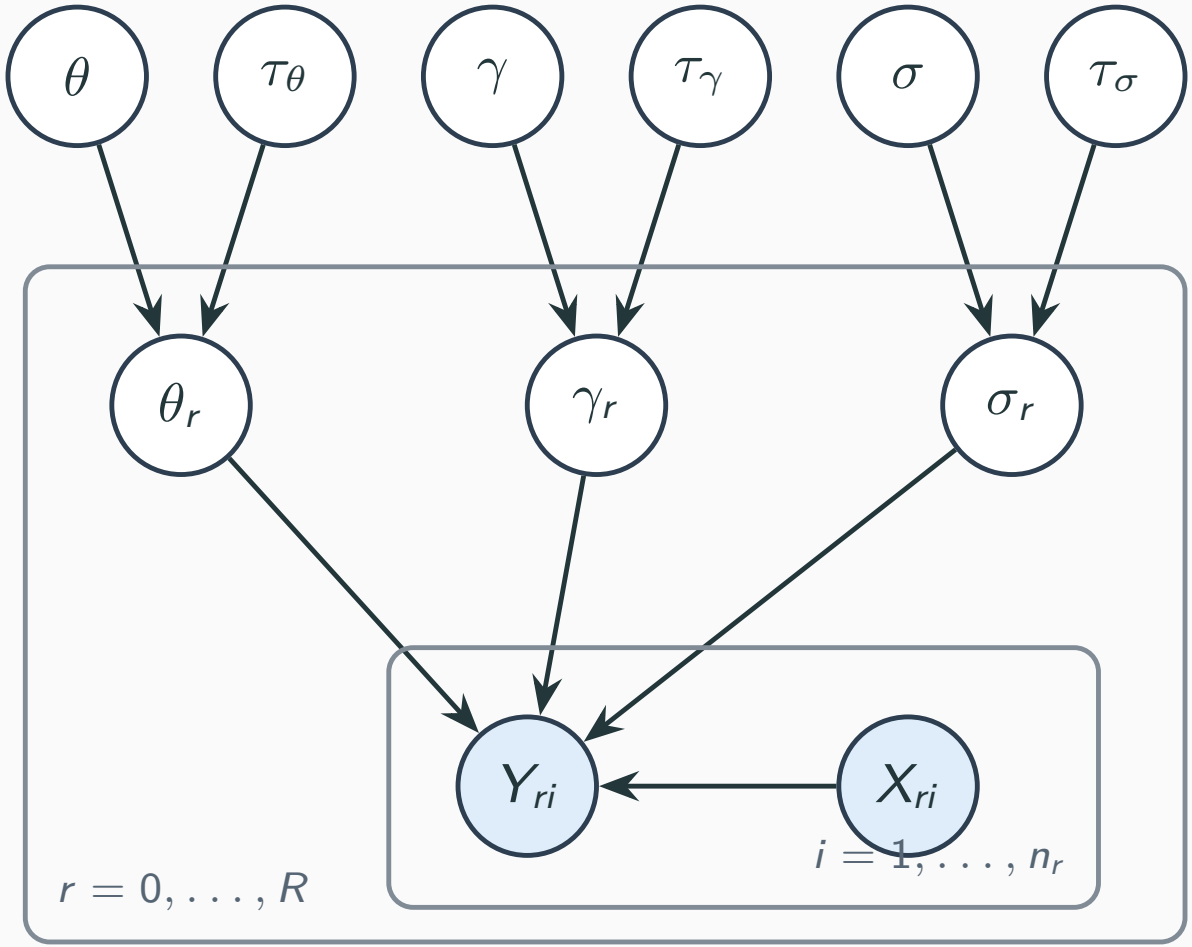
# Some Challenges with Existing Methodologies

- **Fragmentation:** Over 50 metrics requiring different models currently exist (Hayes, 2013);
- **Dichotomization:** Replication assessment is collapsed into binary "success/failure" thresholds (Pawel et al., 2024);
- **Empirical Focus:** Studies are compared through horizontal empirical comparisons (Machery, 2020).

# Principles

- Predictions about replicability are important prospectively, as we plan new studies; reflections on whether replication has occurred are best carried out conditionally on data, based on statements on effects or other parameters
- Whether replication experiments are extensions or not is also an empirical question.
- The same model can provide the backbone for a answering, in an integrated way, many questions of replicability.

# Bayesian Hierarchical Model

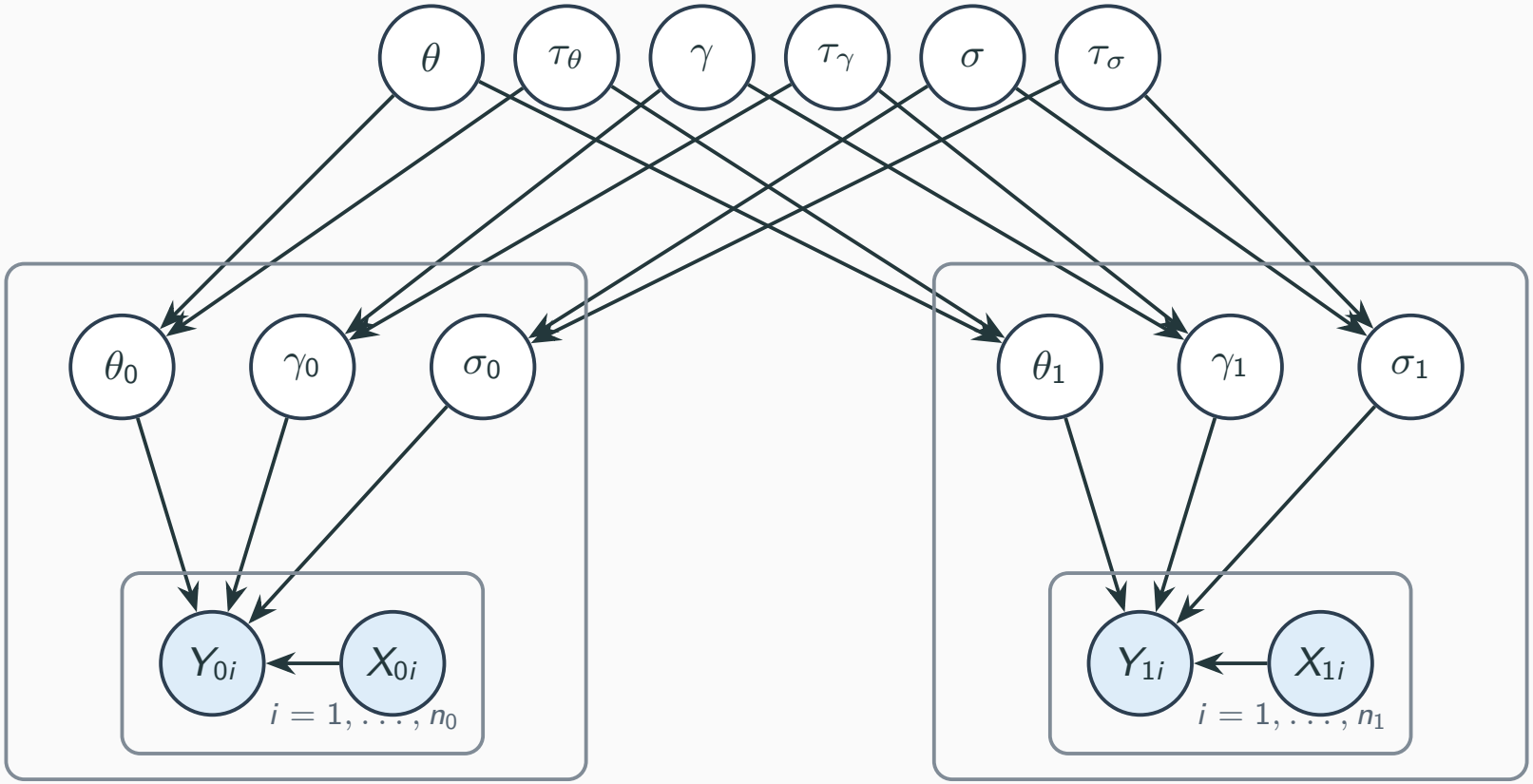



**Generative model (generic).** $Y_{ri} \mid \theta_r, \gamma_r, \sigma_r, X_{ri} \sim p_Y\Big(\cdot \mid \gamma_r + \theta_r X_{ri},\, \sigma_r^2\Big)$, with population-level draws

$$\theta_r \mid \theta, \tau_\theta \sim p_\theta, \quad \gamma_r \mid \gamma, \tau_\gamma \sim p_\gamma, \quad \sigma_r \mid \sigma, \tau_\sigma \sim p_\sigma.$$

Hyperpriors on $\theta, \tau_\theta, \gamma, \tau_\gamma, \sigma, \tau_\sigma$ are set by empirical Bayes (next slide).

# Hierarchical model with one replication ($R = 1$)

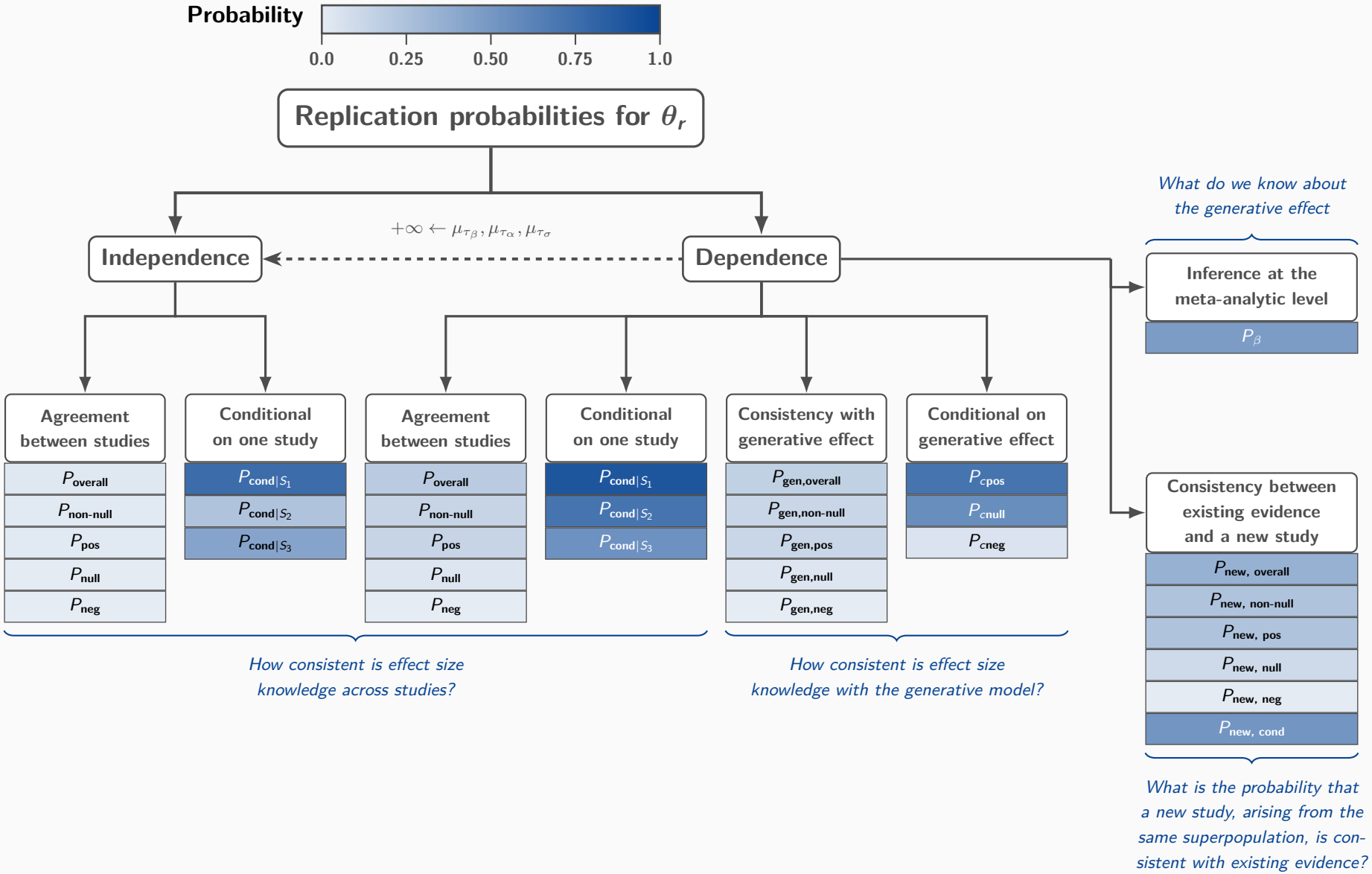



Each study draws its own $(\theta_r, \gamma_r, \sigma_r)$ from the shared population-level distributions; precise replication corresponds to $\tau_\theta, \tau_\gamma, \tau_\sigma \to 0$, so $\theta_0 = \theta_1$ a.s.

# Facets

- Dependent vs Independent Analyses of Studies
- Rebplicability of Study Conclusions versus Agreement of Studies with generative mode (inference on which capture the field's knowlege about the effect).
- Prospective planning of new studies for replication or extension

# Multiple Facets of Replicability



# Region of practical equivalence

**Definition**

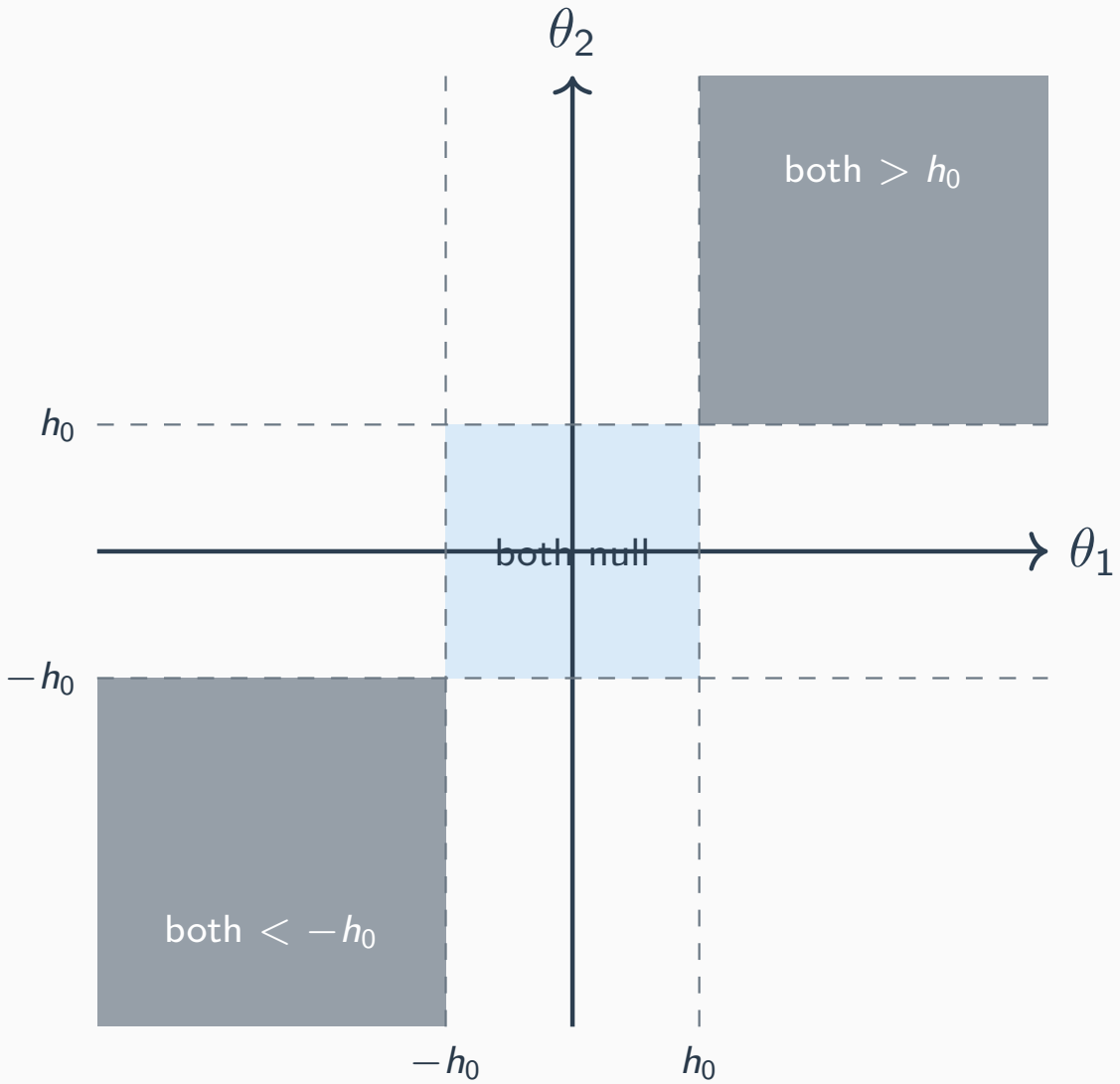
Given a threshold $h_0 > 0$, the **region of practical equivalence (ROPE)** is the interval $(-h_0, h_0)$. The null hypothesis

$$H_0 : \ |\theta| < h_0$$

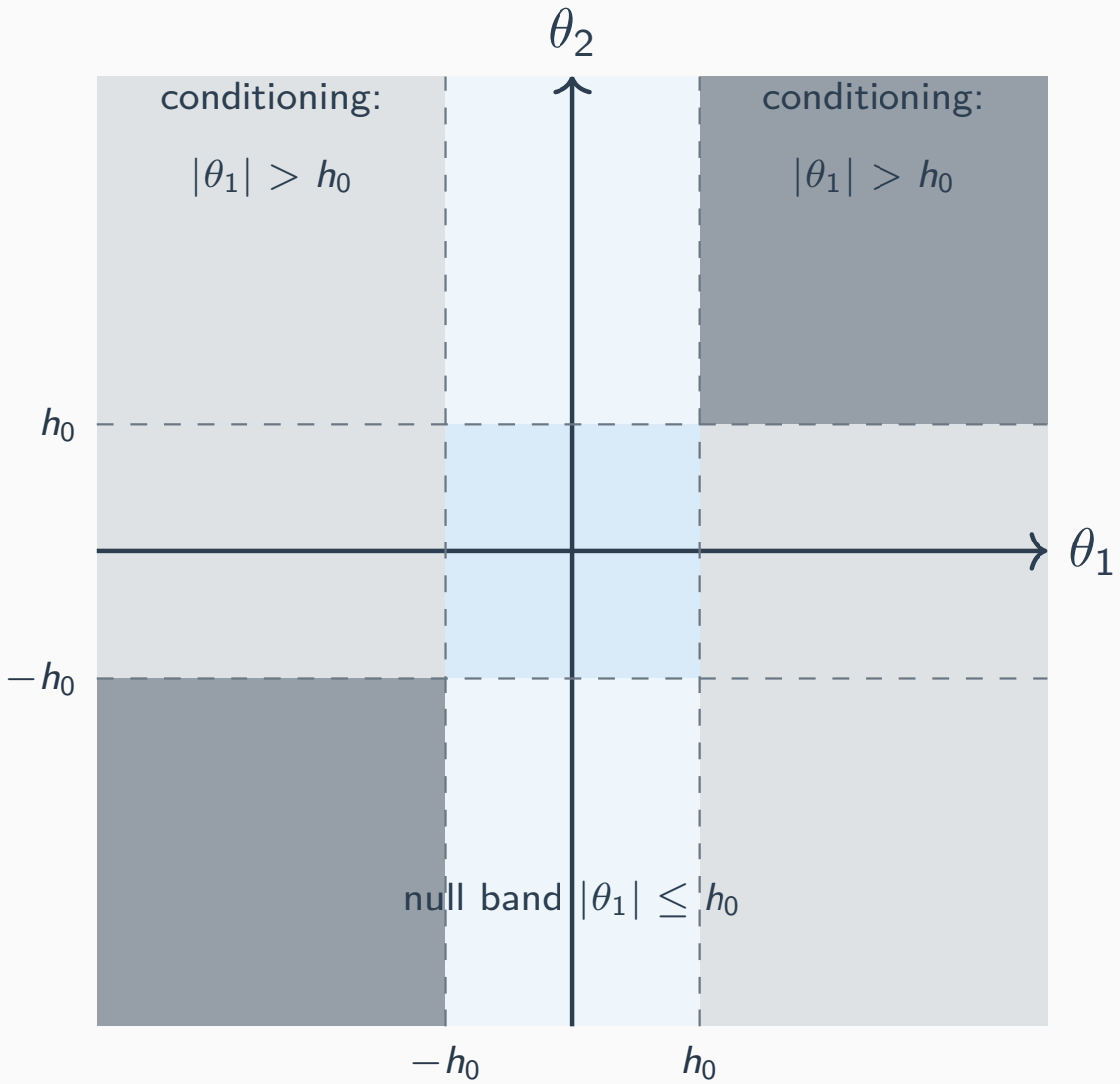
asserts that the effect is too small to matter scientifically; the alternative $H_1 : \ |\theta| \geq h_0$ asserts that it lies outside the ROPE.

- $h_0$ is fixed a priori from substantive knowledge (clinical, biological, economic).
- Bayesian view: posterior mass inside $(-h_0, h_0)$ summarises support for the null.
- Frequentist counterpart: equivalence testing against the bounds $\pm h_0$.

# Overall Agreement

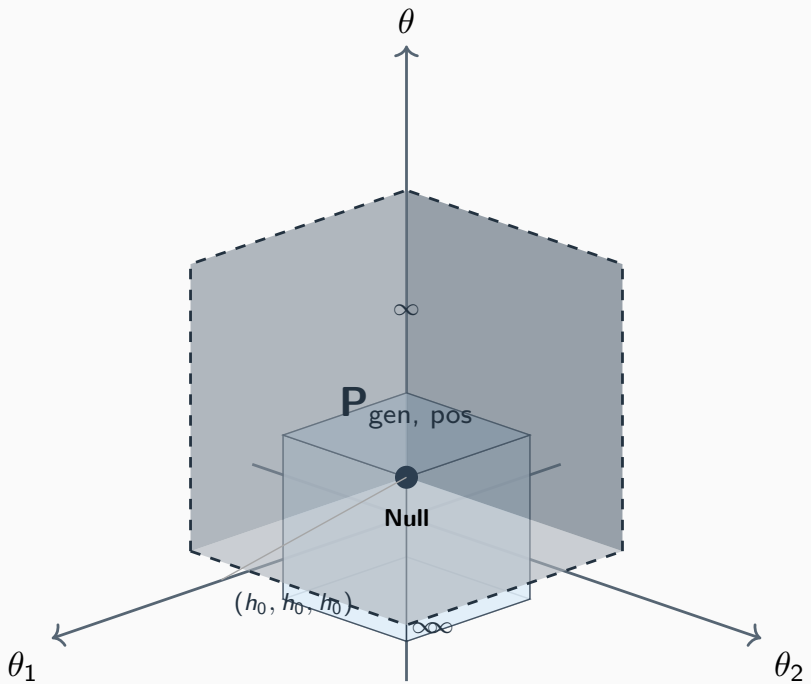


# Conditional probabilities



# Generative Consistency

$$(P_{\text{gen, pos}})$$



# References



Consortium, G. (2020).**The gtex consortium atlas of genetic regulatory effects across human tissues.** *Science*, *369*(6509), 1318–1330.


Hayes, A. F. (2013). ***Introduction to mediation, moderation, and conditional process analysis: A regression-based approach*.** Guilford Press.


Machery, E. (2020).**What is a replication?** *Philosophy of Science*, *87*(4), 545–567.


National Academies of Sciences, E., & Medicine. (2019). ***Reproducibility and replicability in science*.** National Academies Press.

# References ii

Pawel, S., Heyard, R., Micheloud, C., & Held, L. (2024). **Replication of null results: Absence of evidence or evidence of absence?** *Elife*, *12*, RP92311.

# Backup slides

# Metrics explanation

Let $t = 1, \ldots, T$ index the posterior samples generated by MCMC. For each study $s$, we obtain posterior draws $\theta_r^{(t)},\ t = 1, \ldots, T$, where $h_0 > 0$ is a prespecified theoretical threshold.

The counts of studies showing positive, negative, or null effects are

$$N_{pos}^{(t)} = \sum_{r=0}^{R} \mathbf{1}\big(\theta_r^{(t)} > h_0\big), \quad N_{neg}^{(t)} = \sum_{r=0}^{R} \mathbf{1}\big(\theta_r^{(t)} < -h_0\big), \quad N_{null}^{(t)} = \sum_{r=0}^{R} \mathbf{1}\big(|\theta_r^{(t)}| \leq h_0$$

# Agreement between studies

## Overall agreement ($P_{\text{overall}}$)

The posterior probability that a consensus majority $k$ (e.g., $k \geq 2$ out of $S = 3$) agree on the direction, or agree on the absence of a relevant effect:

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

*Interpretation:* How often do the populations qualitatively agree?

## Signed agreement ($P_{\text{pos}}$, $P_{\text{neg}}$, $P_{\text{null}}$)

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

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

# Conditional Replication Probability on one study

**Conditional replication probability on one study ($P_{\text{cond}|O}$)**

The conditional posterior probability that the effect discovered in the original study $O$ replicates in at least $k-1$ other studies:

$$P_{\text{cond}|O} = \frac{\sum_{t=1}^{T} \mathbf{1}\Big(\big(\theta_O^{(t)} > h_0 \wedge N_{pos}^{(t)} \geq k\big) \ \vee \ \big(\theta_O^{(t)} < -h_0 \wedge N_{neg}^{(t)} \geq k\big)\Big)}{\sum_{t=1}^{T} \mathbf{1}\big(|\theta_O^{(t)}| > h_0\big)}$$

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

# Consistency with the Generative Effect

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

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

$$P_{\text{gen, overall, k}} = \frac{1}{T}\sum_{t=1}^{T} \mathbf{1}\begin{pmatrix} (N_{pos}^{(t)} \geq k \wedge \theta^{(t)} > h_0) \vee \\ (N_{neg}^{(t)} \geq k \wedge \theta^{(t)} < -h_0) \vee \\ (N_{null}^{(t)} \geq k \wedge |\theta^{(t)}| \leq h_0) \end{pmatrix}$$

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

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

$$P_{\text{gen, pos, k}} = \frac{1}{T}\sum_{t=1}^{T} \mathbf{1}\left(N_{pos}^{(t)} \geq k \ \wedge \ \theta^{(t)} > h_0\right)$$

*Interpretation:* Does the global generative effect $\theta$ represent a robust biological consensus across most populations?

# Conditional Replication Probability on the generative effect

**Conditional replication probability on the generative effect ($P_{c\text{pos}}$, $P_{c\text{neg}}$, $P_{c\text{null}}$)**

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

$$P_{c\text{pos, k}} = \frac{\sum_{t=1}^{T} \mathbf{1}\left(N_{pos}^{(t)} \geq k \ \wedge \ \theta^{(t)} > h_0\right)}{\sum_{t=1}^{T} \mathbf{1}\left(\theta^{(t)} > h_0\right)}$$

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

# Simulation Strategy: Isolating Sources of Heterogeneity

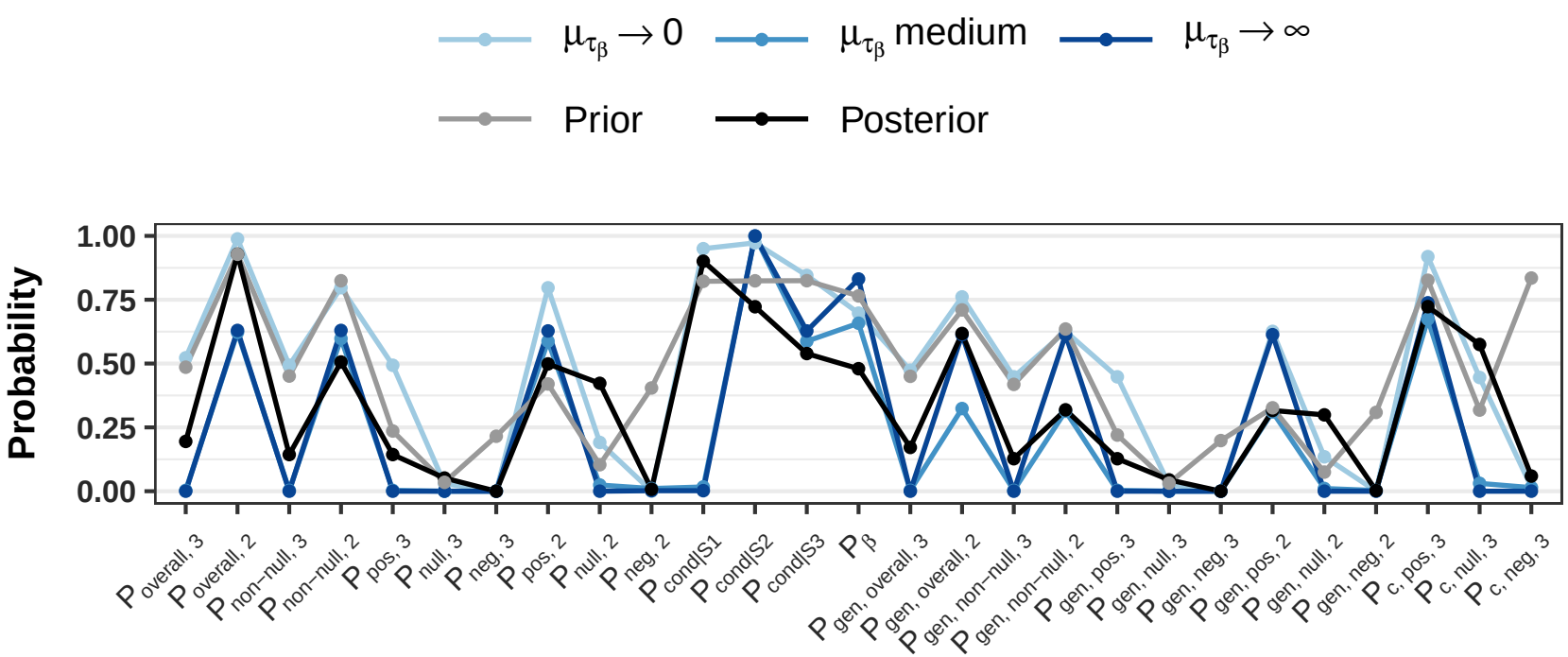
| Simulation Scenario | Perturbed Heterogeneity | Fixed Parameters (Real Posterior Means) | | |
|---|---|---|---|---|
| | | **Gen. Effect ($\theta$)** | **Baseline Exp. ($\gamma$)** | **Noise ($\sigma$)** |
| **1. Effect** | Effect Heterogeneity ($\mu_{\tau_\theta}$) | – | $\hat{\gamma}_{r,real}$ | $\hat{\sigma}_{r,real}$ |
| **2. Baseline** | Intercept Heterogeneity ($\mu_{\tau_\gamma}$) | $\hat{\theta}_{r,real}$ | – | $\hat{\sigma}_{r,real}$ |
| **3. Residual Variance** | Noise Heterogeneity ($\mu_{\tau_\sigma}$) | $\hat{\theta}_{r,real}$ | $\hat{\gamma}_{r,real}$ | – |

*Note: For the perturbed parameter, the 3 simulated studies are forced at the 10%, 50%, and 90% quantiles of the $\tau$ distribution.*

# Scenario 1: Genetic Effect Heterogeneity

## Research Question

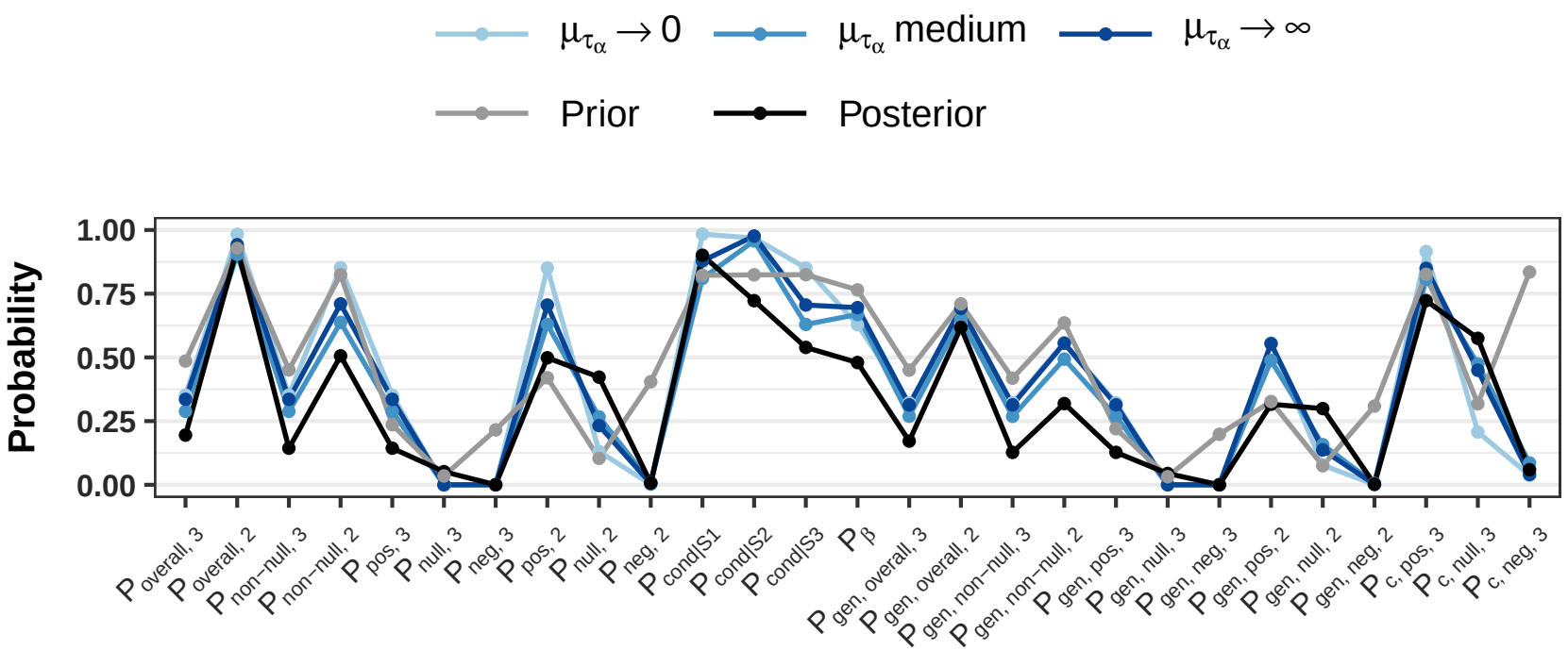
How would replication metrics change if the genetic effect was perfectly consistent across cohorts ($\mu_{\tau_\theta} \to 0$), assuming baselines ($\gamma$) and noise ($\sigma$) remain at their real-world values?



# Scenario 2: Baseline Expression Shifts

## Research Question

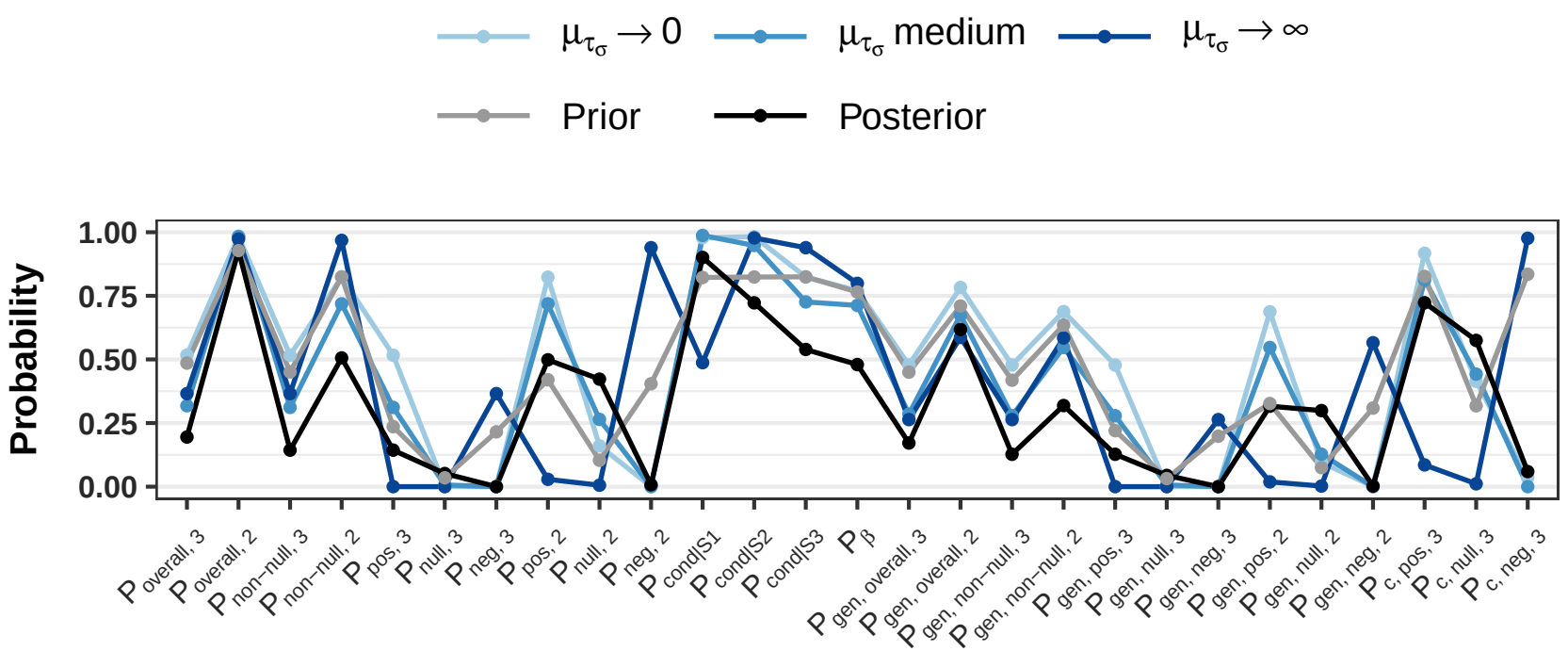
How sensitive is the replication of a genetic effect to systematic differences in baseline gene expression ($\gamma$) across populations, assuming constant $\theta$ and $\sigma$?



# Scenario 3: Unmeasured Noise Heterogeneity

## Research Question

If cohorts have vastly different levels of unmeasured environmental/technical noise ($\sigma$), what impact does this have on the replicability of the effect of interest?



# Empirical Bayes: Hyperpriors distributions

To match empirical variances to a Student-t distribution ($\nu = 3$), we define the scale conversion factor $c = \sqrt{(\nu - 2)/\nu}$.

## Path Coefficients and Intercepts

For each triplet $k$, using weights $w_{rk} = 1/SE_{rk}^2$:

$$\theta_k = \frac{\sum_r w_{rk}\,\hat{\theta}_{rk}}{\sum_r w_{rk}} \qquad \tau_{\theta,k} = c \cdot \sqrt{\frac{\sum_r w_{rk}(\hat{\theta}_{rk} - \theta_k)^2}{\sum_r w_{rk}}}$$

## Residual Standard Deviations

For each triplet $k$, using weights $w_{rk} = df_{rk}$:

$$\sigma_k = \frac{\sum_r w_{rk}\,\hat{\sigma}_{rk}}{\sum_r w_{rk}} \qquad \tau_{\sigma,k} = c \cdot \sqrt{\frac{\sum_r w_{rk}(\hat{\sigma}_{rk} - \sigma_k)^2}{\sum_r w_{rk}}}$$

## Global Empirical Bayes Hyperparameters

$$\mu_\theta = \text{Mean}(\theta_k) \qquad \eta_\theta = c \cdot \text{SD}(\theta_k)$$

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

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

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

# Multiple Facets of Replicability