

How do questionable research practices affect inferences of heterogeneity? A computer simulation.

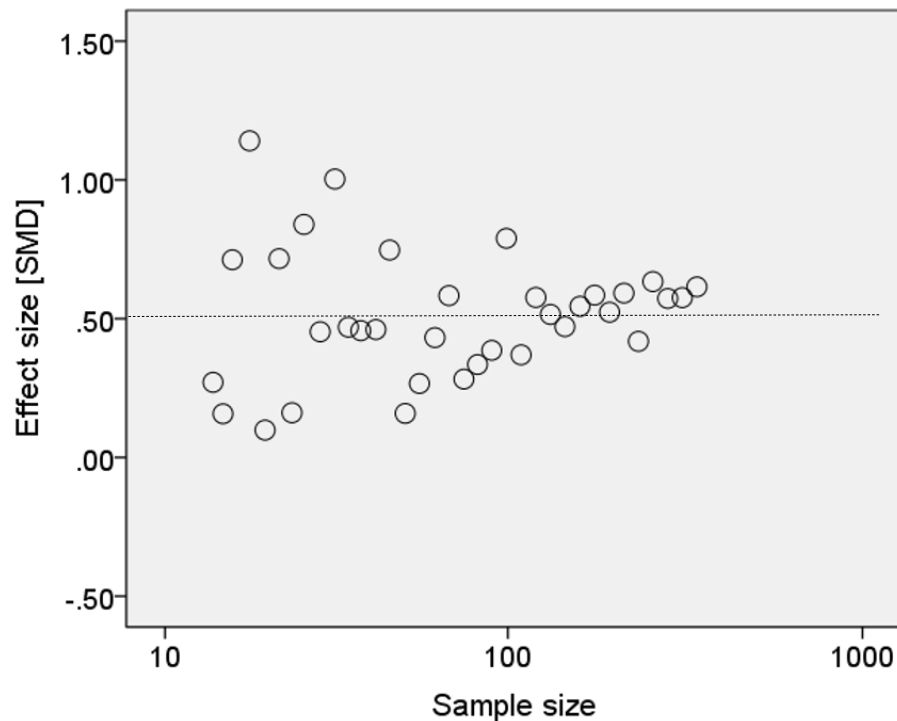
Johannes Hönekopp
(johannes.honekopp@unn.ac.uk)

Audrey Linden

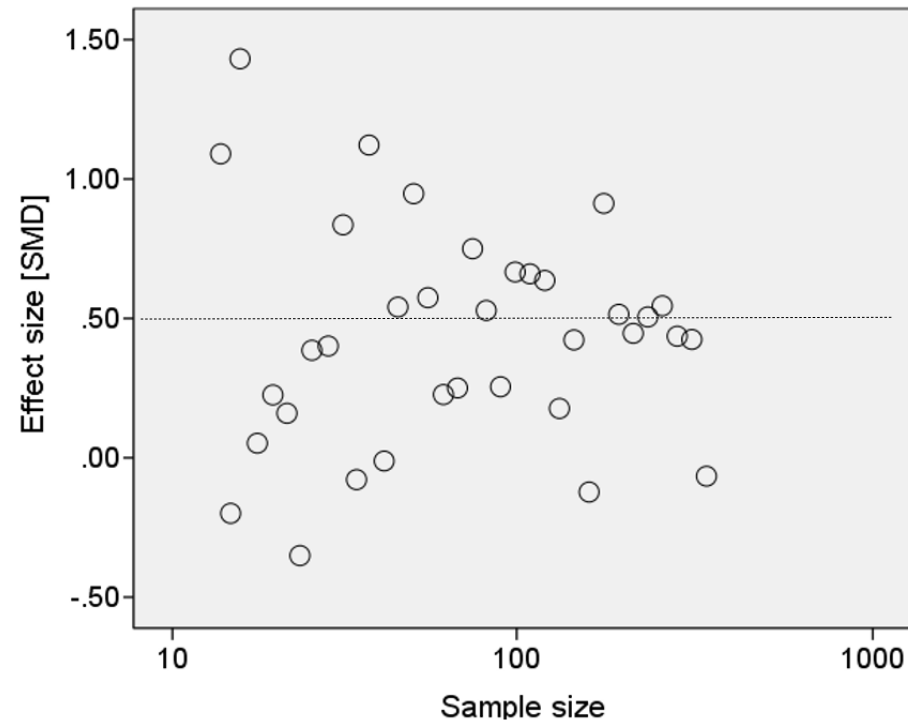
Open Science, Trier, April 2019

Background

Estimated **heterogeneity** central to result of meta-analysis.



Computer simulation: 35 between subjects-experiments tap into the **same population effect** (SMD = 0.5).



35 experiments tap into **variable population effect** ($M = 0.5$, $SD = 0.3$).

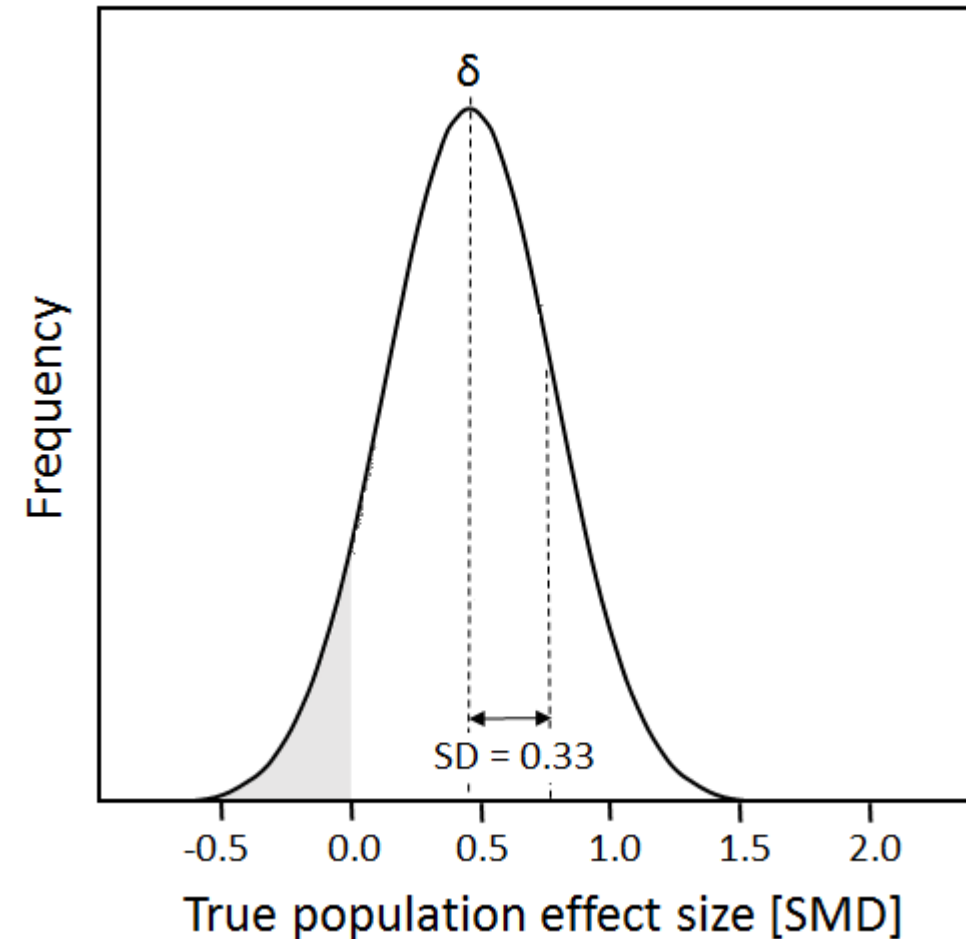
Background

Heterogeneity can be inferred when variability in observed effect sizes exceeds expectation.

Population effect size can no longer be described as single value.

Better described as **variable** with particular M and SD .

τ (SD of population effect size) good measure of heterogeneity.



Background and Aim

Can we trust heterogeneity estimates in meta-analyses? Distorted by **publication bias** (PB) and **questionable research practices** (QRPs)?

Estimates distorted by **publication bias** (PB) and **questionable research practices** (QRPs)?

(PB and QRPs distort effect size estimates in meta-analysis.)

Aim: Find out via computer simulation.

Basic Simulation Idea

Simulation of **simple between-subjects experiments**. Mean difference between CG and EG expressed as d .

Multiple independent studies are

- conducted (with or without QRPs);
- apply two-tailed testing;
- published (or not);
- meta-analysis (Hunter-Schmidt) of all published results;
- heterogeneity estimate T can be compared against true τ .

Factors Manipulated in Simulation

Factors manipulated (and their levels):

- True population effect (0, 0.2, 0.5, 0.8)
- True heterogeneity (0, 0.2, 0.4)
- Number of studies in meta-analysis (10, 30, 60, 100)
- % of non-significant studies suppressed by PB (0, 40, 80)

Factors fully crossed = 144 unique combinations.

Run for 3 different QRP environments (none, medium, high).

Some Additional Details

If a study uses QRPs:

- All possible QRP options are combined.
- Only results with lowest p-value is submitted for publication.
- ($p < .05$ always published; publication of $p > .05$ depends on PB.)

Study N s were drawn from a realistic distribution.

For each specific factor combination, 2,000 meta-analyses were simulated.

$T_{\text{bias}} = T - \tau$ serves as our DV.

Results

No QRPs

Medium QRPs

High QRPs

Publication bias level

0

0.4

0.8

0

0.4

0.8

0

0.4

0.8

tau

--- .00
- - - .20
— .40

10

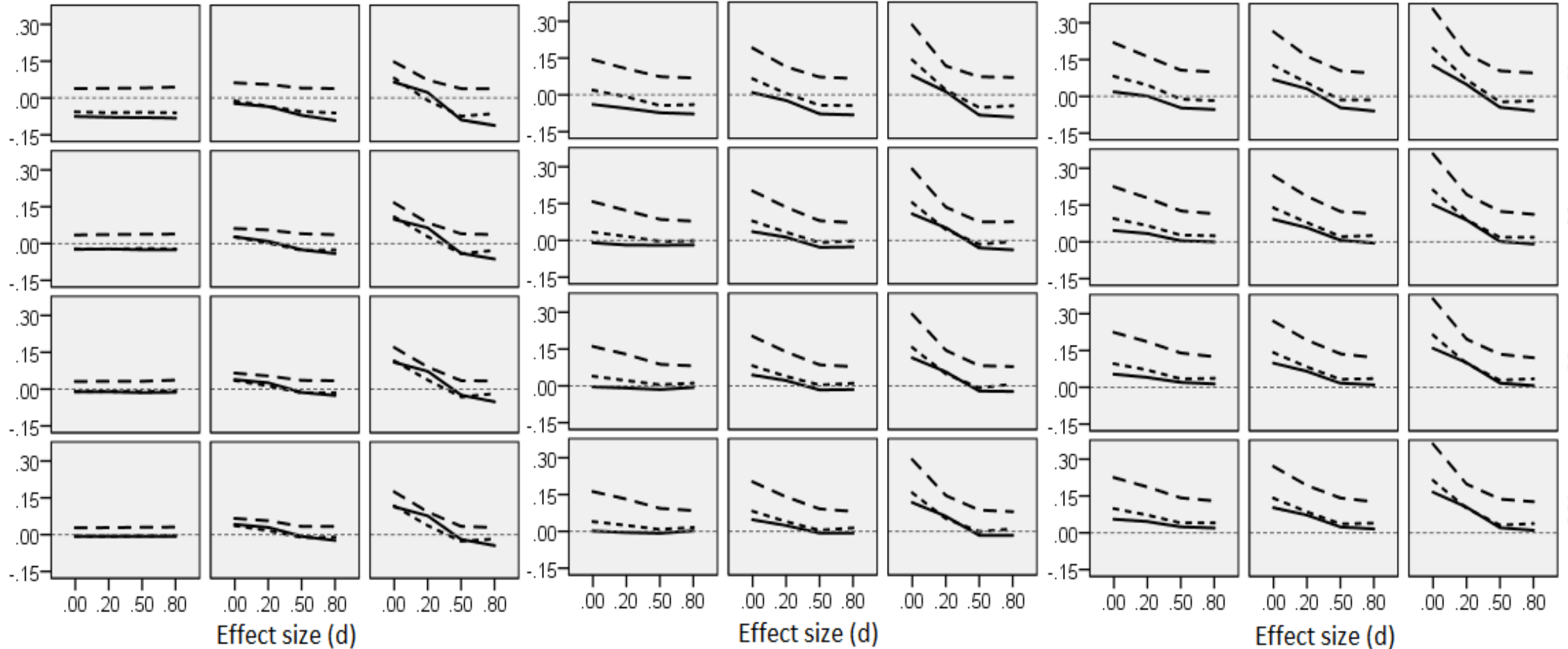
Number of studies (k)

30

60

100

T_{bias}



Results

ANOVA reveals relative importance of manipulated factors.

True population effect size and true population heterogeneity most influential.

(All higher-order interactions proved trivial.)

	No QRPs	Medium QRPs	High QRPs
True effect size (δ)	.22	.21	.18
True heterogeneity (τ)	.29	.37	.23
Number of studies (k)	.07	.03	.02
Publication bias (PB)	.09	.03	.01
$\delta \times \tau$.01	.01	.01
$\delta \times k$.00	.00	.00
$\delta \times \text{PB}$.20	.06	.03
$\tau \times k$.04	.01	.00
$\tau \times \text{PB}$.00	.00	.00
$k \times \text{PB}$.00	.00	.00

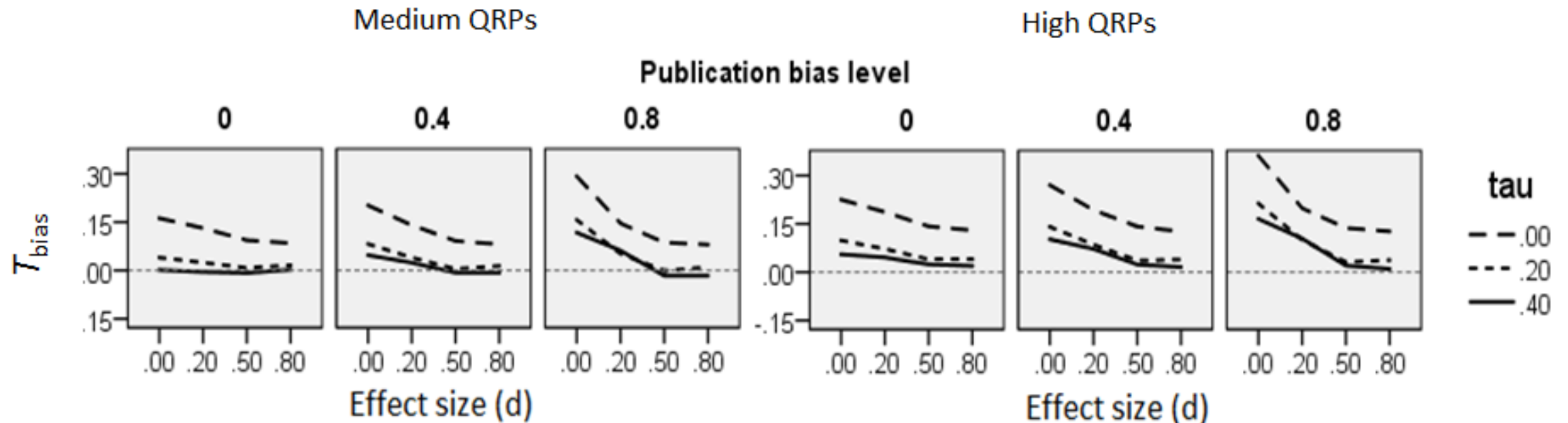
Results

True heterogeneity:

- $\tau = 0.0$: Mean $T_{\text{bias}} = 0.12$
- $\tau = 0.2$: Mean $T_{\text{bias}} = 0.03$
- $\tau = 0.4$: Mean $T_{\text{bias}} = 0.01$

True effect size:

- Zero/S effects \rightarrow bigger T_{bias} .
- M/L effects $\rightarrow T_{\text{bias}}$ small.

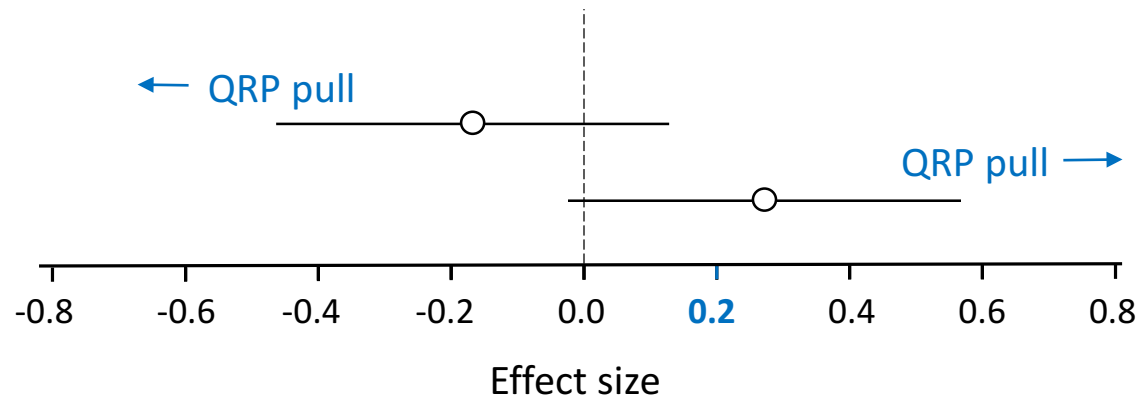


Making Sense of Results

Why is heterogeneity overestimated?

Why is this strongest for $\tau = 0$ and for absent/small effects?

- Where $\tau = 0$, heterogeneity cannot be underestimated.



What Do Results Mean for Observed Heterogeneity?

Do the observed levels of bias matter?

Best addressed in comparison to actual levels of heterogeneity in meta-analyses.

Stay tuned!



Implications

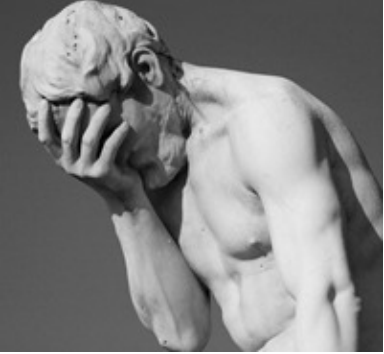


Research planning

Heterogeneity adversely affects power.

Good sample size calculations will take this into account (Kenny & Judd, forthcoming).

Now we know what levels of heterogeneity to expect.



Implications



Empirical cumulativeness

Science = quest to explain apparent complexity in observations through simpler fundamental principles.

(Unexplained) heterogeneity is a measure of how much this quest fails.

On this count, we fail badly.



Implications



Falsification of theories

Weak tools undermine falsification and thereby theoretic progress.

Say test of theory X requires induction of good mood. We use mood induction procedure Y .

When effectiveness of Y is debatable (large heterogeneity), failed test of theory X becomes meaningless.

When knowledge (Y) is used as a tool, we need to **replicate as closely as possible**.



Implications



Testing theories

Large heterogeneity implies that particular finding might not readily generalize.

We therefore need **conceptual replications to test theory** properly.

E.g., don't use standard stimulus set but new stimuli that should work according to theory.



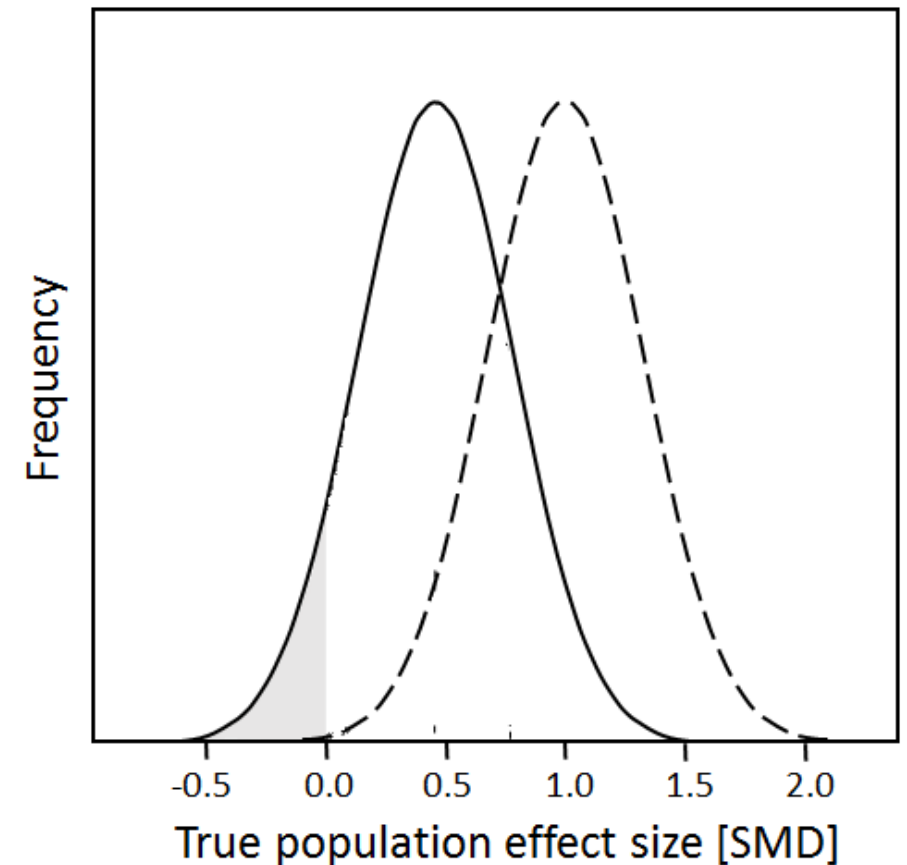
Implications

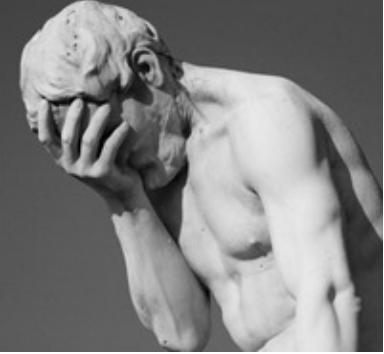


Design of practical applications

Design of practical application might be thought of as 'next study'.

Heterogeneity (and effect size) determine how predictable its result is.





Conclusion



Heterogeneity is a useful but underappreciated tool to evaluate research.

Particularly useful to reflect our lack of understanding.

Considering heterogeneity more routinely should help us to 'fail better'.

**“Ever tried. Ever failed. No matter.
Try again. Fail again. Fail better.”**

