2019-03-20

Correcting for bias in the literature

A comprehensive comparison of meta-analytic methods for bias-correction

Felix Schönbrodt, Evan Carter, Will Gervais, Joe Hilgard

DAGStat 2019

Felix Schönbrodt
Ludwig-Maximilians-Universität
München

www.nicebread.de
www.researchtransparency.org
@nicebread303

LMU Open Science Center

Meta-analysis is at the top of the evidence pyramid - the pinnacle of evidence-based medicine.

Cochrane Collaboration

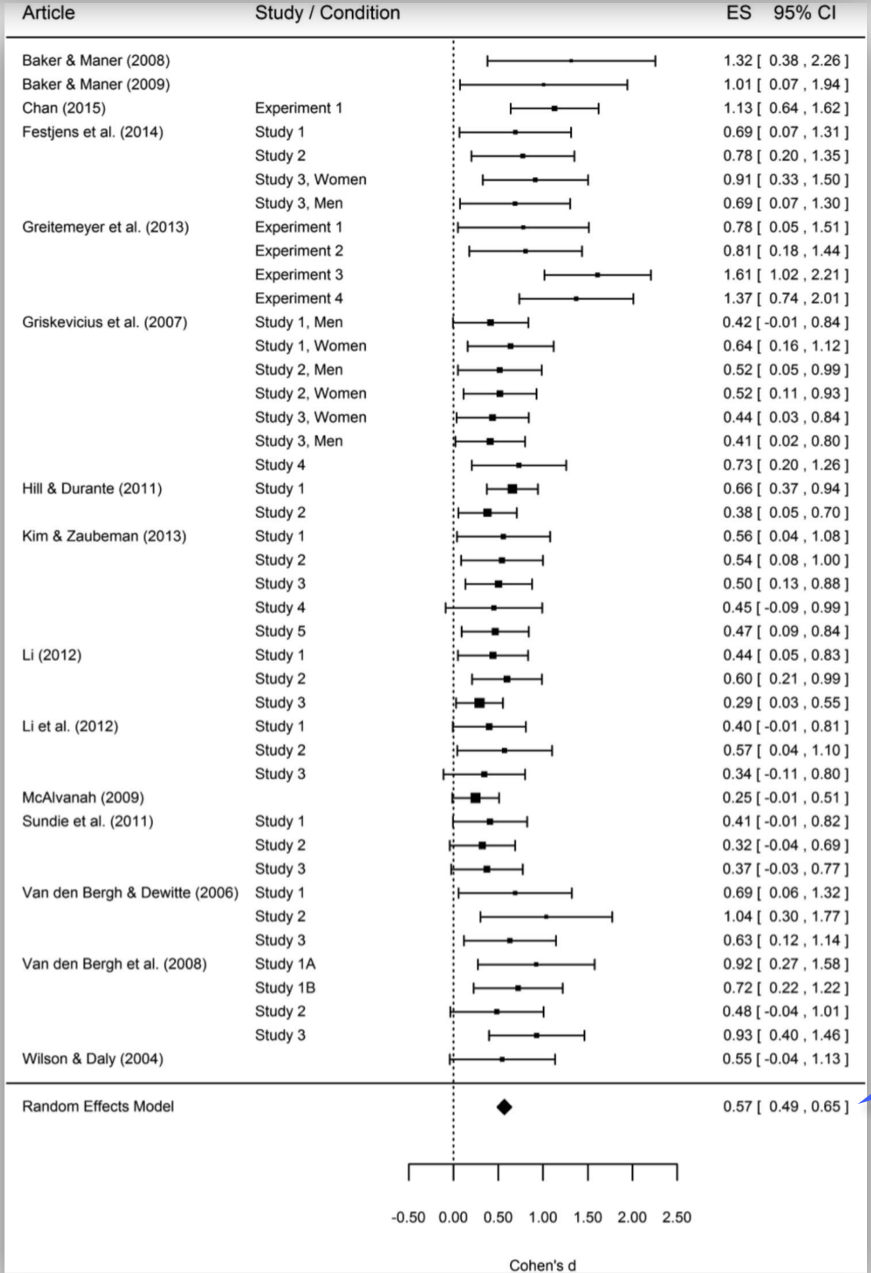
https://uk.cochrane.org/news/meta-analysis-what-why-and-how

Meta-analyses are fucked.

Michael Inzlicht

http://www.slate.com/articles/health_and_science/cover_story/2016/03/ego_depletion_an_influential_theory_in_psychology_may_have_just_been_debunked.single.html

Romance, Risk, and Replication: Can Consumer Choices and Risk-Taking Be Primed by Mating Motives?

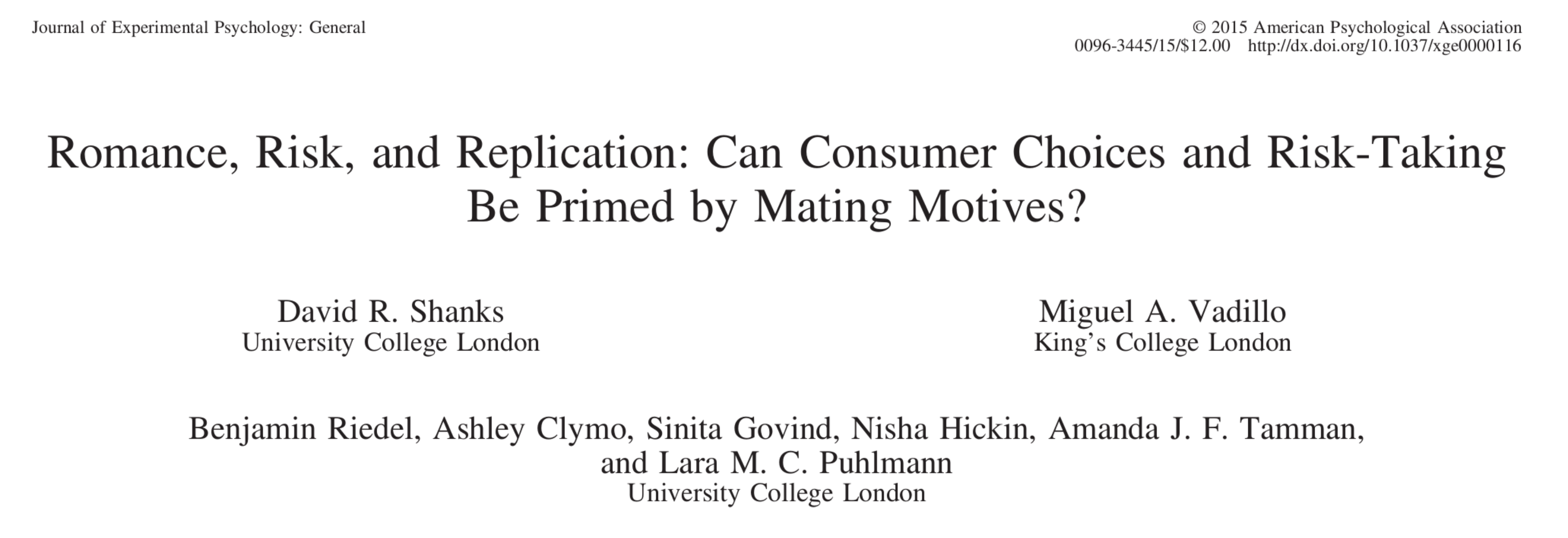


Article	Study / Condition	ES	95% CI
Baker & Maner (2008)		1.32	[0.38 , 2.26]
Baker & Maner (2009)		1.01	[0.07 , 1.94]
Chan (2015)	Experiment 1	1.13	[0.64 , 1.62]
Festjens et al. (2014)	Study 1	0.69	[0.07 , 1.31]
	Study 2	0.78	[0.20 , 1.35]
	Study 3, Women	0.91	[0.33 , 1.50]
	Study 3, Men	0.69	[0.07 , 1.30]
Greitemeyer et al. (2013)	Experiment 1	0.78	[0.05 , 1.51]
	Experiment 2	0.81	[0.18 , 1.44]
	Experiment 3	1.61	[1.02 , 2.21]
	Experiment 4	1.37	[0.74 , 2.01]
Griskevicius et al. (2007)	Study 1, Men	0.42	[-0.01 , 0.84]
	Study 1, Women	0.64	[0.16 , 1.12]
	Study 2, Men	0.52	[0.05 , 0.99]
	Study 2, Women	0.52	[0.11 , 0.93]
	Study 3, Women	0.44	[0.03 , 0.84]
	Study 3, Men	0.41	[0.02 , 0.80]
	Study 4	0.73	[0.20 , 1.26]
Hill & Durante (2011)	Study 1	0.66	[0.37 , 0.94]
	Study 2	0.38	[0.05 , 0.70]
Kim & Zaubeman (2013)	Study 1	0.56	[0.04 , 1.08]
	Study 2	0.54	[0.08 , 1.00]
	Study 3	0.50	[0.13 , 0.88]
	Study 4	0.45	[-0.09 , 0.99]
	Study 5	0.47	[0.09 , 0.84]
Li (2012)	Study 1	0.44	[0.05 , 0.83]
	Study 2	0.60	[0.21 , 0.99]
	Study 3	0.29	[0.03 , 0.55]
Li et al. (2012)	Study 1	0.40	[-0.01 , 0.81]
	Study 2	0.57	[0.04 , 1.10]
	Study 3	0.34	[-0.11 , 0.80]
McAlvanah (2009)		0.25	[-0.01 , 0.51]
Sundie et al. (2011)	Study 1	0.41	[-0.01 , 0.82]
	Study 2	0.32	[-0.04 , 0.69]
	Study 3	0.37	[-0.03 , 0.77]
Van den Bergh & Dewitte (2006)	Study 1	0.69	[0.06 , 1.32]
	Study 2	1.04	[0.30 , 1.77]
	Study 3	0.63	[0.12 , 1.14]
Van den Bergh et al. (2008)	Study 1A	0.92	[0.27 , 1.58]
	Study 1B	0.72	[0.22 , 1.22]
	Study 2	0.48	[-0.04 , 1.01]
	Study 3	0.93	[0.40 , 1.46]
Wilson & Daly (2004)		0.55	[-0.04 , 1.13]
Random Effects Model		0.57	[0.49 , 0.65]

Random effects meta-analytic estimate:
d = 0.57 [0.49; 0.65]

42/43 studies are significant (98% success rate)

Shanks, D. R., Vadillo, M. A., Riedel, B., Clymo, A., Govind, S., Hickin, N., et al. (2015). Romance, Risk, and Replication: Can Consumer Choices and Risk-Taking Be Primed by Mating Motives? Journal of Experimental Psychology: General, 1–18. http://doi.org/10.1037/xge0000116



Journal of Experimental Psychology: General

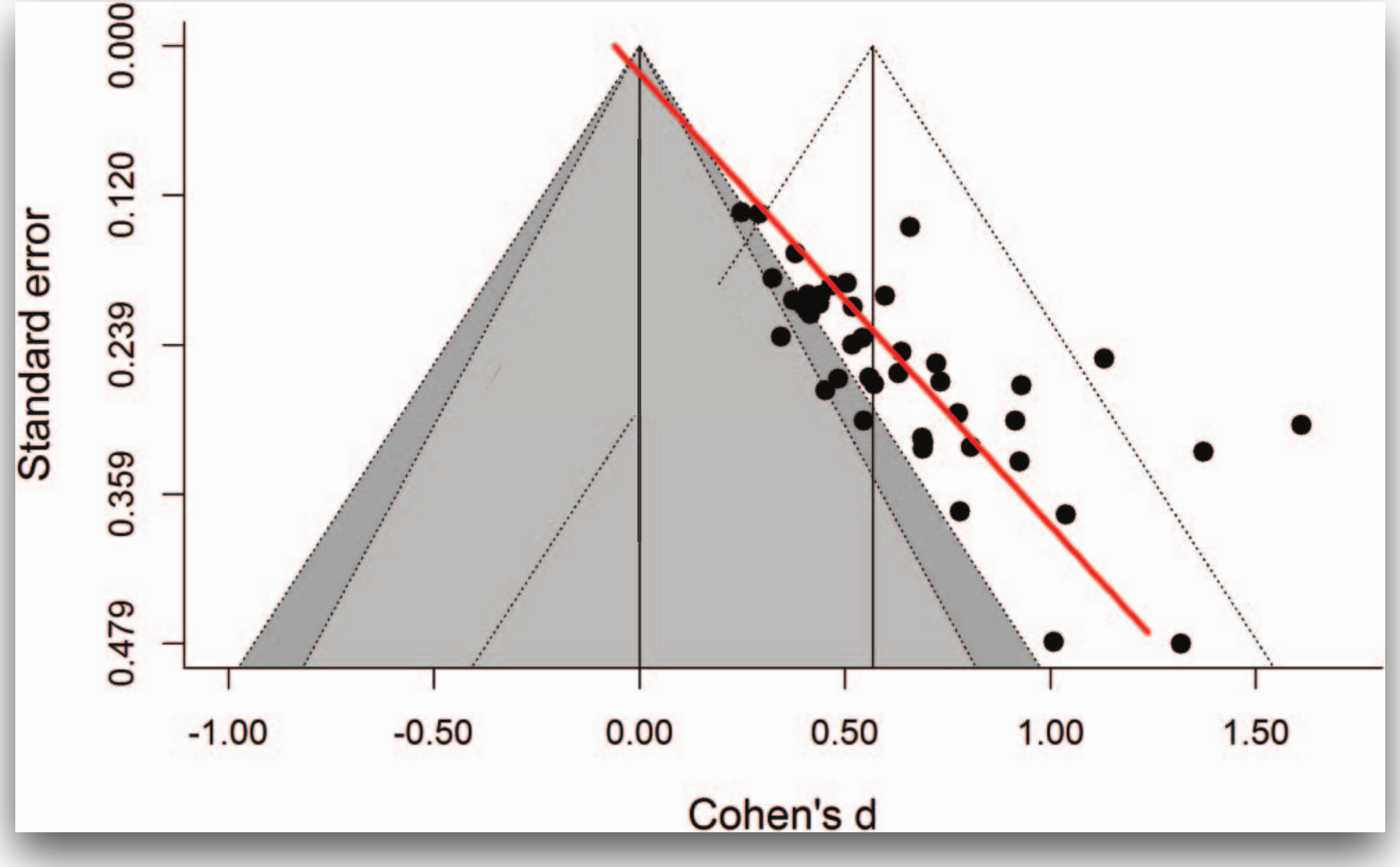
© 2015 American Psychological Association
0096-3445/15/$12.00 http://dx.doi.org/10.1037/xge0000116

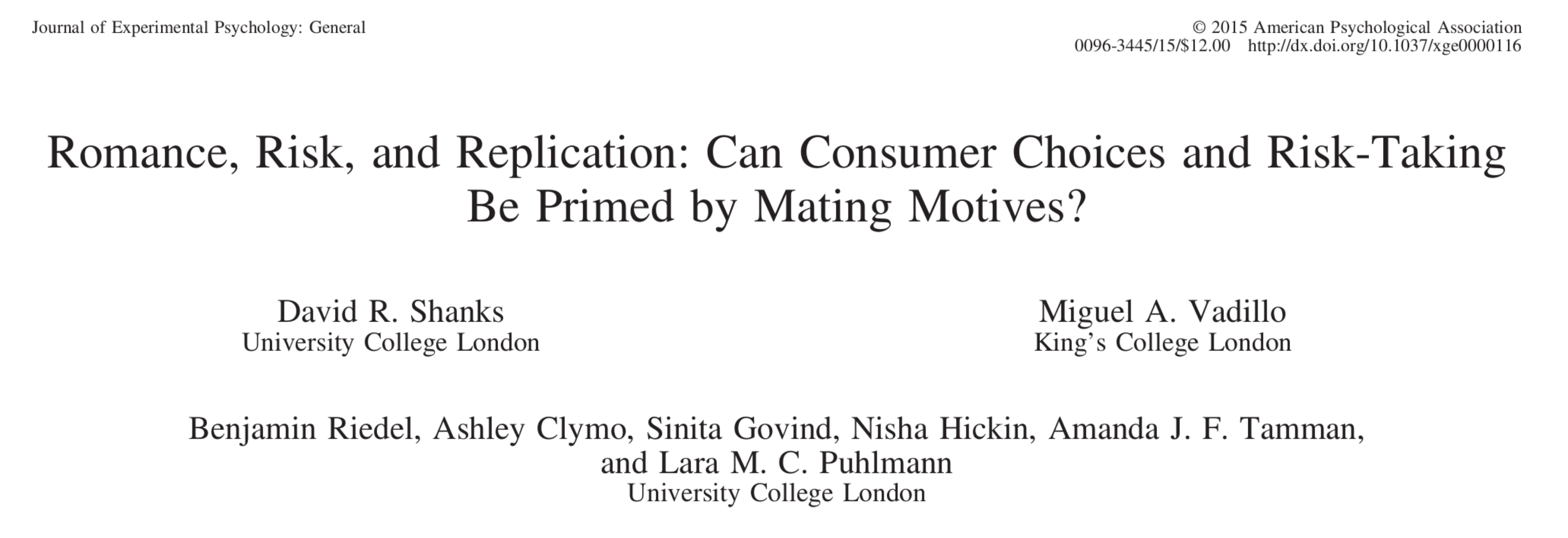
Romance, Risk, and Replication: Can Consumer Choices and Risk-Taking Be Primed by Mating Motives?

David R. Shanks
University College London

Miguel A. Vadillo
King's College London

Benjamin Riedel, Ashley Clymo, Sinita Govind, Nisha Hickin, Amanda J. F. Tamman, and Lara M. C. Puhlmann
University College London





Journal of Experimental Psychology: General

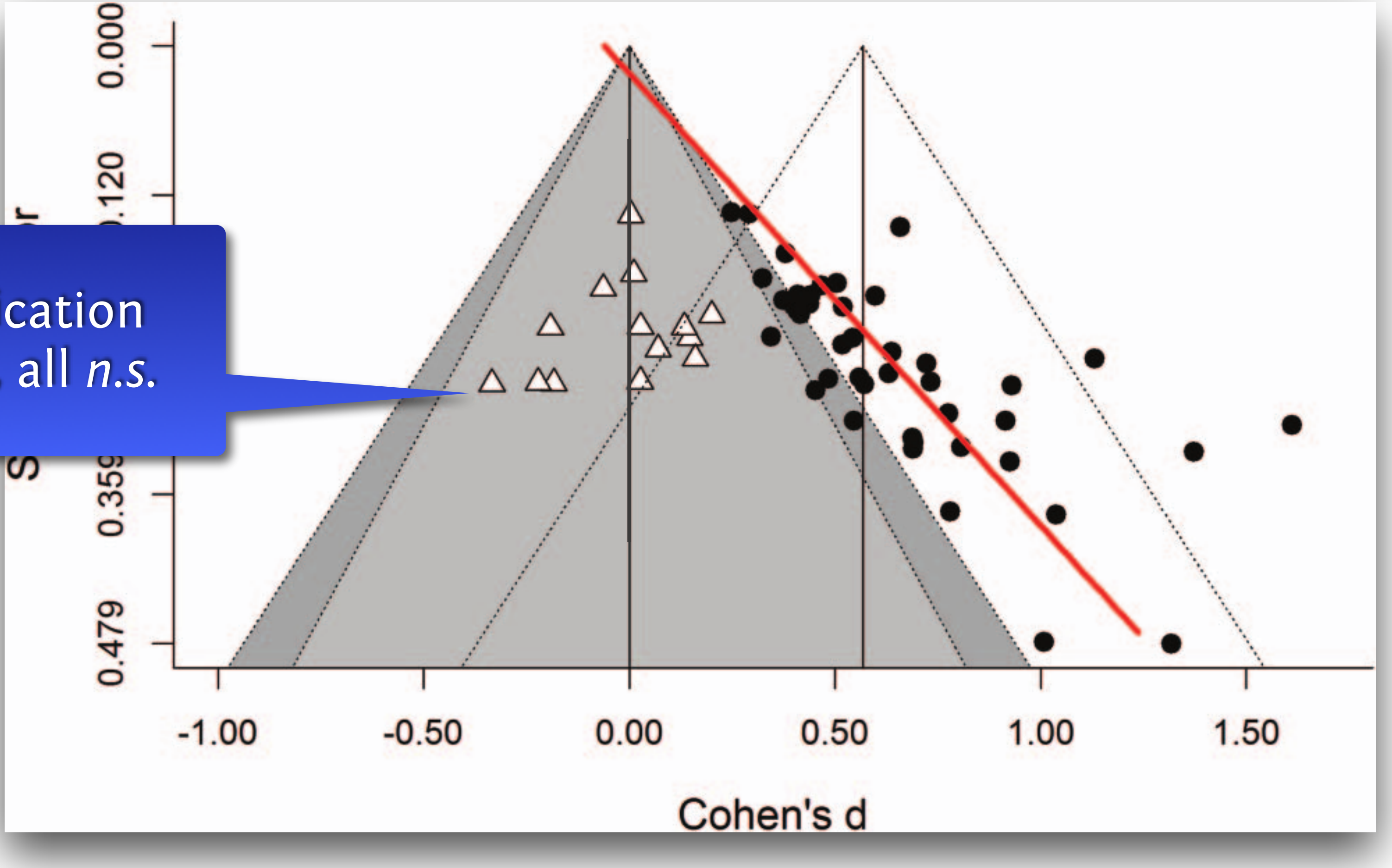
© 2015 American Psychological Association
0096-3445/15/$12.00 http://dx.doi.org/10.1037/xge0000116

Romance, Risk, and Replication: Can Consumer Choices and Risk-Taking Be Primed by Mating Motives?

David R. Shanks
University College London

Miguel A. Vadillo
King's College London

Benjamin Riedel, Ashley Clymo, Sinita Govind, Nisha Hickin, Amanda J. F. Tamman, and Lara M. C. Puhlmann
University College London



14 replication studies, all *n.s.*

Correcting for publication bias (PB) and *p*-hacking

(aka. fishing for significance, data dredging, questionable research practices)

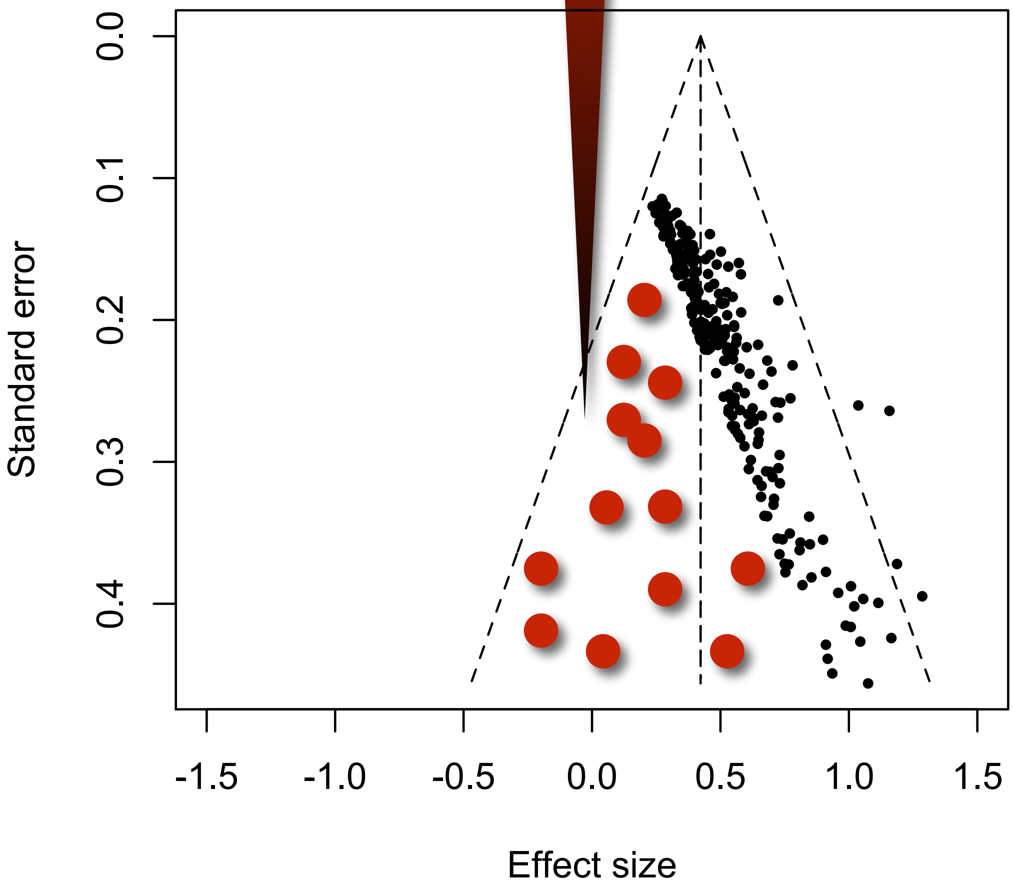
or

Can we clean up the mess,
if we only had the right tool?

Trim & Fill

- Originally designed as a *test* for PB, but also used to *correct* for PB
- Algorithmically fill in missing studies to achieve a symmetric funnel plot
- Compute meta-analysis on the data set including imputed studies

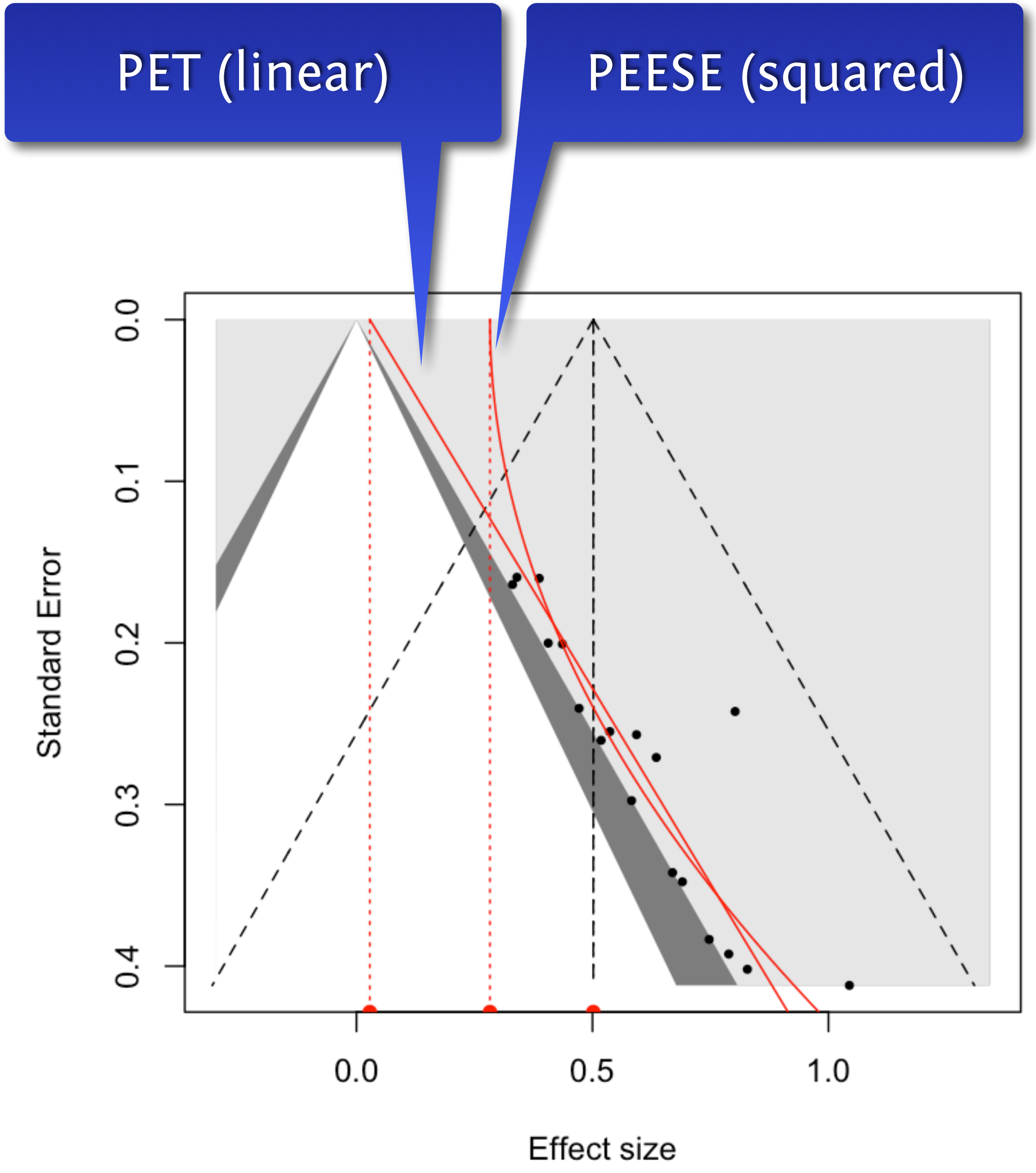
There seem to be some studies missing!
→ trim-and-fill



Duval, S. & Tweedie, R. (2000). Trim and fill: a simple funnel-plot–based method of testing and adjusting for publication bias in meta-analysis. *Biometrics, 56 (2)*, 455–463.

PET / PEESE

- Extrapolates the „small study effect“ to samples with ∞ sample size
- What would be the effect size if we had an infinitely large sample?
- PET – „precision effect test“: linear regression
- PEESE – „precision-effect estimate with standard errors“: squared slope



PET (linear)

PEESE (squared)

ELDERLY PRIMING by D. Lakens

p-curve analysis from Lakens, D. (2014). Professors are not elderly: Evaluating the evidential value of two social priming effects through p-curve analyses. Available at SSRN: http://ssrn.com/abstract=2381936 or http://dx.doi.org/10.2139/ssrn.2381936

https://osf.io/3urp2/

Stanley, T. D., & Doucouliagos, H. (2013). Meta-regression approximations to reduce publication selection bias. Research Synthesis Methods, 5(1), 60–78. http://doi.org/10.1002/jrsm.1095

Selection models

- Explicitly model the functional form of publication bias
- Three-parameter SM:
  - Effect size model (plain random effects meta-analysis):

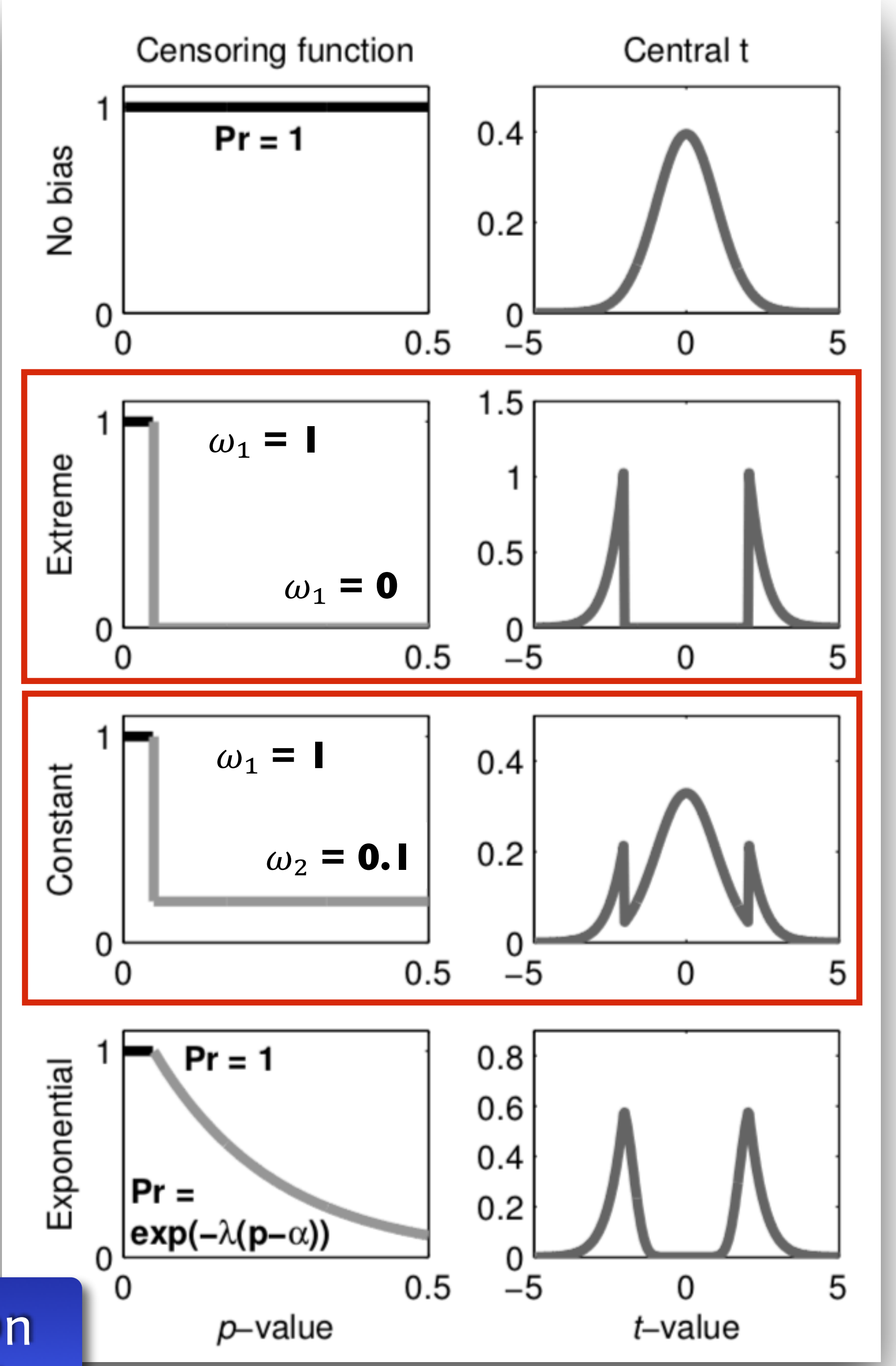
$$Y_i \sim N(\Delta_i, \sigma_i^2 + \sigma^2)$$

heterogeneity

  - Selection model:
    ω = probability of publication
    p_i = p-value of study i

$$w(p_i) = \begin{cases} \omega_1, & \text{if } 0 < p_i \leq .025, \\ \omega_2, & \text{if } .025 < p_i \leq 1. \end{cases}$$

step function with 1 cutpoint



from Guan & Vandekerckhove, 2015)

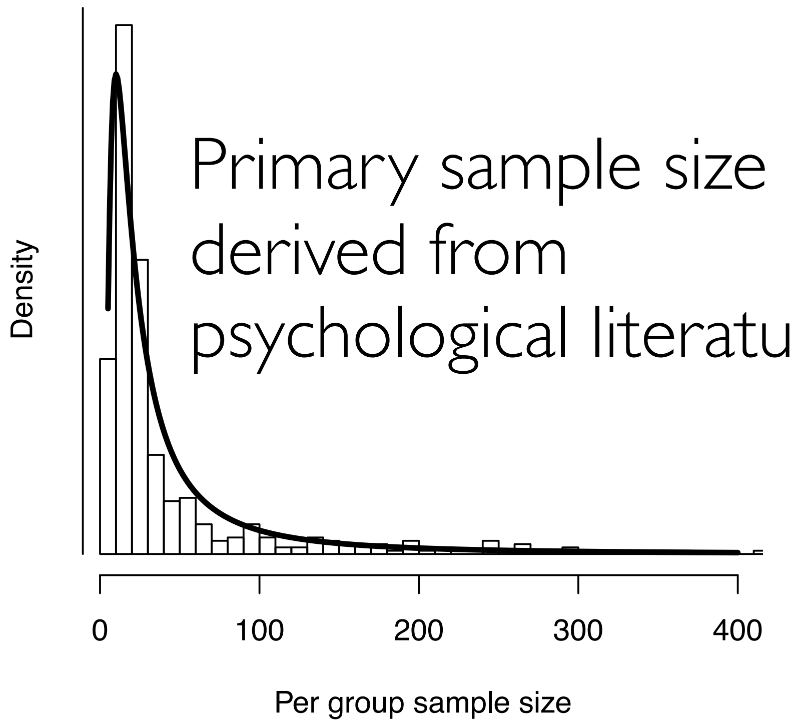
Hedges, L. V. (1984); Vevea & Hedges (1995)
Iyengar, S. & Greenhouse, J. B. (1988)
McShane, B. B., Böckenholt, U., & Hansen, K. T. (2016).

More models …

- **Four parameter selection model**: Adds a cutpoint at $p_{\text{onetailed}} = .5$ (i.e, reversal of sign)
- **WAAP-WLS**: Weighted average of adequately powered studies (Stanley, Doucouliagos, & Ioannidis, 2016): Find which studies have >= 80% power, run meta-analysis only on those.
- ***p*-curve** (Simonsohn et al., 2014) and ***p*-uniform** (van Assen et al., 2015): Variations of selection models
- All of these techniques only model publication bias, but not *p*-hacking!

Performance of bias correcting methods

Simulation study



Primary sample size derived from psychological literature

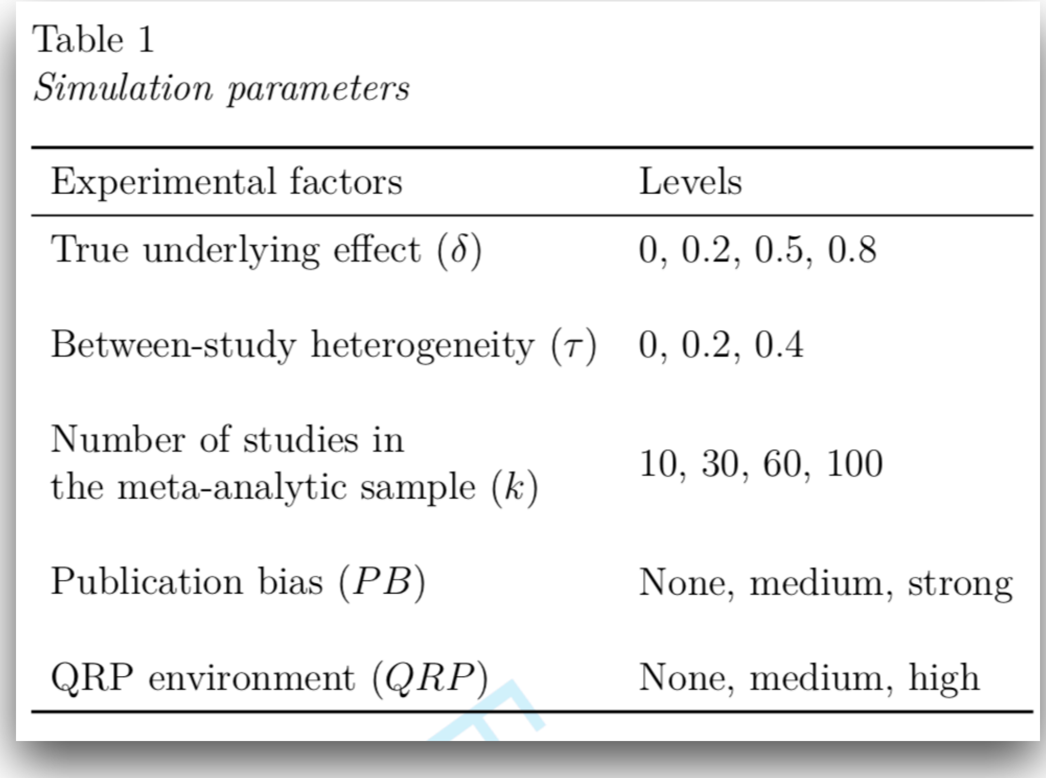


Table 1
Simulation parameters

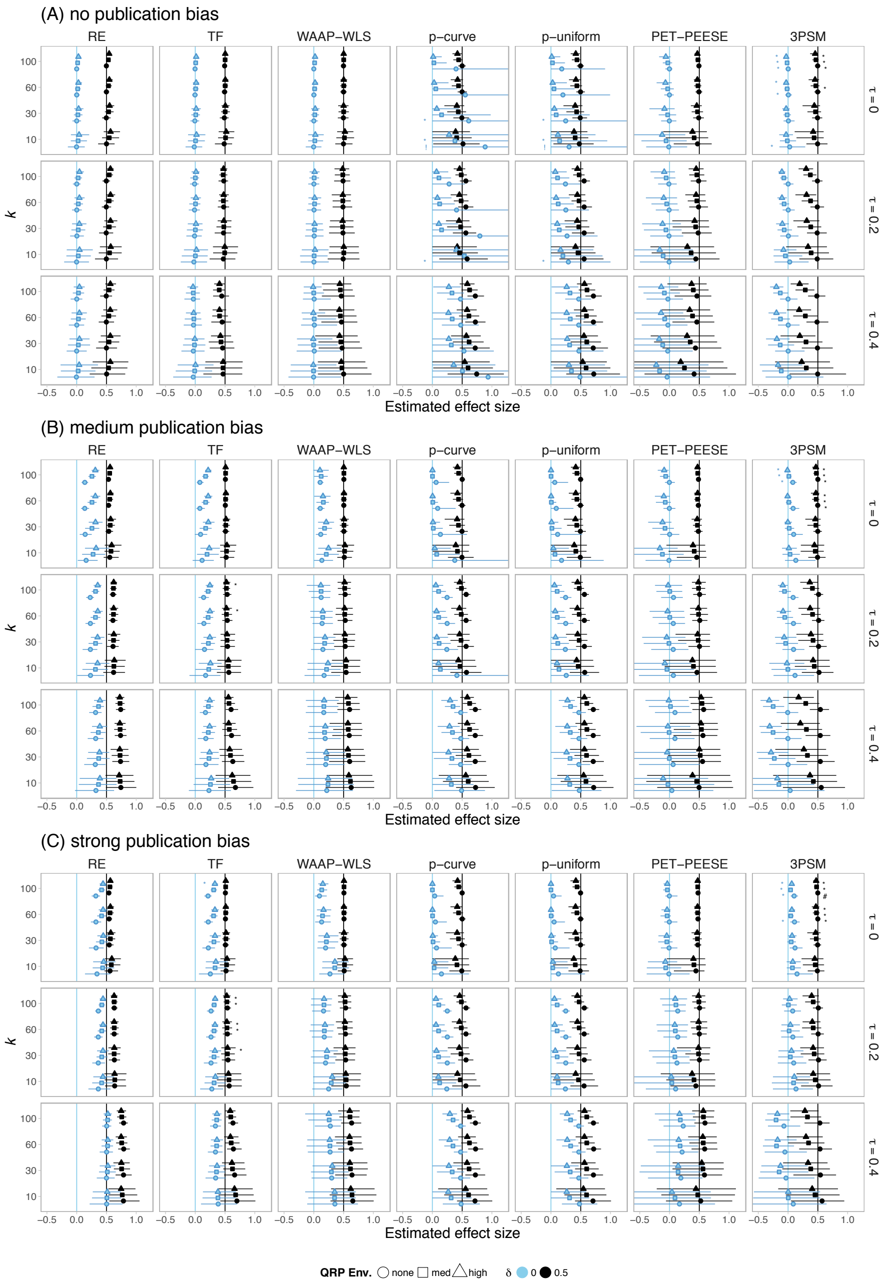
Experimental factors	Levels
True underlying effect (δ)	0, 0.2, 0.5, 0.8
Between-study heterogeneity (τ)	0, 0.2, 0.4
Number of studies in the meta-analytic sample (k)	10, 30, 60, 100
Publication bias (PB)	None, medium, strong
QRP environment (QRP)	None, medium, high

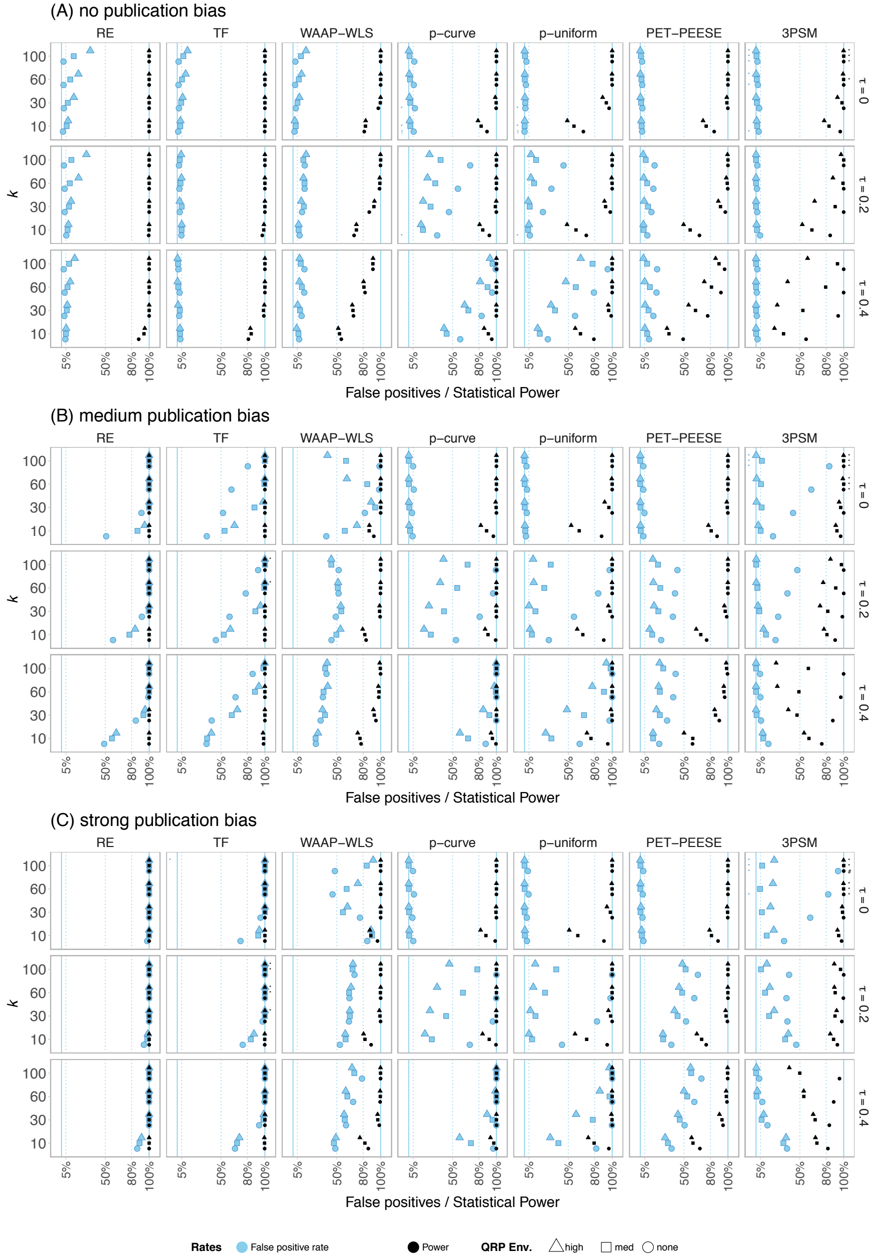
fully crossed:
432 conditions

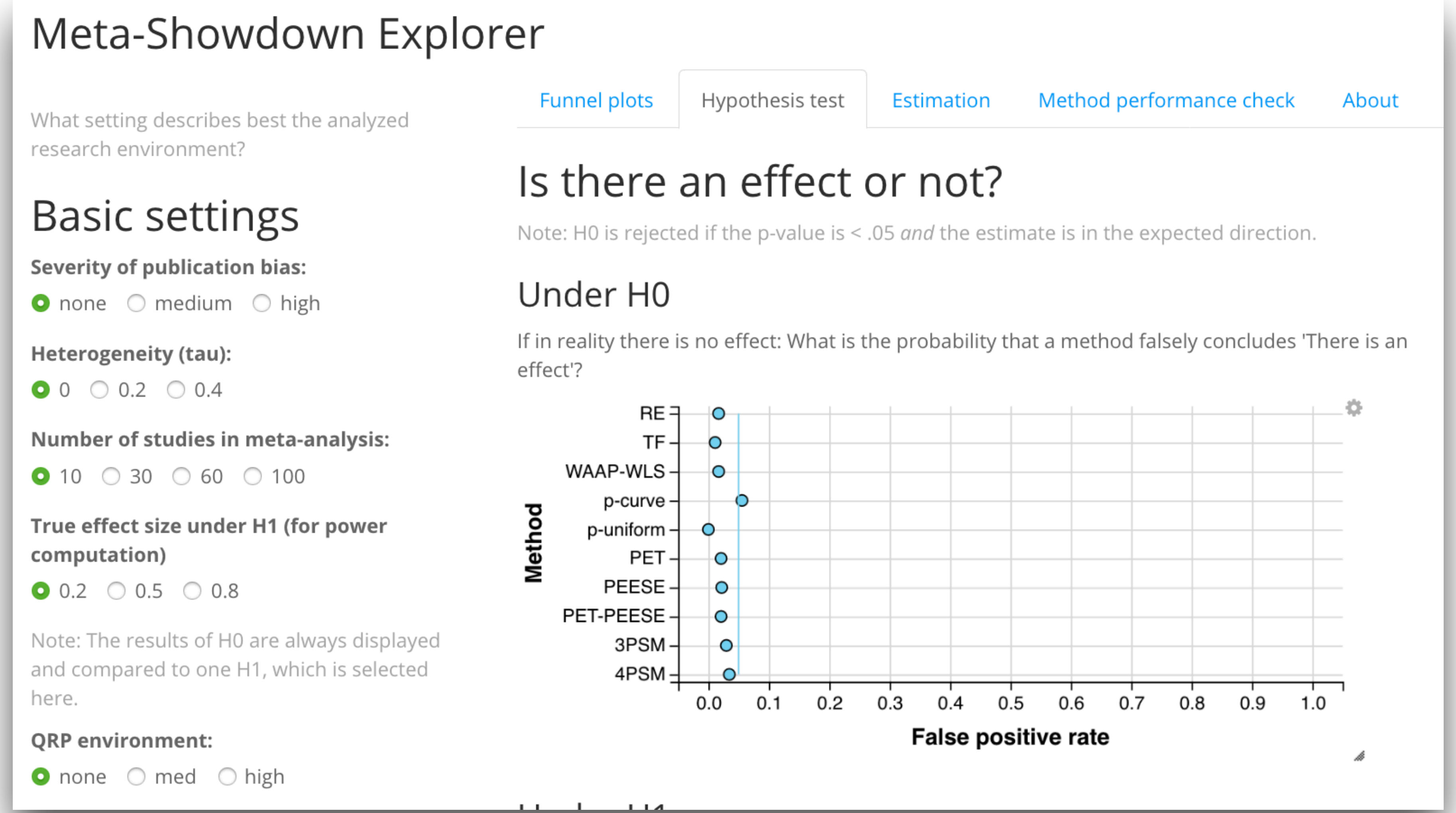
Estimators:

(naive) Random effects meta-analysis, Trim&Fill, PET, PEESE, PET-PEESE, three-parameter selection model (3PSM), four-parameter selection model (4PSM), *p*-curve, *p*-uniform, WAAP-WLS

Results (a selection)







Meta-Showdown Explorer

What setting describes best the analyzed research environment?

Basic settings

Severity of publication bias:
● none ○ medium ○ high

Heterogeneity (tau):
● 0 ○ 0.2 ○ 0.4

Number of studies in meta-analysis:
● 10 ○ 30 ○ 60 ○ 100

True effect size under H1 (for power computation)
● 0.2 ○ 0.5 ○ 0.8

Note: The results of H0 are always displayed and compared to one H1, which is selected here.

QRP environment:
● none ○ med ○ high

Is there an effect or not?

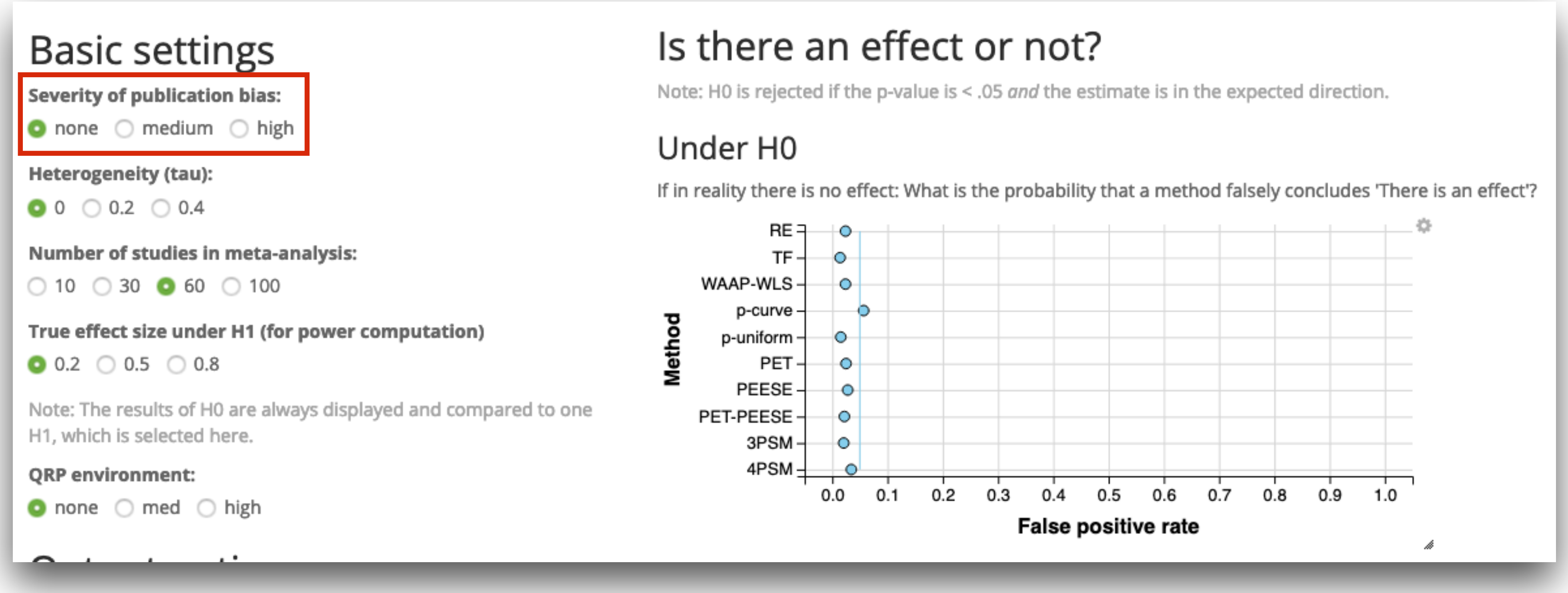
Note: H0 is rejected if the p-value is < .05 *and* the estimate is in the expected direction.

Under H0

If in reality there is no effect: What is the probability that a method falsely concludes 'There is an effect'?

http://shinyapps.org/apps/metaExplorer/

Limits of generalizability: The results are conditional on our implementation of QRPs (not all p-hacking is alike; see van Aert et al. , 2016), our model of publication bias, typical primary study sample sizes and designs in psychology.



Basic settings

Severity of publication bias:

● none ○ medium ○ high

Heterogeneity (tau):

● 0 ○ 0.2 ○ 0.4

Number of studies in meta-analysis:

○ 10 ○ 30 ● 60 ○ 100

True effect size under H1 (for power computation)

● 0.2 ○ 0.5 ○ 0.8

Note: The results of H0 are always displayed and compared to one H1, which is selected here.

QRP environment:

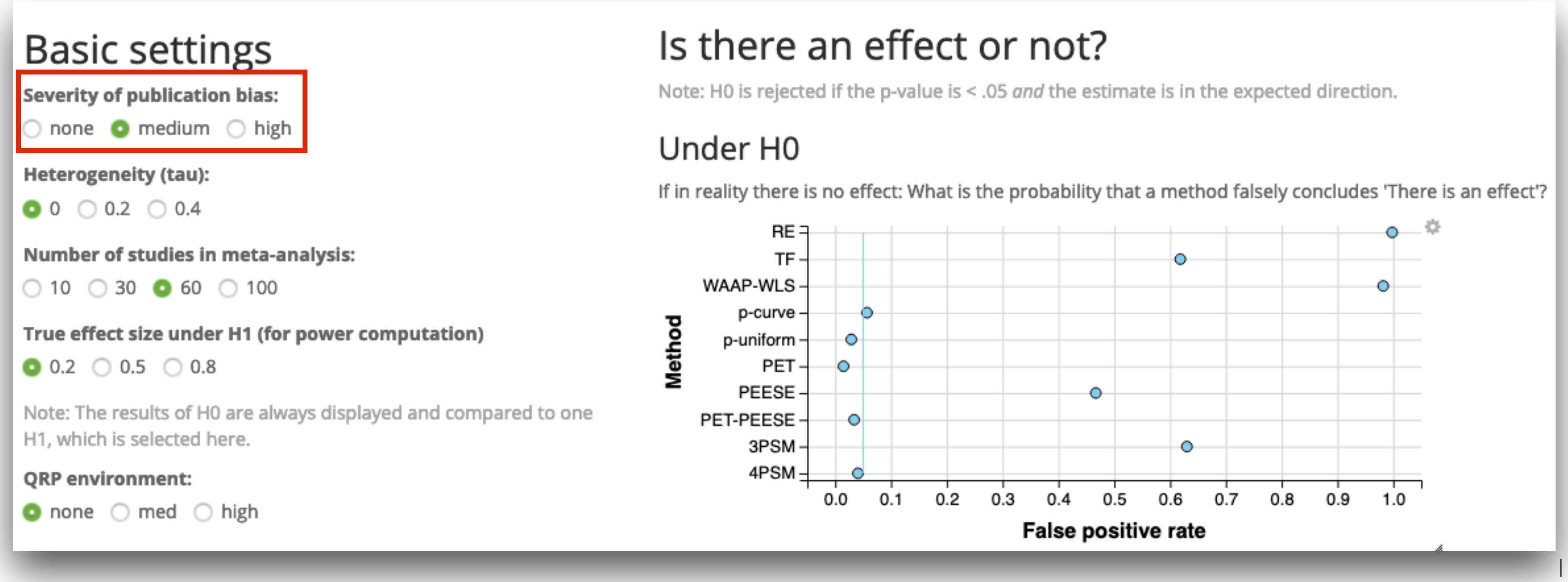
● none ○ med ○ high

Is there an effect or not?

Note: H0 is rejected if the p-value is < .05 *and* the estimate is in the expected direction.

Under H0

If in reality there is no effect: What is the probability that a method falsely concludes 'There is an effect'?



Basic settings

Severity of publication bias:

○ none ● medium ○ high

Heterogeneity (tau):

● 0 ○ 0.2 ○ 0.4

Number of studies in meta-analysis:

○ 10 ○ 30 ● 60 ○ 100

True effect size under H1 (for power computation)

● 0.2 ○ 0.5 ○ 0.8

Note: The results of H0 are always displayed and compared to one H1, which is selected here.

QRP environment:

● none ○ med ○ high

Is there an effect or not?

Note: H0 is rejected if the p-value is < .05 *and* the estimate is in the expected direction.

Under H0

If in reality there is no effect: What is the probability that a method falsely concludes 'There is an effect'?

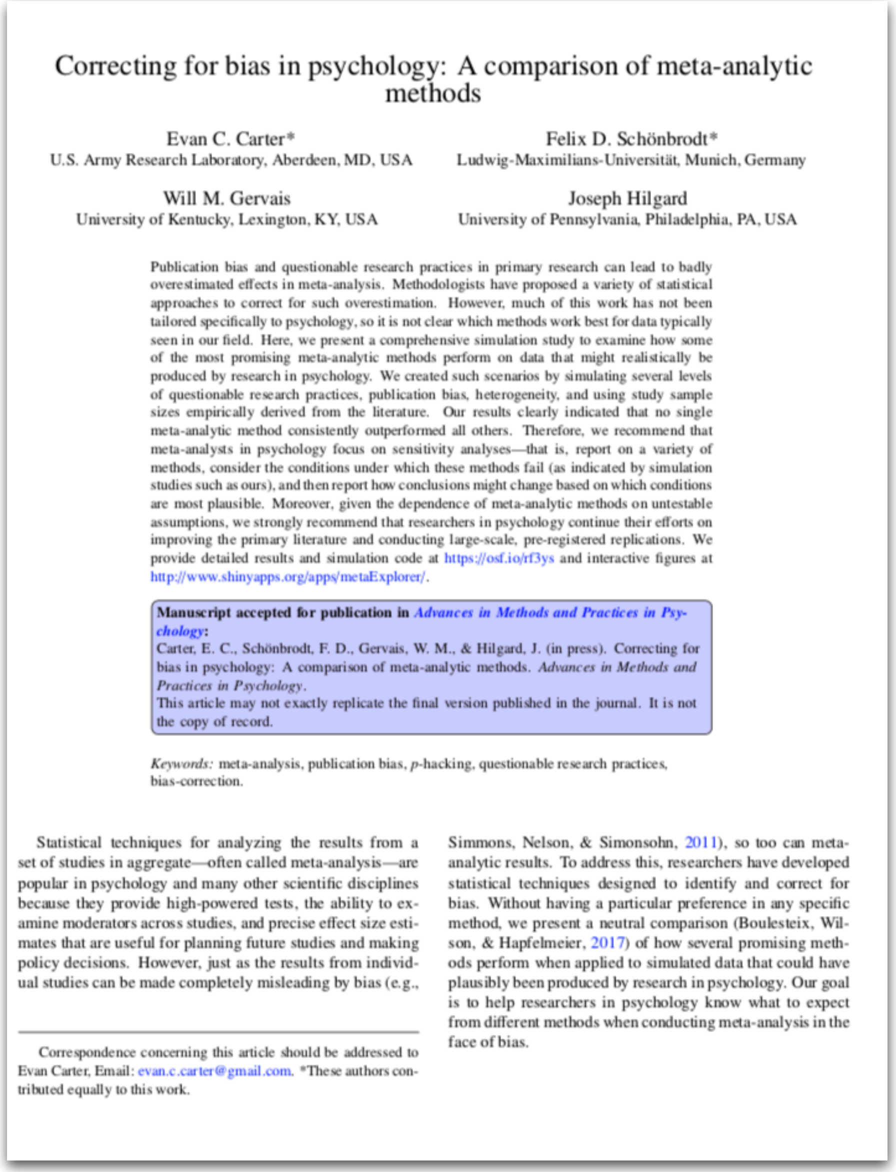
Method performance check

- Each type of bias-correction works in some conditions, but fails in others.
  Problem: Researchers usually don't know which condition they are in.
- Hope that all bias-correcting methods will converge on the same value? Usually that does not happen
- Use the app (or do your own simulations) to see which bias-correcting methods perform well in plausible conditions for the meta-analysis at hand
- ➞No vote counting - no triangulation:
  - Even if three out of four methods converge on a value, this can be irrelevant when those three are known to perform badly in this condition
- Do a sensitivity analysis - but only including methods that passed the performance check!

Meta-analysis – the pinnacle of evidence-based research?

Meta-analyses are fucked?

- Publication bias and *p*-hacking massively distort the evidence: **Garbage in – garbage out / bias in – bias out**
- Even meta-analyses of dozens of significant primary studies can come from a null effect.
- Each type of bias-correction works in some conditions, but fails in others. *Problem*: Don't know which condition we are in.
  - But: Reverting to naive meta-analysis probably is the worst „solution"!
  - Our recommendation: Do a method performance check + sensitivity analysis
- Systematic review is much more than just synthesizing statistics: Rate primary studies for bias, define strict inclusion criteria, etc.
- **Doing biased research and hoping to correct it afterward *does not* work** (at least with the available methods).



Correcting for bias in psychology: A comparison of meta-analytic methods

Evan C. Carter*
U.S. Army Research Laboratory, Aberdeen, MD, USA

Felix D. Schönbrodt*
Ludwig-Maximilians-Universität, Munich, Germany

Will M. Gervais
University of Kentucky, Lexington, KY, USA

Joseph Hilgard
University of Pennsylvania, Philadelphia, PA, USA

Publication bias and questionable research practices in primary research can lead to badly overestimated effects in meta-analysis. Methodologists have proposed a variety of statistical approaches to correct for such overestimation. However, much of this work has not been tailored specifically to psychology, so it is not clear which methods work best for data typically seen in our field. Here, we present a comprehensive simulation study to examine how some of the most promising meta-analytic methods perform on data that might realistically be produced by research in psychology. We created such scenarios by simulating several levels of questionable research practices, publication bias, heterogeneity, and using study sample sizes empirically derived from the literature. Our results clearly indicated that no single meta-analytic method consistently outperformed all others. Therefore, we recommend that meta-analysts in psychology focus on sensitivity analyses—that is, report on a variety of methods, consider the conditions under which these methods fail (as indicated by simulation studies such as ours), and then report how conclusions might change based on which conditions are most plausible. Moreover, given the dependence of meta-analytic methods on untestable assumptions, we strongly recommend that researchers in psychology continue their efforts on improving the primary literature and conducting large-scale, pre-registered replications. We provide detailed results and simulation code at https://osf.io/rf3ys and interactive figures at http://www.shinyapps.org/apps/metaExplorer/.

Manuscript accepted for publication in *Advances in Methods and Practices in Psychology*:
Carter, E. C., Schönbrodt, F. D., Gervais, W. M., & Hilgard, J. (in press). Correcting for bias in psychology: A comparison of meta-analytic methods. *Advances in Methods and Practices in Psychology*.
This article may not exactly replicate the final version published in the journal. It is not the copy of record.

Keywords: meta-analysis, publication bias, *p*-hacking, questionable research practices, bias-correction.

Statistical techniques for analyzing the results from a set of studies in aggregate—often called meta-analysis—are popular in psychology and many other scientific disciplines because they provide high-powered tests, the ability to examine moderators across studies, and precise effect size estimates that are useful for planning future studies and making policy decisions. However, just as the results from individual studies can be made completely misleading by bias (e.g., Simmons, Nelson, & Simonsohn, 2011), so too can meta-analytic results. To address this, researchers have developed statistical techniques designed to identify and correct for bias. Without having a particular preference in any specific method, we present a neutral comparison (Boulesteix, Wilson, & Hapfelmeier, 2017) of how several promising methods perform when applied to simulated data that could have plausibly been produced by research in psychology. Our goal is to help researchers in psychology know what to expect from different methods when conducting meta-analysis in the face of bias.

Correspondence concerning this article should be addressed to Evan Carter, Email: evan.c.carter@gmail.com. *These authors contributed equally to this work.

Carter, E. C., Schönbrodt, F. D., Gervais, W. M., & Hilgard, J. (in press). Correcting for bias in psychology: A comparison of meta-analytic methods. *Advances in Methods and Practices in Psychology.*

https://psyarxiv.com/9h3nu/

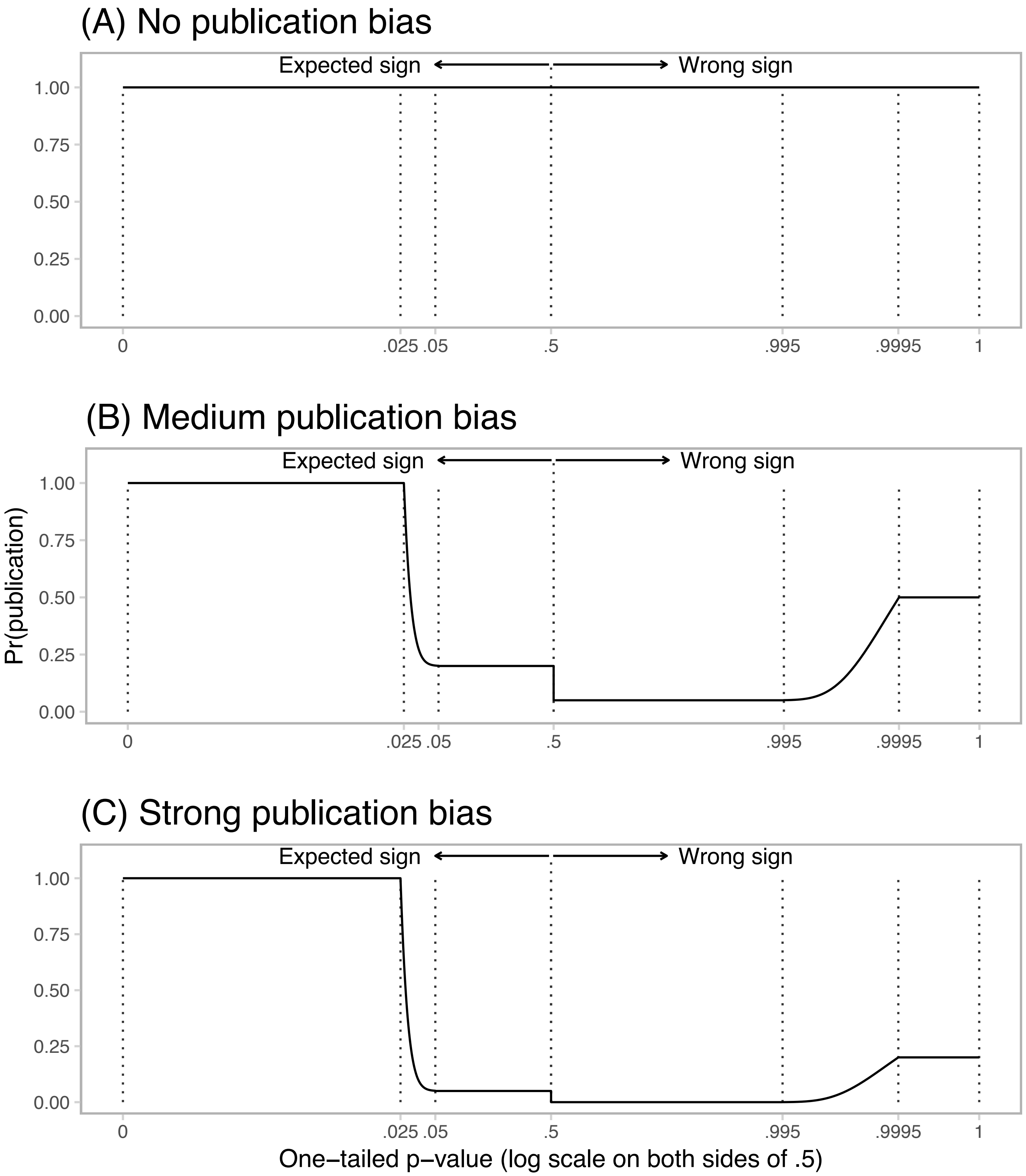
- „Researchers should **not expect** to produce a conclusive, **debate-ending result** by conducting a meta-analysis on an existing literature“
- There is no alternative to making primary studies more transparent, credible, and reproducible.

Interactive results visualization: http://shinyapps.org/apps/metaExplorer/

OPEN MATERIALS

Fully reproducible code (open license): https://github.com/nicebread/meta-showdown

Speicher



(A) No publication bias

(B) Medium publication bias

(C) Strong publication bias