2019-01-25



Correcting for bias in the literature

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

Felix Schönbrodt, Evan Carter, Will Gervais, Joe Hilgard

This presentation is licensed under a <u>CC-BY 4.0 license</u>. You may copy, distribute, and use the slides in your own work, as long as you give attribution to the original author at each slide that you use.



OSC LMU Open Science Center

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

Felix Schönbrodt Ludwig-Maximilians-Universität München



<u>Meta-analysis</u> is at the top of the evidence-based medicine pyramid - the pinnacle of evidence-based medicine.

Cochrane Collaboration

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

Meta-analyses are fucked.

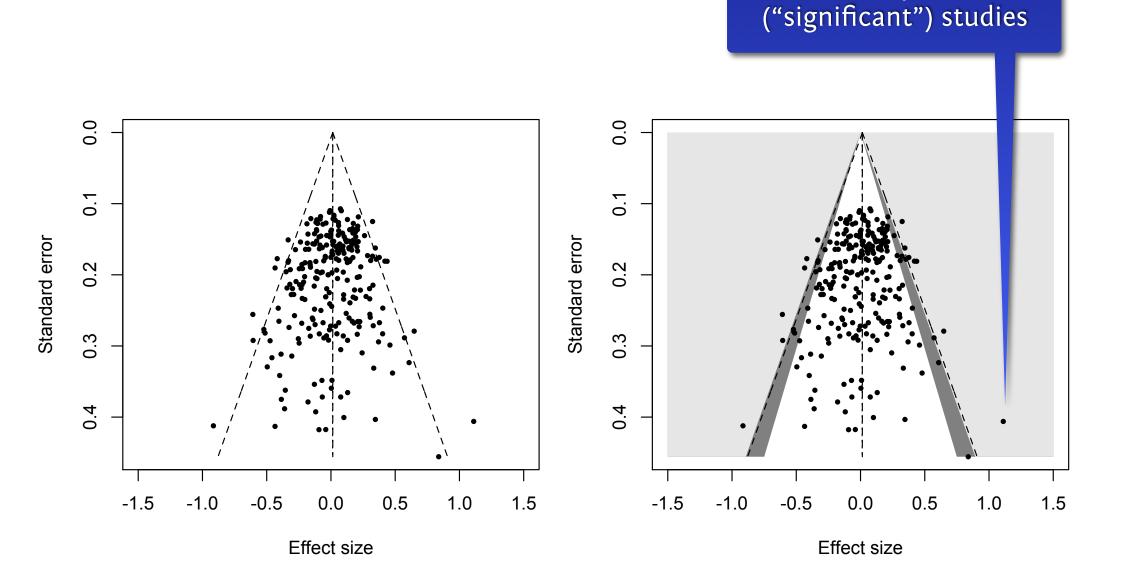
Mickey Inzlicht

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

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]	
Hill & Durante (2011)	Study 4 Study 1		0.73 [0.20 , 1.26] 0.66 [0.37 , 0.94]	
Hill & Durante (2011)				
Kim & Zouhomon (2012)	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]	
1 (0010)	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	⊨ − 1	0.37 [-0.03 , 0.77]	
Van den Bergh & Dewitte (2006)	Study 1	· · · · · · · · · · · · · · · · · · ·	0.69 [0.06 , 1.32]	
	Study 2	i	1.04 [0.30 , 1.77]	
	Study 3	· · · · · ·	0.63 [0.12 , 1.14]	
Van den Bergh et al. (2008)	Study 1A	i i	0.92 [0.27 , 1.58]	
	Study 1B	i ⊢I	0.72 [0.22 , 1.22]	
	Study 2	<u>⊨</u>	0.48 [-0.04 , 1.01]	
	Study 3	i	0.93 [0.40 , 1.46]	
Wilson & Daly (2004)		I	0.55 [-0.04 , 1.13]	
Random Effects Model		•	0.57 [0.49 , 0.65]	
	I			
	-0.	50 0.00 0.50 1.00 1.50 2.00 2.50		
		Cohen's d		

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

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



True H₀ samples*

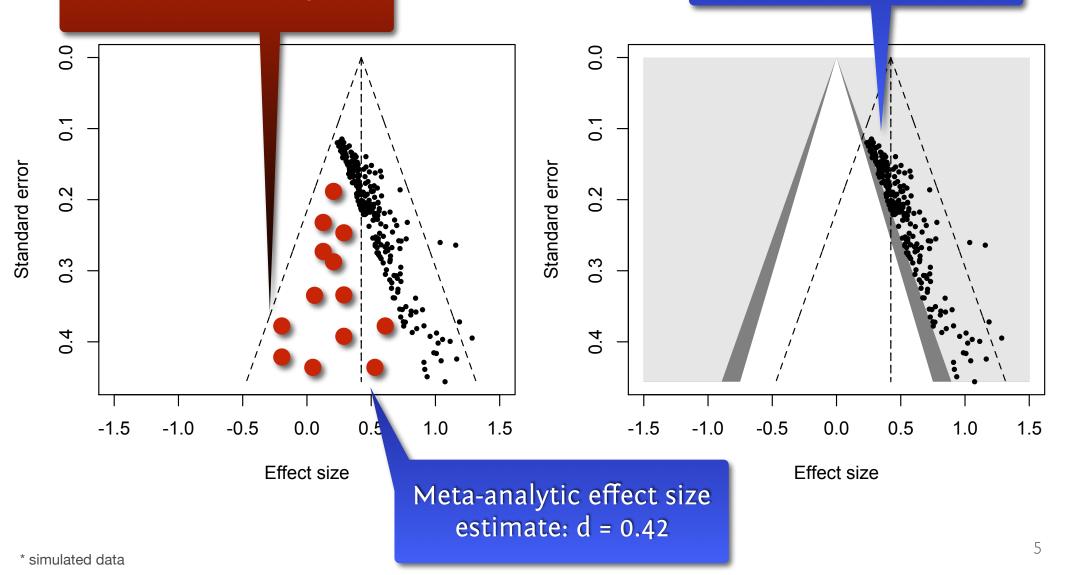
* simulated data

5% false positive

True H_0 + directional publication bias

There seem to be some studies missing!

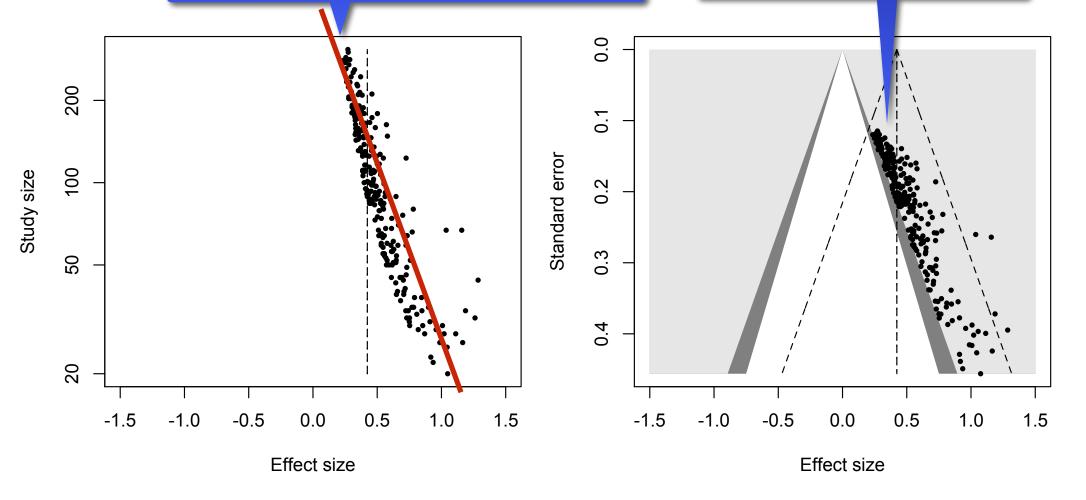
Studies "huddle" against the significance threshold



True H_0 + publication bias

Negative correlation of study size & estimated effect size: Smaller studies have larger effects

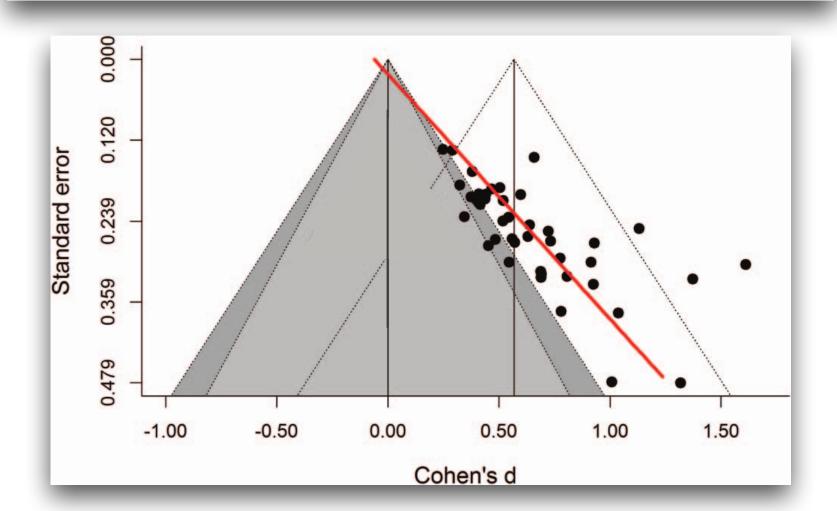
Studies "huddle" against the significance threshold



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

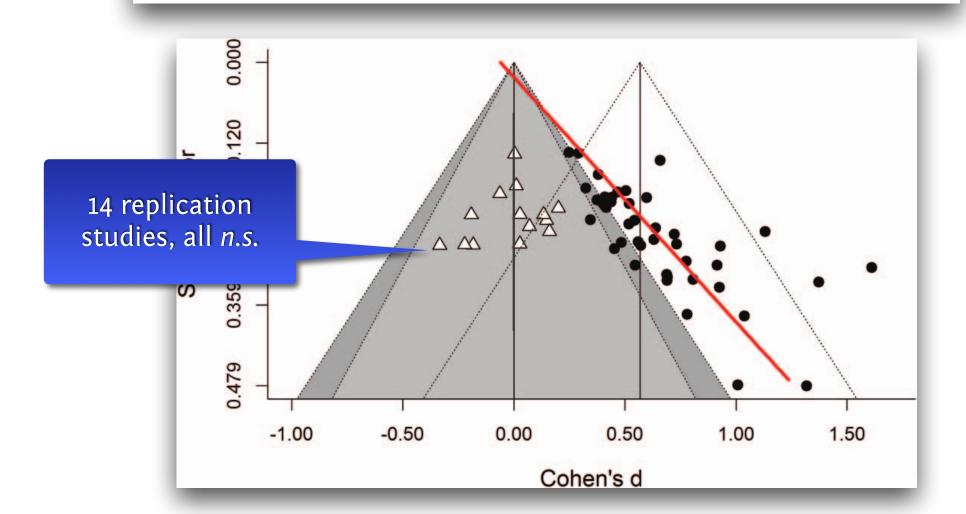
Benjamin Riedel, Ashley Clymo, Sinita Govind, Nisha Hickin, Amanda J. F. Tamman, and Lara M. C. Puhlmann University College London



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



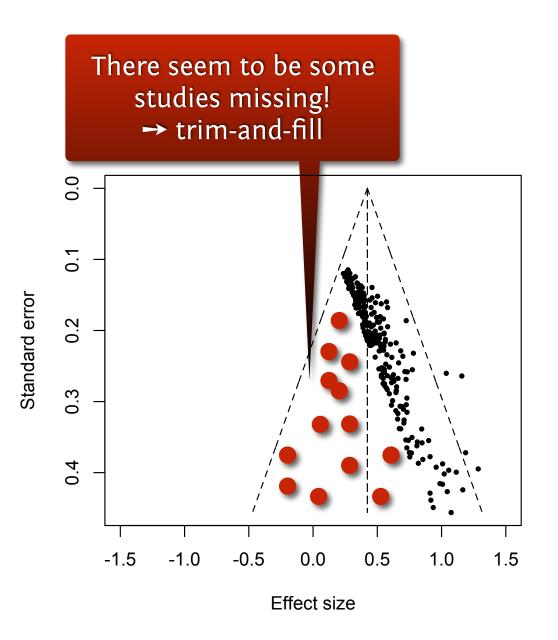
Correcting for publication bias (PB)

Or

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

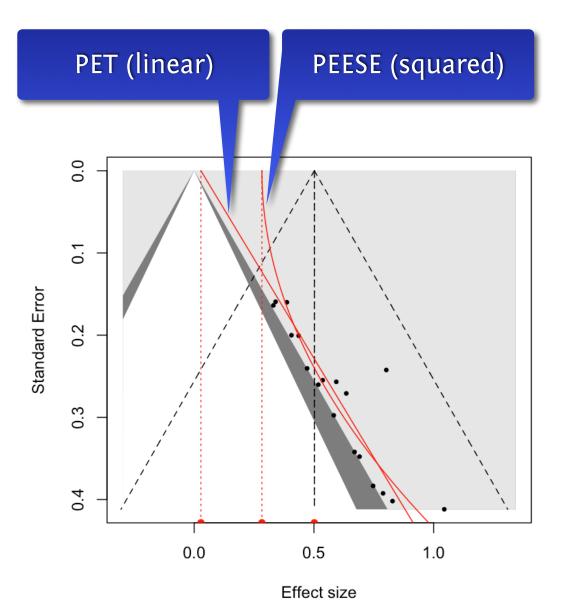
Trim & Fill

- 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



PET / PEESE

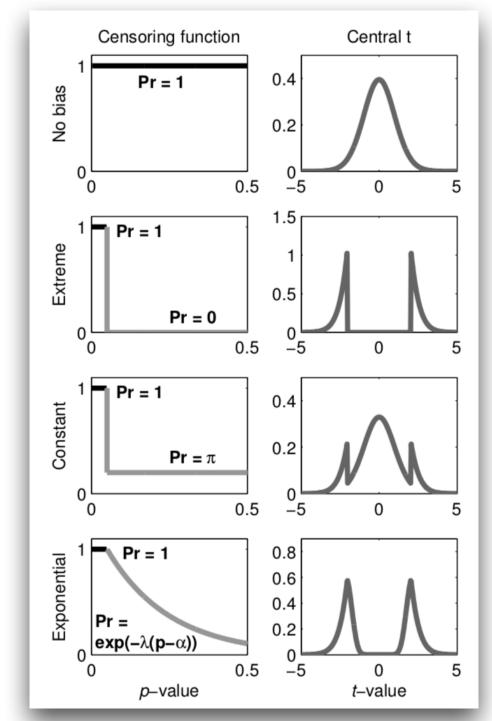
- Extrapolates the ,,small study effect⁺⁺ to samples with ∞ sample size
- What would be the effect size if we had an infinitely large sample?
- PET: linear regression
- PEESE: squared slope



Selection models

- Explicitly model the functional form of publication bias
- Provide estimates for, e.g., *Prob*(published | n.s.)
- Three-parameter SM: μ, τ, and Prob(published | n.s.)
- Four-parameter SM: μ, τ, and *Prob*(pub | n.s. & correct direction) and *Prob*(pub | wrong direction)

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



from Guan & Vandekerckhove, 2015)

Performance of bias correcting methods

Simulation study

Table 1 Simulation parameters		Density		deriv	ved fro		size erature
Experimental factors	Levels						
True underlying effect (δ)	0, 0.2, 0.5, 0.8	с О)	100	200	300	400
Between-study heterogeneity (τ)	0,0.2,0.4			Per	group samp	le size	
Number of studies in the meta-analytic sample (k)	10, 30, 60, 100						
Publication bias (PB)	None, medium, strong		fu	lly cro	ossed:		
QRP environment (QRP)	None, medium, high		43	32 со	nditio	ns	

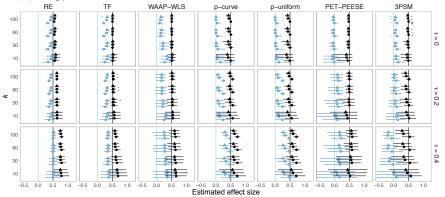
Estimators:

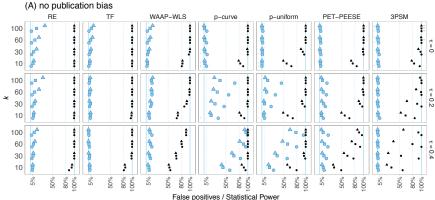
(naive) Random effects meta-analysis, Trim&Fill, PET, PEESE, PET-PEESE, threeparameter selection model (3PSM), four-parameter selection model (4PSM), *p*-curve, *p*-uniform, WAAP-WLS

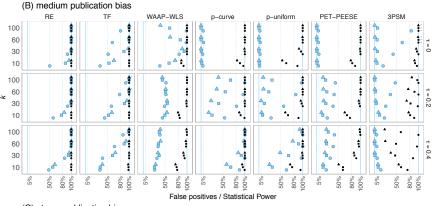
Results (a selection)

(A) no publication bias BF TE WAAP-WLS p-uniform PET-PEESE 3PSM p-curve -* 60 14 **4**0 ÷ 30 -Ē \$ -# - 4 ŧ **≵** ≢ ŧ ***** \$ ŧ 10 = 1 ≢ ***** 1 **\$**____ * 圭 100 --# ≢ ** Ē 60 \$ 主 1 ŧ ≇≢ 4 + ŧ * 30 **≢**≢ **___**≢ **⇒**≢ 4 10 ŧ 4 --*** *** \$ \$ 100 -\$ -÷ --60 \$ + _____ ± 30 * 0.0 0.5 1.0 -0.5 0.0 0.5 1.0 -0.5 0.0 0.5 4 **__** -0.5 0.0 0.5 1.0 -0.5 0.0 0.5 1.0 -0.5 0.0 0.5 1.0 -0.5 -05 00 05 Estimated effect size (B) medium publication bias WAAP-WLS PET-PEESE RF TE n_uniform 3PSM n_curve ***** 1 4 <u></u> \$ 쇸 송 4 *** -* - **1** 4 * 60 4 €_₹ * <u>*</u> <u></u>≢ ‡ 30 ***** -= **t** <u>*</u> 4 10 <u></u> -<u>_</u> ±‡ **a a \$**_____ -100 송 *** *** * ***____ ____**‡** * * 2 I ŧ₽ 60 * = ***** --* **= *** 30 <u>_____</u> • 10 --<u>+</u> ***** * <u>___</u> 100 ***= ≵** ‡ * *** *** 60 ⇒主 ***** --4 4 30 5 10 -0.5 0.0 0.5 1.0 -0.5 0.0 0.5 1.0 Estimated effect size -0.5 0.0 0.5 1.0 -0.5 0.0 0.5 1.0 -0.5 0.0 0.5 1.0 -0.5 0.0 0.5 -05 00 05

(C) strong publication bias







(C) strong publication bias TF WAAP-WLS p-curve PET-PEESE 3PSM RE p-uniform 100 1 1 1 . **a**≜ ‡ **.** 60 1 1 1 4 30 1 ÷ 1 4 ÷., 10 ÷, 1 1 • 100 4 t 1 60 1 â. ŧ 1 \$ ÷. t 4 ÷. 30 t 1 <u>ن</u> 1 \$ 1 4 4 ***** A, 10 ŝ ٩, t ÷ 4 1 ۹. **^** 100 **1 ^** 4. 1 60 1 R. 1 1 . ٩. A. 1 4 4 ٩. 30 4 1 é 🛔 £. **A** = **a** ▲. ▲ •. ۰. 10 8 ê 🗄 80% 00% 2% %0 2% 50% 30% % %0 30% 5% %0% 80% 00% % %09 30% % %00 30% 5% %09 30% False positives / Statistical Power

Meta-Showdown Explorer

What setting describes best the analyzed research environment?

Basic settings

Severity of publication bias:

O none ○ medium ○ high

Heterogeneity (tau):

0 0 0.2 0.4

Number of studies in meta-analysis:



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:

O none ○ med ○ high

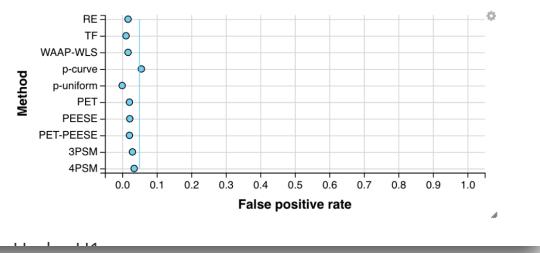
Funnel plots	Hypothesis test	Estimation	Method performance check	About
s thoro	an offoct	or pot?		

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/

Hypothesis test

How many % of original studies are submitted to publication bias?:

€ 0% ○ 60% ○ 90%

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)

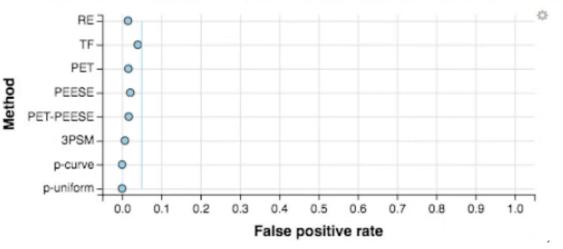
0.2 0.5 0.8

QRP environment:

o none ○ med ○ high

Under H0

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



Effect size estimation

Basic settings

How many % of original studies are submitted to publication bias?:

• 0% ○ 60% ○ 90%

Heterogeneity (tau):

○ 0 ○ 0.2 ○ 0.4

Number of studies in meta-analysis:

○ 10 ○ 30 ○ 60 ○ 100

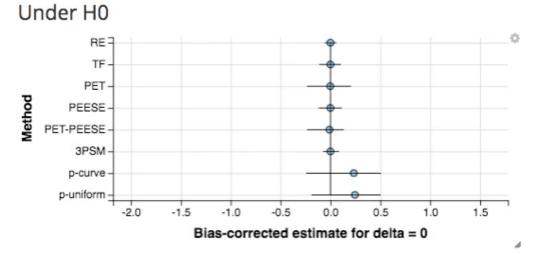
True effect size under H1 (for power computation)

○ 0.2 ○ 0.5 ○ 0.8

QRP environment:

none 🔿 med 🔿 high

Bias-corrected estimates of the true effect



17

Method performance check

- Hope that all bias-correcting methods will converge on the same value? Usually that does not happen
- → No vote counting no triangulation:
 - Even if three out of four methods converge on a value this is irrelevant, when those three are known to perform badly in plausible conditions.
- Use the app to see which bias-correcting methods perform well in plausible conditions for the metaanalysis at hand
- 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 distorts the evidence:
 Garbage in garbage out.
- Even meta-analyses of many dozen significant primary studies can come from a null effect.
- Each type of bias-correction works in some conditions, but fails in other conditions.
 Problem: We do not know which condition we are in.
- Doing biased research and hoping to correct it afterward *does not work*.
- Better put efforts into improving primary studies themselves (e.g., by using registered reports which combat both *p*-hacking and publication bias)

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/.

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 metaanalytic 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 methor address the simulated data that could have

is to help researchers in psychology know what to expect from different methods when conducting meta-analysis in the face of bias.

plausibly been produced by research in psychology. Our goal

Meta-analysis

Meta-analytic techniques involve synthesizing a set of results from studies investigating the same empirical phenomenon (Borenstein, Hedges, Higgins, & Rothstein, 2011). Most often, the results from the individual studies take the form of effect size estimates, and because meta-analyses are usually applied to studies with dependent variables measured on different scales, effect size estimates are typically standardized. The typical goal of a meta-analysis is to produce a single summary estimate of the hypothetical true underlying effect, δ_i estimated by each effect size in the dataset. This is usually called fixed-effect meta-analysis (Cooper, Hedges, & Valentine, 2009) and can be modeled as $d_i = \delta + e_i$, where d_i is the observed effect size for study *i* that differs from the true underlying effect, δ_i by some amount of sampling er-

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



 "Researchers should not expect to produce a conclusive, debateending result by conducting a meta-analysis on an existing literature"

• "Instead, we imagine meta-analyses may serve best to draw attention to the existing strengths and/or weaknesses in a literature and these results can then inspire a careful reexamination of methodology and theory followed by, if necessary, large-scale, preregistered replication efforts."