Bayesian Bias Correction: Critically evaluating sets of studies in the presence of publication bias

Alexander Etz

UC Irvine

Introduction

- When we read a published paper, we have two broadly related goals
- Goal 1: Evaluate the evidence in favor of a claim
- Goal 2: Estimate how big the reported effect could reasonably be

Introduction

- Solution (in principle): Compute a Bayes factor (goal 1), or compute the posterior distribution (goal 2)

Challenge

- Challenge: Literature is biased
- You only see a subset of all studies performed, because only the studies that "worked" make it past the gatekeepers (editors, reviewers)

Publication bias

- Statistical significance filter: Only studies with $p < .05$ get published
- Censored data, based on the stochastic outcome of the study

Publication bias

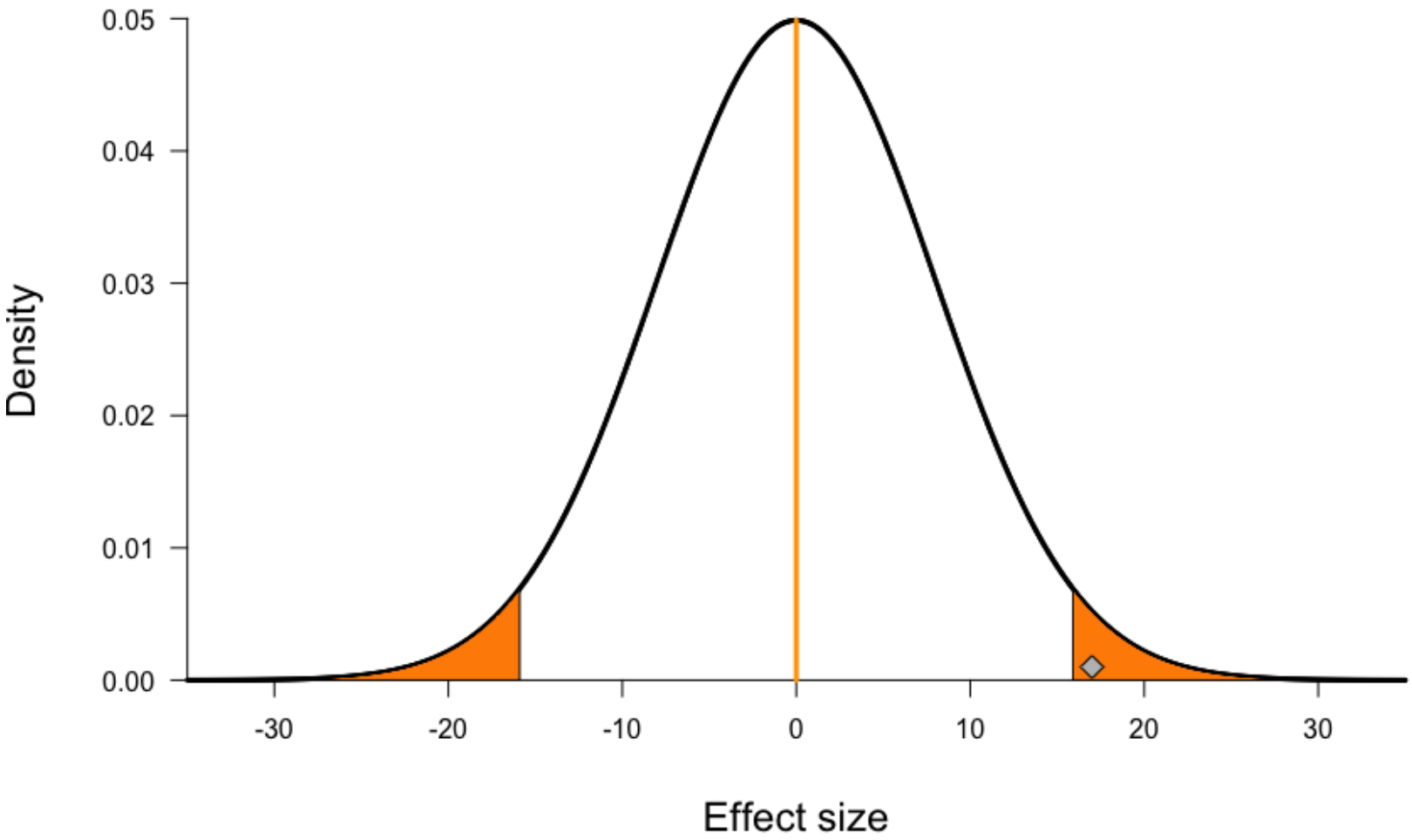
Publication bias ruins our ability to evaluate the claims in a paper

Example

- Imagine we estimate a group mean and perform a one-sample test
- We find observed mean of 17, with a standard error of 8

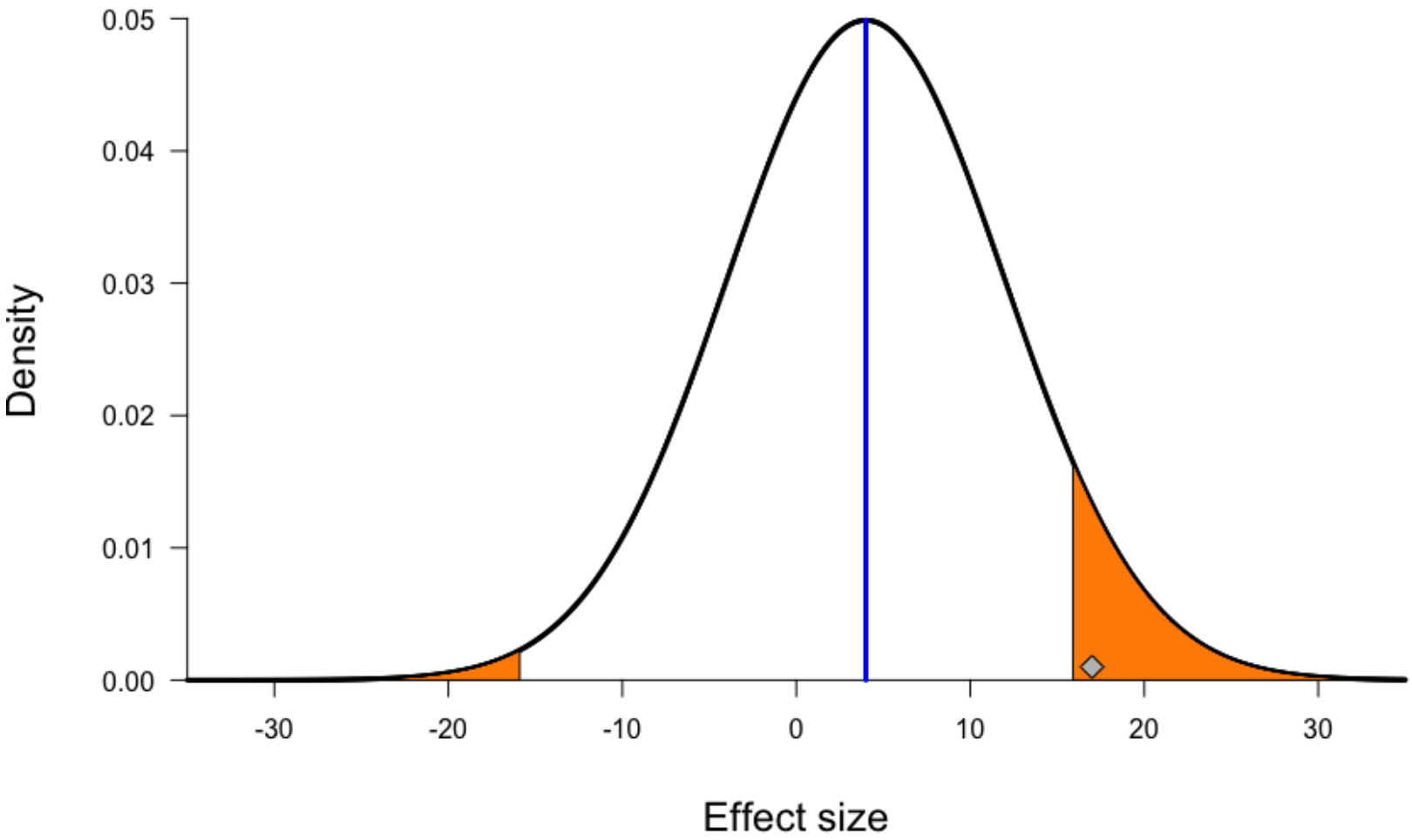
Example

If the null hypothesis is true, this would be what we predict:



Example

If the effect is small and positive, this is what we predict:



Implications

- What gets published if we only acknowledge significant results? Only studies with observed mean greater than 16
- If there is no effect, any published result is a type-1 error (i.e., false positive)
- If there is non-null effect, our effect size estimate will certainly be inflated
- Publication bias is clearly a problem for evaluating single studies, and its effects can only compound if we meta-analyze many studies at once

Modeling the bias

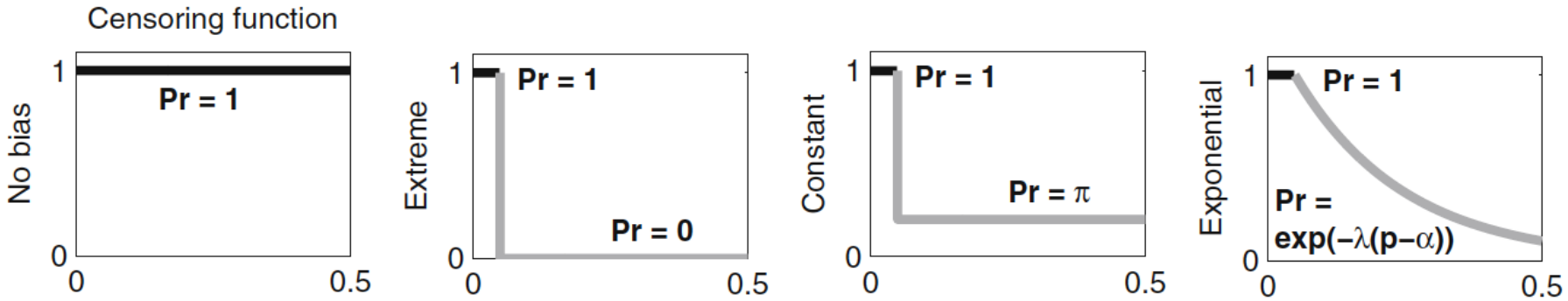
- We are modelers, so naturally we try to model this publication filtering process
- Try to patch up our inferences using Bayesian bias correction

Modeling the bias

- Key idea: Modify the sampling distribution to more accurately describe the results that will be making it to the literature

Modeling the bias

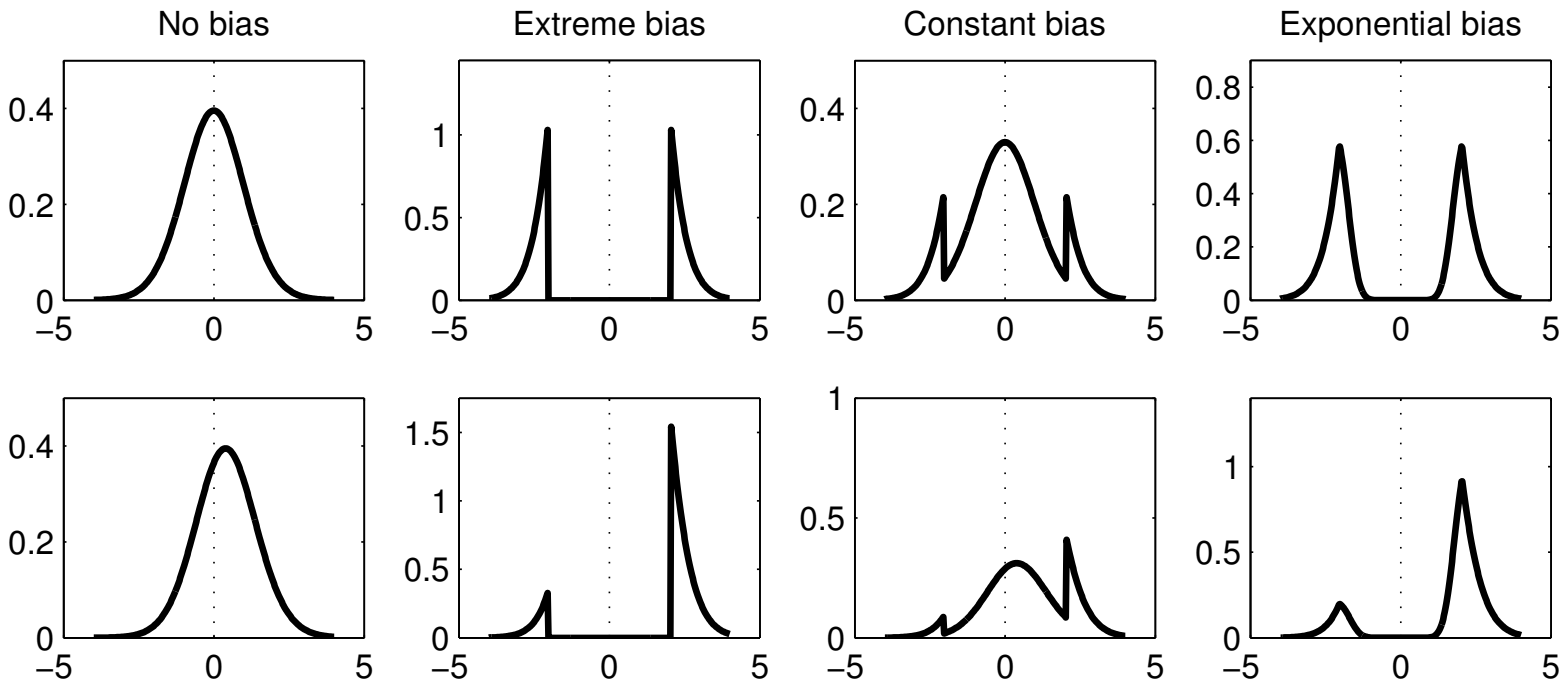
Guan and Vandekerckhove (2016) suggested the following plausible censoring processes...



(Note added: The x-axis is the p-value and y-axis is relative probability of publication. The x-axes are truncated at .5 merely for aesthetics)

Modeling the bias

... which make the following predictions about the t-values that make it to the literature



Modeling the bias

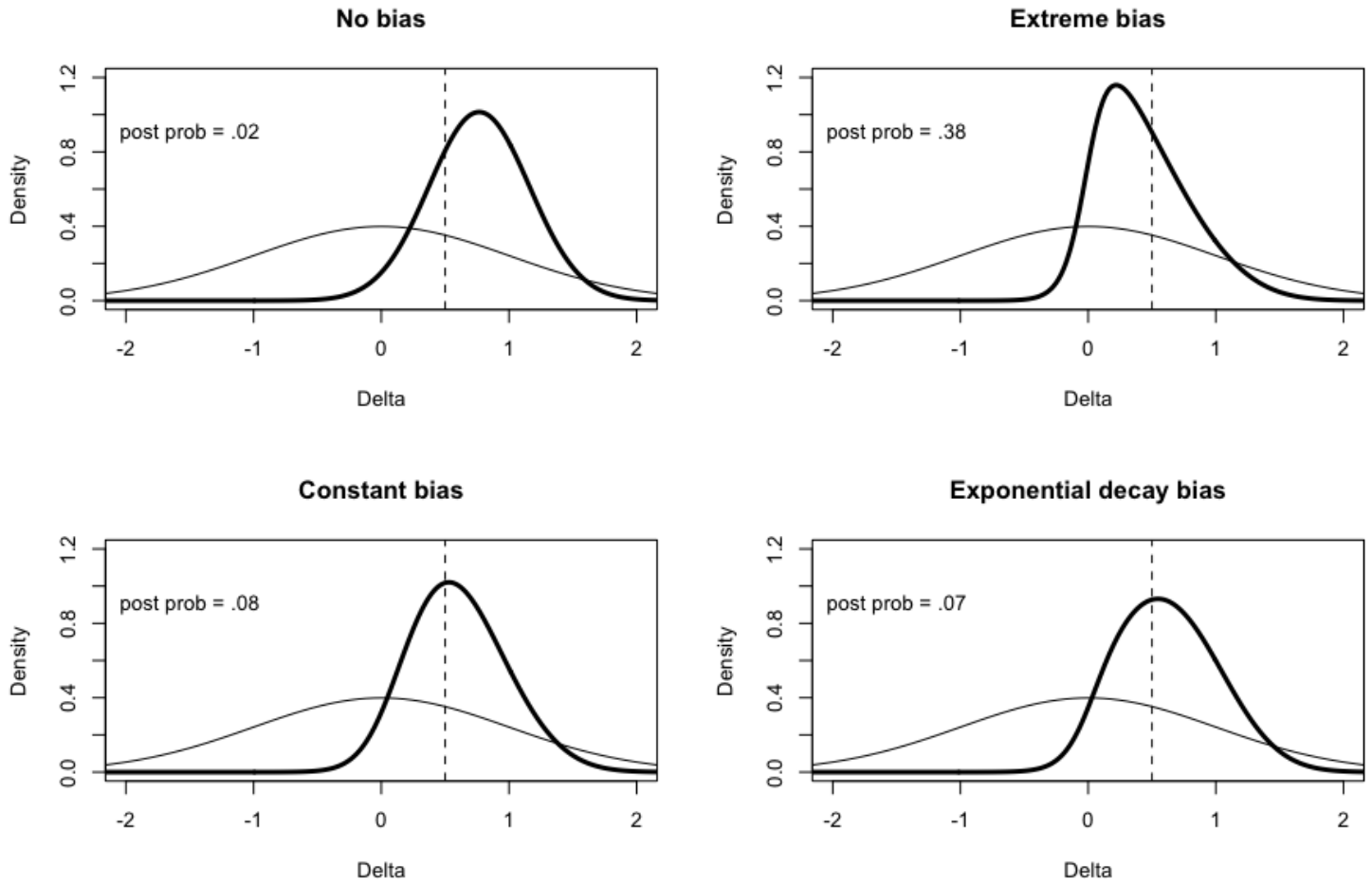
(Note added: The first model corresponds to the typical central and non-central t distributions, where all results are published. The second says n.s. results are never published, so all mass from the middle of the distribution is pushed to the tails (hence the name "extreme"). The third says all n.s. results are published at some rate π that does not depend on the p-value (hence the name, "constant"). The fourth says n.s. results with p-values close to .05 have higher publication rate than higher p-values.)

Estimating and testing

- Let's consider a single study that finds $t = 2.2$ with $n = 25$.
- Start with a prior on δ of $p(\delta|\mathcal{H}_1) = \mathcal{N}(0, 1)$
- Goal 1: Calculate Bayes factor in favor of nonzero effect size
- Goal 2: Compute posterior distribution for δ, $p(\delta|\text{data}, \mathcal{H}_1)$

Estimating and testing

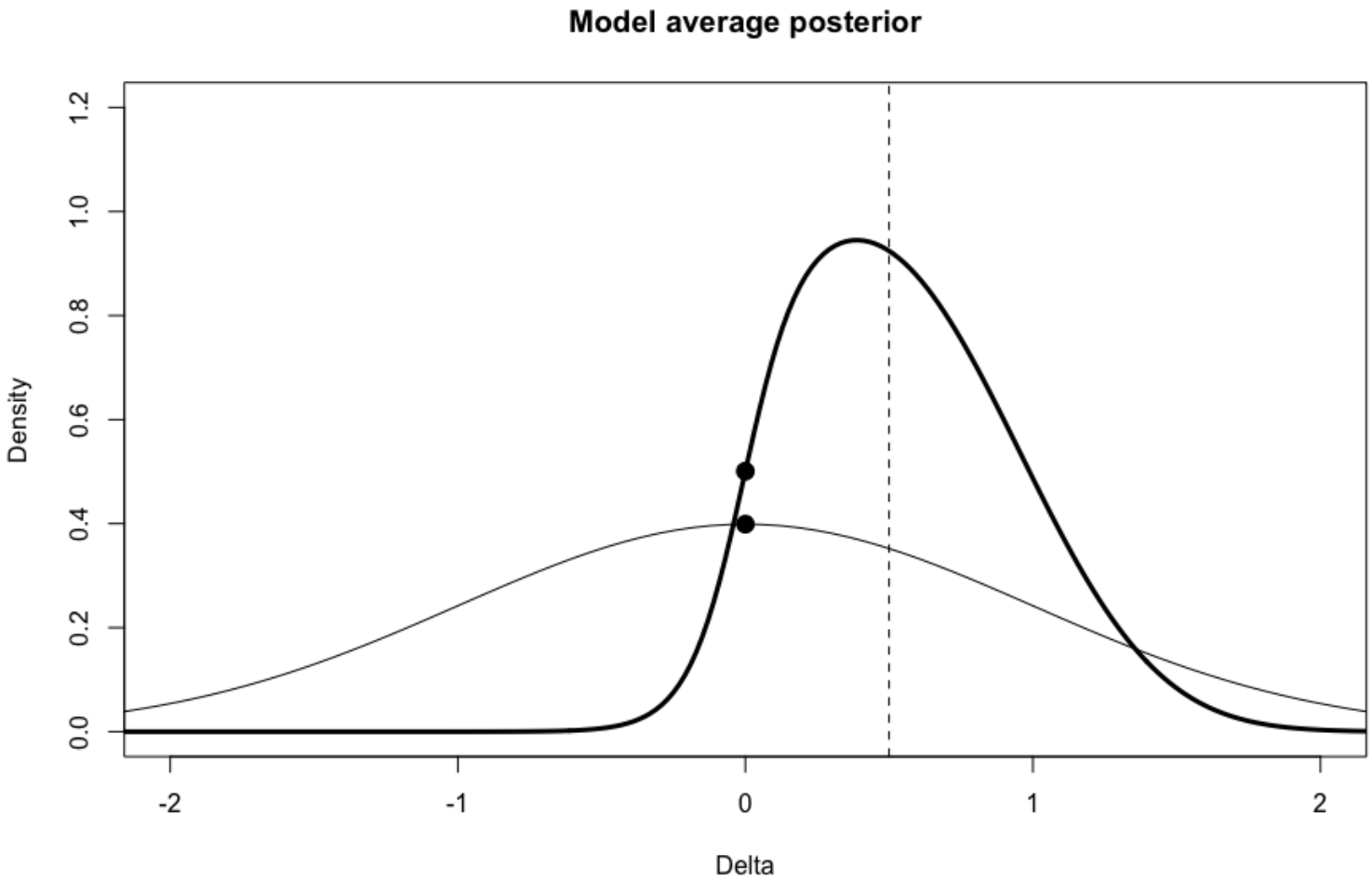
We can compute the posterior distribution[1] for each biasing model:



[1]Note: Displayed probs do not sum to 1 since we do not show the probabilities of the respective null models. A reference line is added at $\delta = .5$ to aid comparisons.

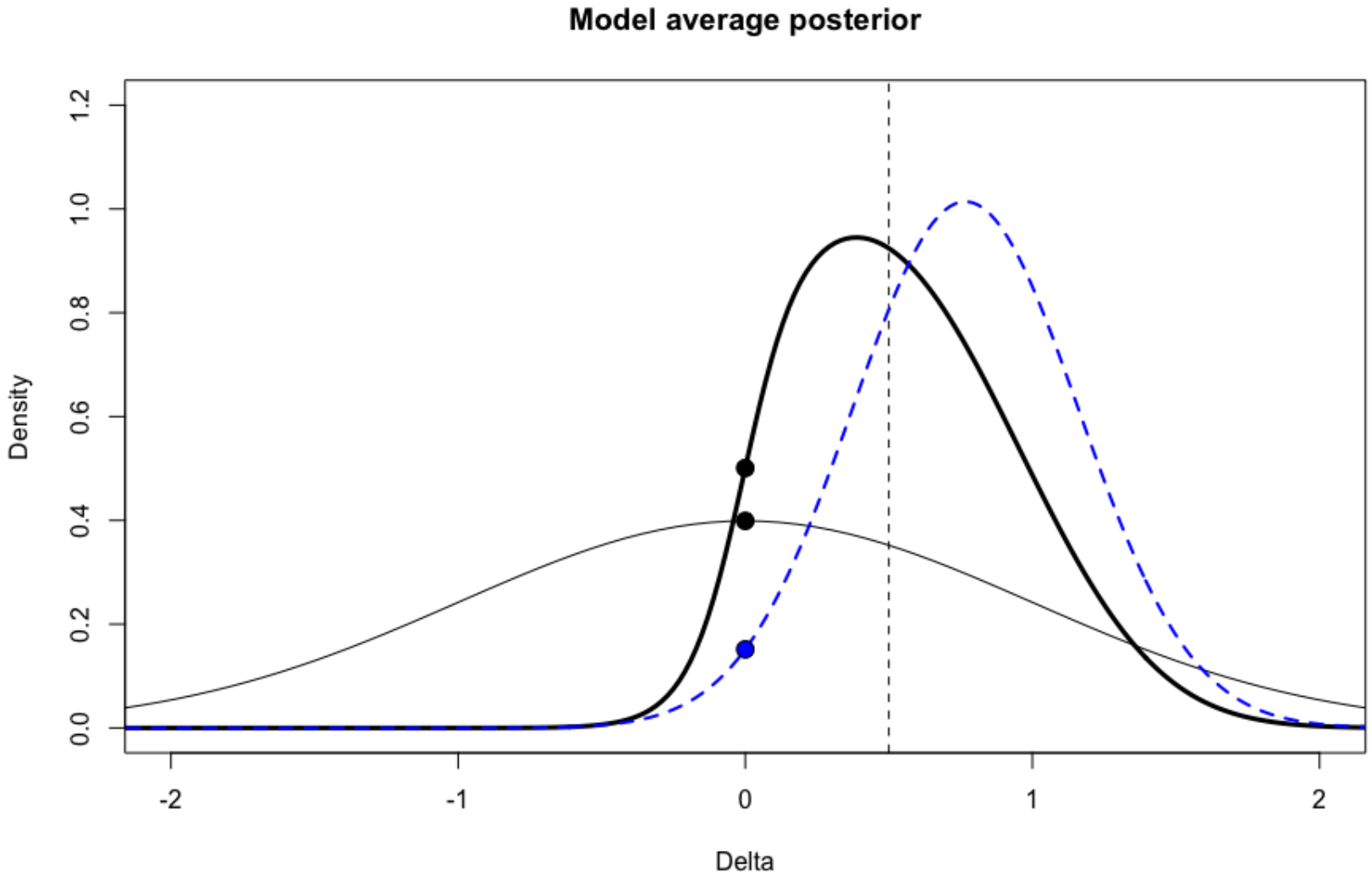
Estimating and testing

We can synthesize the posteriors according to their posterior model probabilities using *Bayesian model averaging* (the ratio of the heights of the black dots gives the model-averaged Bayes factor)



Estimating and testing

Compare to results from the naive posterior, assuming no bias (blue)



Estimating and testing

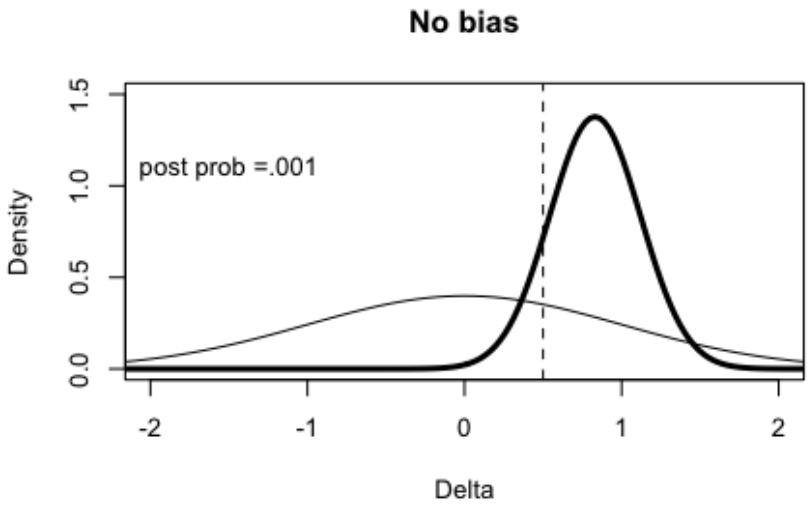
To summarize, if we observe $t = 2.2$ with $n = 25$:

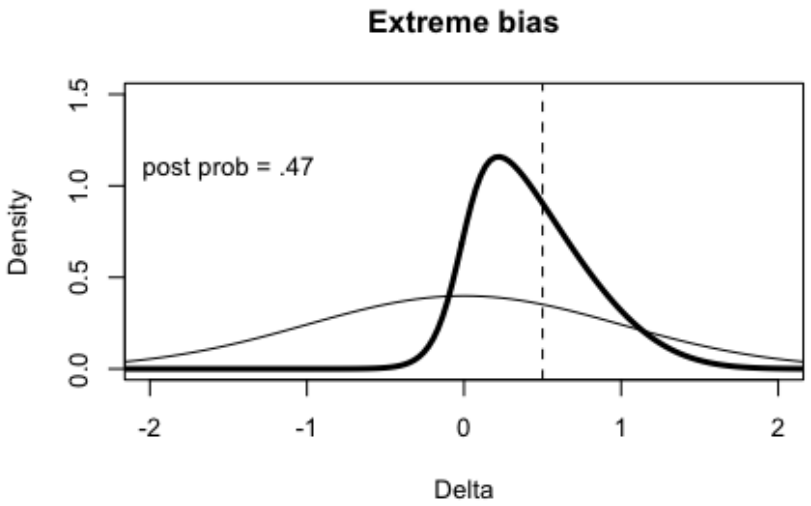
- The posterior distribution is shifted to smaller values
- The mitigated Bayes factor is $BF_{01} = 1.4$, where the original Bayes factor $BF_{10} = 2.5$ slightly favored the alternative. But neither are too conclusive

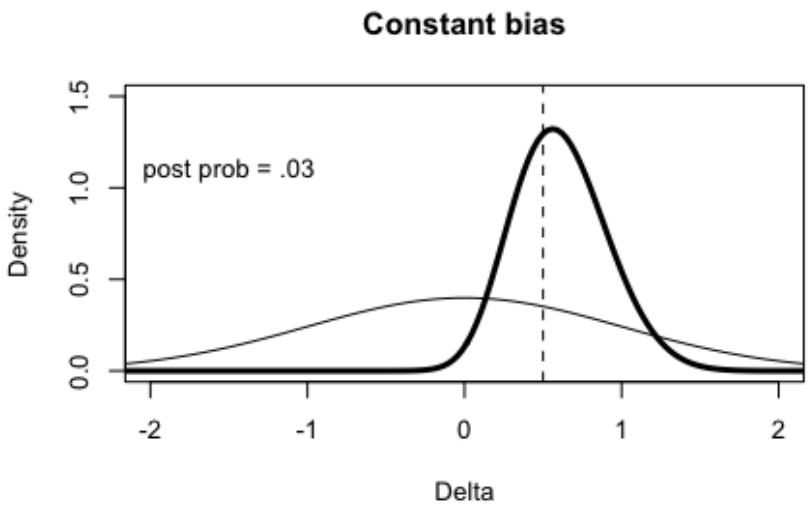
Estimating and testing

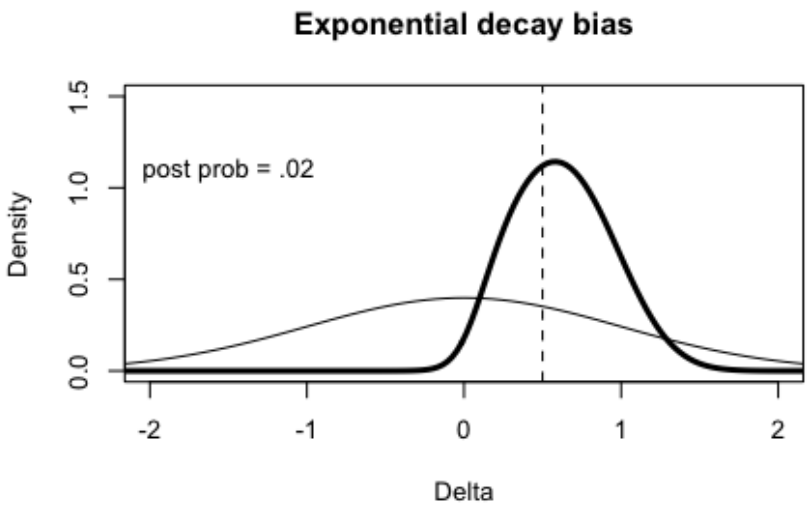
- Consider: The authors publish a followup study, with two new (direct) replications of the effect: In total we have $t = (2.2, 2.3, 2.1)$ and $n = (25, 20, 35)$
- Now we want to do a fixed-effects meta-analysis to synthesize the three sets of observations

Estimating and testing

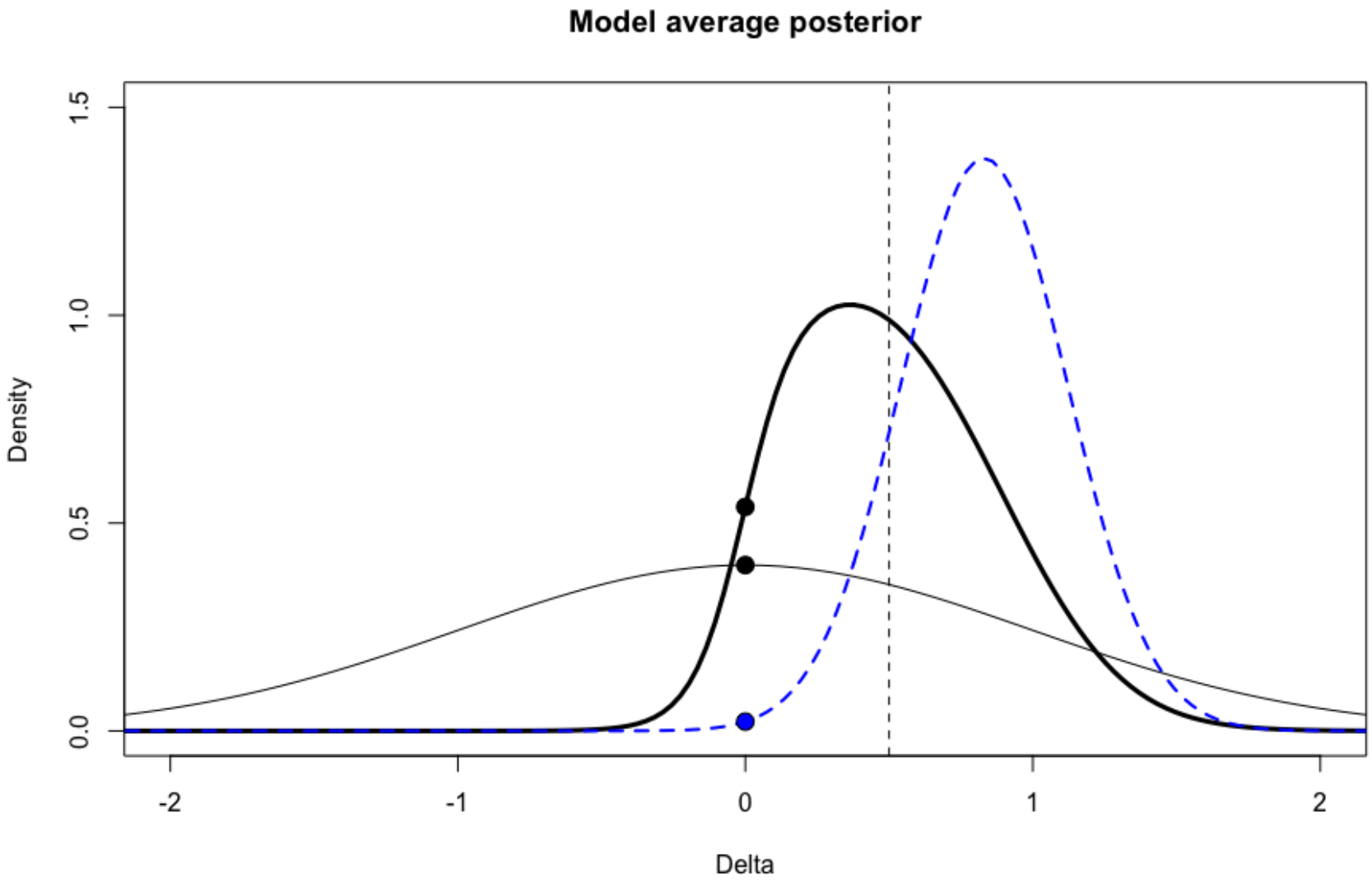








Estimating and testing



Estimating and testing

To summarize, for $t = (2.2, 2.3, 2.1)$ and $n = (25, 20, 35)$:

- The cumulative posterior distribution is largely shifted to smaller values
- The mitigated Bayes factor is $BF_{01} = 1.2$, again indecisive, whereas the original Bayes factor $BF_{10} = 17$ largely favored the alternative

Extending to random effects

- Previously, the method could only apply to single studies or fixed effects meta-analysis
- We have extended it to apply to random effects meta-analysis, by introducing study-specific effect sizes drawn from a higher level population
- Allows for there to be heterogeneity in true effects across studies (e.g., conceptual replications with new measures, replications on different populations, etc.)

We now demonstrate this extension on some recently published data

The studies

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

Interventions aimed at influencing spending behavior and risk-taking have considerable practical importance. A number of studies motivated by the costly signaling theory within evolutionary psychology have reported that priming inductions (such as looking at pictures of attractive opposite sex members) designed to trigger mating motives increase males' stated willingness to purchase conspicuous consumption items and to engage in risk-taking behaviors, and reduce loss aversion. However, a meta-analysis of this literature reveals strong evidence of either publication bias or *p*-hacking (or both). We then report 8 studies with a total sample of over 1,600 participants which sought to reproduce these effects. None of the studies, including one that was fully preregistered, was successful. The results question the claim that romantic primes can influence risk-taking and other potentially harmful behaviors.

Keywords: risk, consumer behavior, decision making, priming, meta-analysis

Supplemental materials: http://dx.doi.org/10.1037/xge0000116.supp

The studies

- "Recently, it has been claimed that risk-taking and spending behavior may be triggered in part by evolutionarily driven motives"
- "Sexual cues in advertising product categories such as casinos, fashion, jewelry, cosmetic surgery, cars, cigarettes, and alcohol, and the evidence for their effectiveness, suggests that controlling such primes in real-world settings might constitute a valuable intervention."

The authors conduct a meta-analysis of the published literature, and subsequently perform various replications of selected studies

The studies

Table 1
Studies of the Influence of Mating Primes on Various Aspects of Decision Making

	Prime method		
Decision-making domain	Opposite-sex pictures	Romantic text	Other formats
Conspicuous consumption	Griskevicius et al. (2007, Study 1) Sundie et al. (2011, Study 1) **Studies 4 and 5**	Griskevicius et al. (2007, Studies 2 and 3) Sundie et al. (2011, Studies 2 and 3) **Studies 1–3**	Chan (2015) Festjens et al. (2014, Study 3)
Benevolence	Griskevicius et al. (2007, Study 1)	Griskevicius et al. (2007, Studies 2–4) **Study 3**	
Gambling	Baker and Maner (2008) Greitemeyer et al. (2013, Experiment 2) Li (2012) McAlvanah (2009) **Studies 6 and 7b**		
Social risk-taking	**Studies 5 and 6**		
Sexual/health risk-taking	Greitemeyer et al. (2013, Experiment 1) Hill and Durante (2011, Study 1) **Studies 6 and 7a**	Hill and Durante (2011, Study 2)	
Driving risk-taking	Greitemeyer et al. (2013, Experiment 3) **Study 5**		
Substance risk-taking	**Study 6**		
Physical risk-taking			Baker and Maner (2009)
Loss aversion		Li et al. (2012, Studies 1–3) **Study 8**	Festjens et al. (2014, Study 2)
Temporal discounting	Kim and Zauberman (2013) Van den Bergh et al. (2008, Study 1A) Wilson and Daly (2004)		Festjens et al. (2014, Study 1) Van den Bergh et al. (2008, Study 1B)
Cooperation (ultimatum game)	Van den Bergh and Dewitte (2006)		

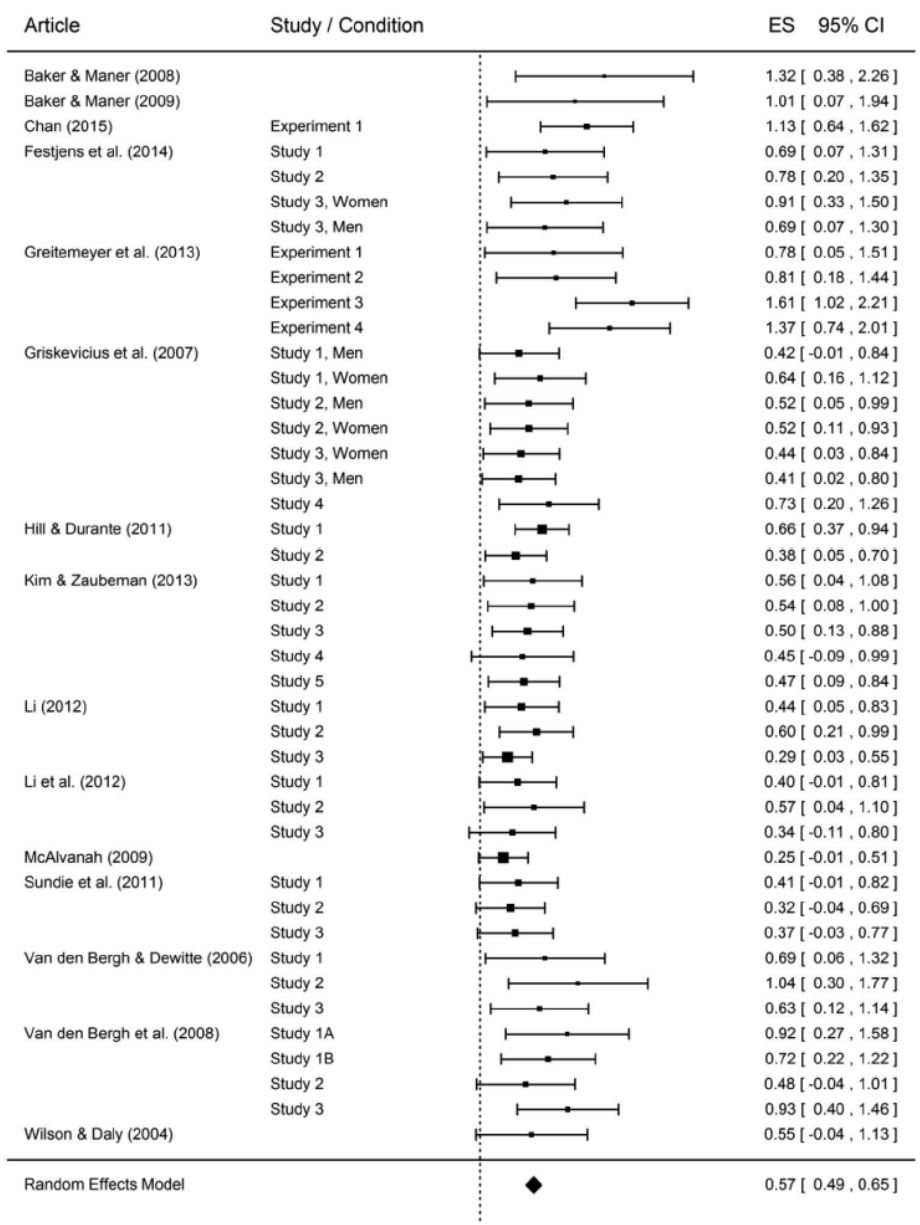
Note. Bold indicates studies reported in this article.

The studies

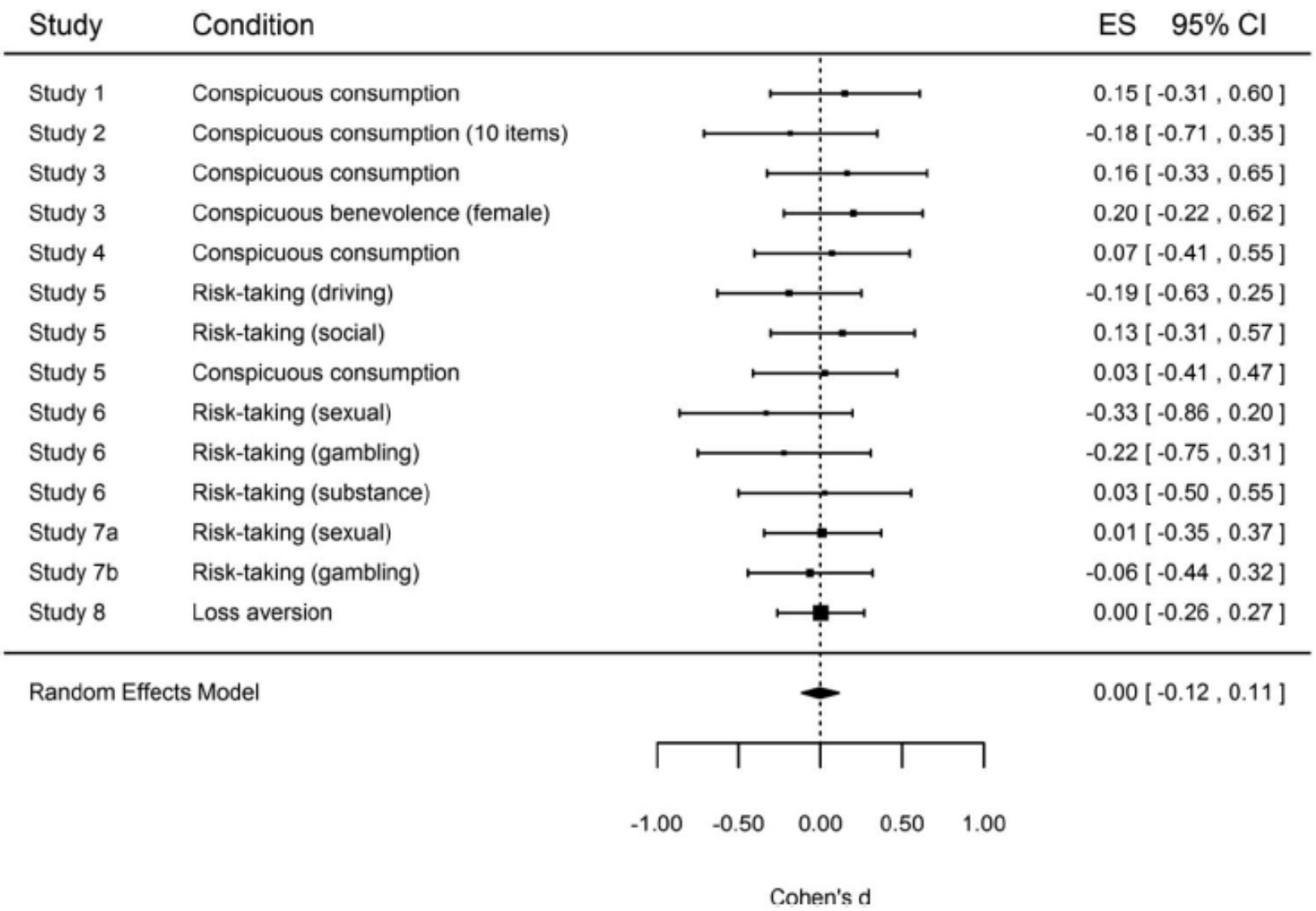
Various designs, manipulations, outcomes, all call for a random effects meta-analysis that allows each study its own effect size:

- Can evaluate each study's individual effect size
- Can evaluate overall population mean
- Can evaluate variability across studies

The published literature



The replications



Synthesis?

How can we synthesize these two sets of results?

- Option: Throw them all together in a joint meta-analysis?
- What about publication bias?
- Option: Throw out the published literature, and only look at the replications?
- Isn't this wasteful? Presumably there is at least some signal in there we can extract

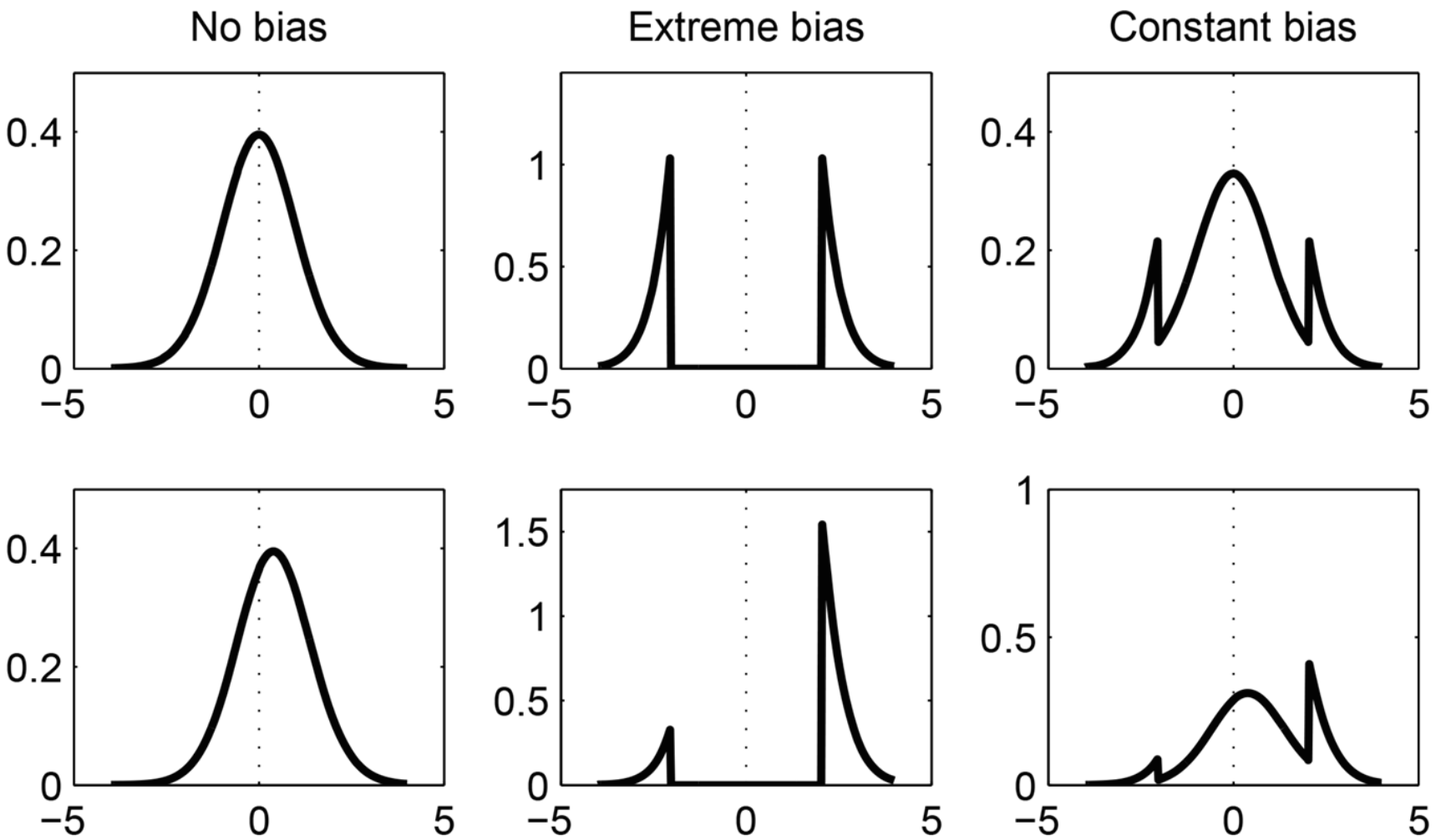
We attempt to jointly analyze all studies using our Bayesian bias correction method

The model

Random effects meta-analysis can be instantiated as a Bayesian hierarchical model, but we would like to modify it to account for bias in the published literature

The model

We can estimate the effect sizes using model 3, of which 1 and 2 are special cases...



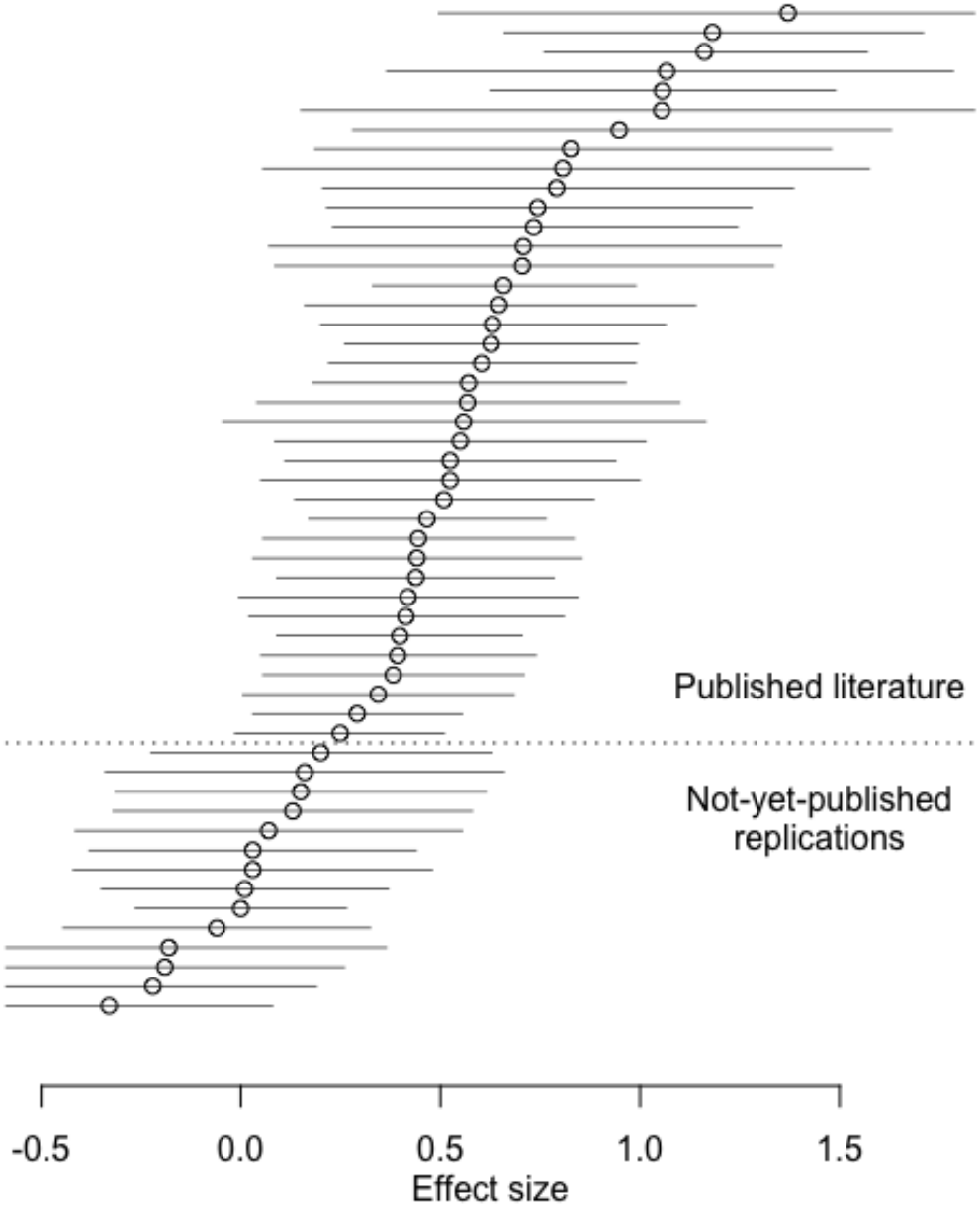
The model

...with an added hierarchical structure:

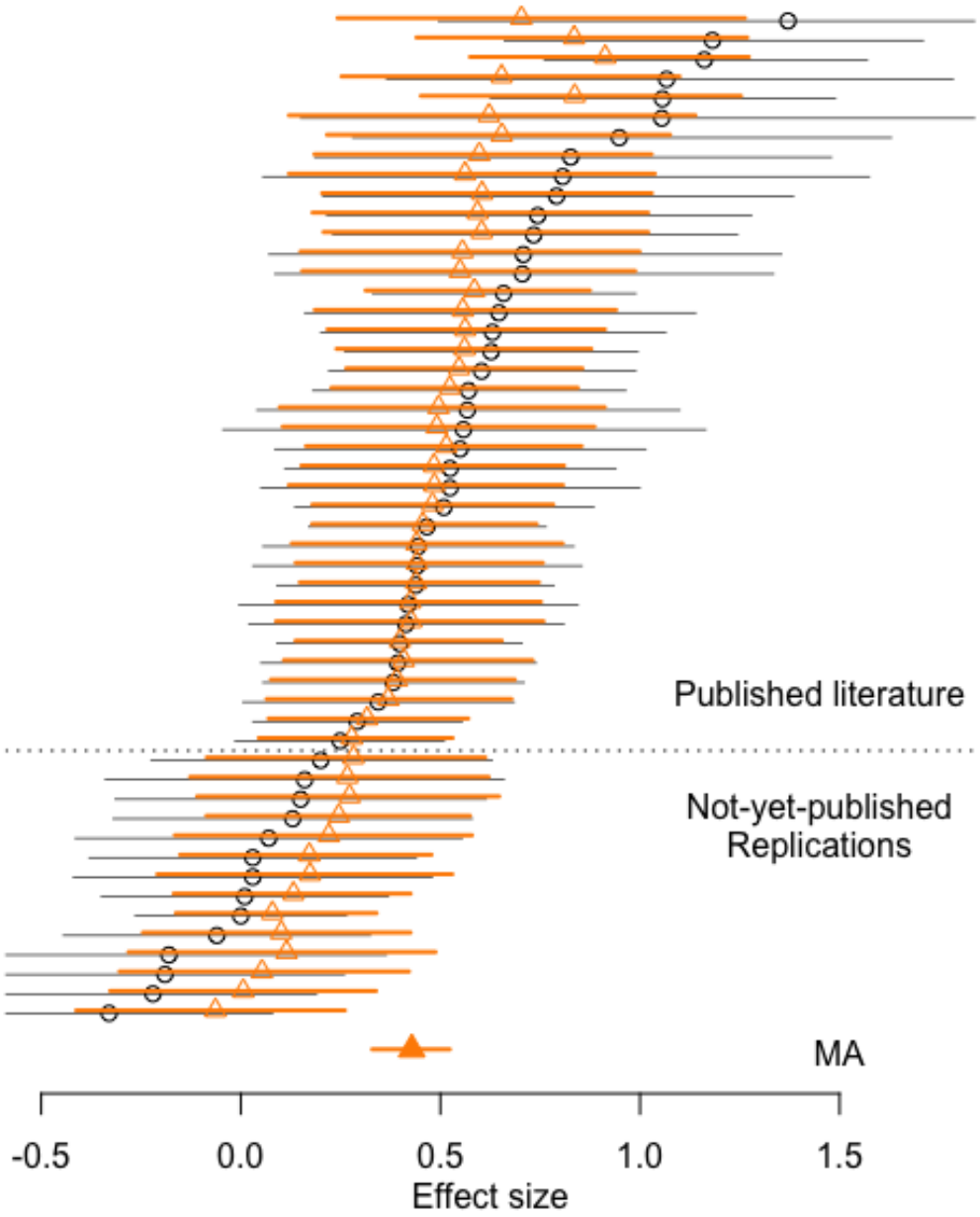
$$
\begin{aligned}
\mu &\sim \mathcal{N}(0, 1) \\
\sigma &\sim \mathcal{G}(4, 4) \\
\pi &\sim \mathcal{B}(5, 95) \\
\delta_i \mid \mu, \sigma &\sim \mathcal{N}(\mu, \sigma^2) \\
t_i \mid \delta_i, U_i = 0 &\sim \text{biased-}t_{\nu,\pi}(\text{ncp} = \delta_i \sqrt{\nu_i}/2) \\
t_i \mid \delta_i, U_i = 1 &\sim t_{\nu}(\text{ncp} = \delta_i \sqrt{\nu_i}/2)
\end{aligned}
$$

where δ_i is the study-specific effect size for study i, U_i indicates whether study i has not been through a publication filter. These priors, like always, are up for debate, but we think they are reasonable

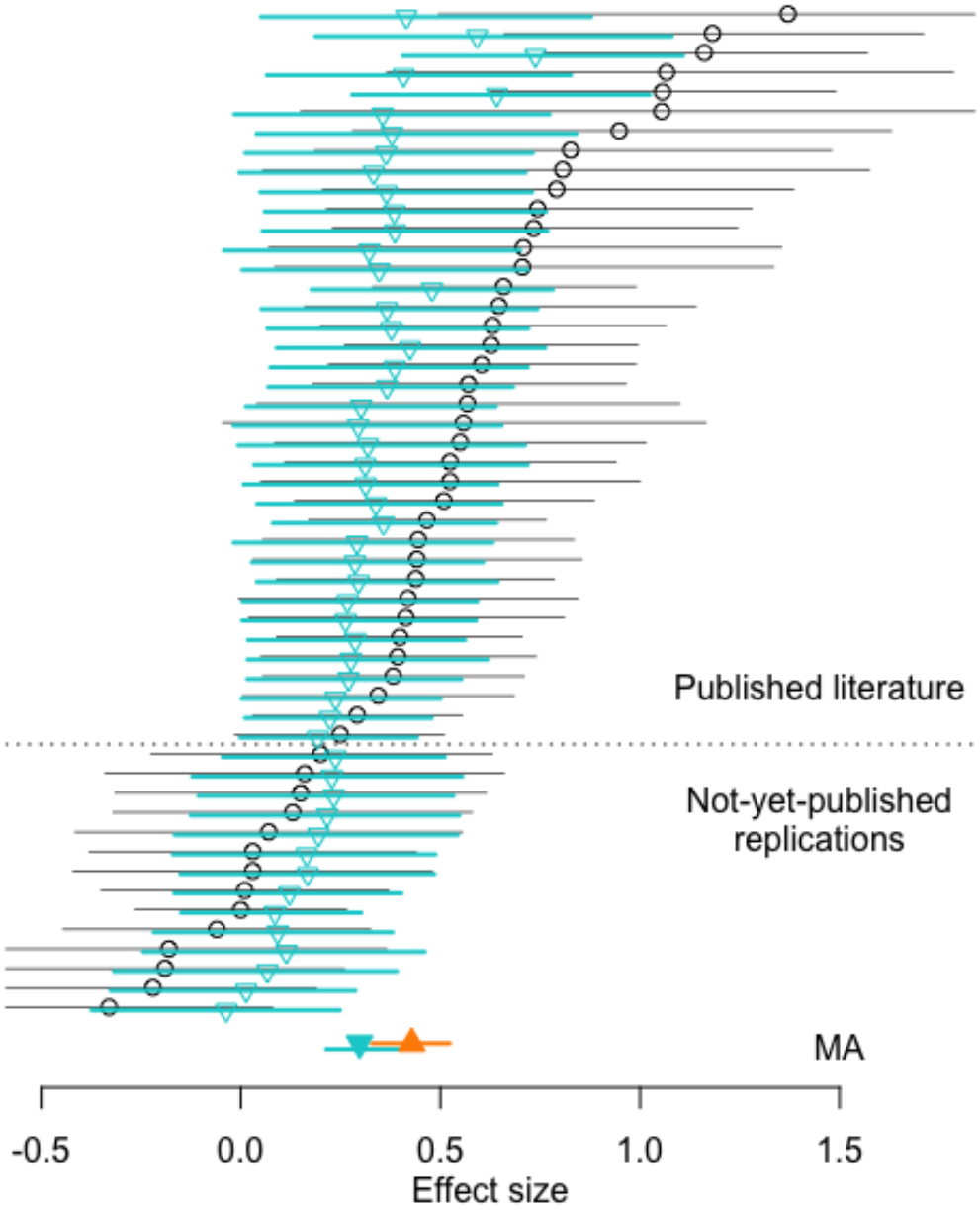
The data, sorted by ascending ES



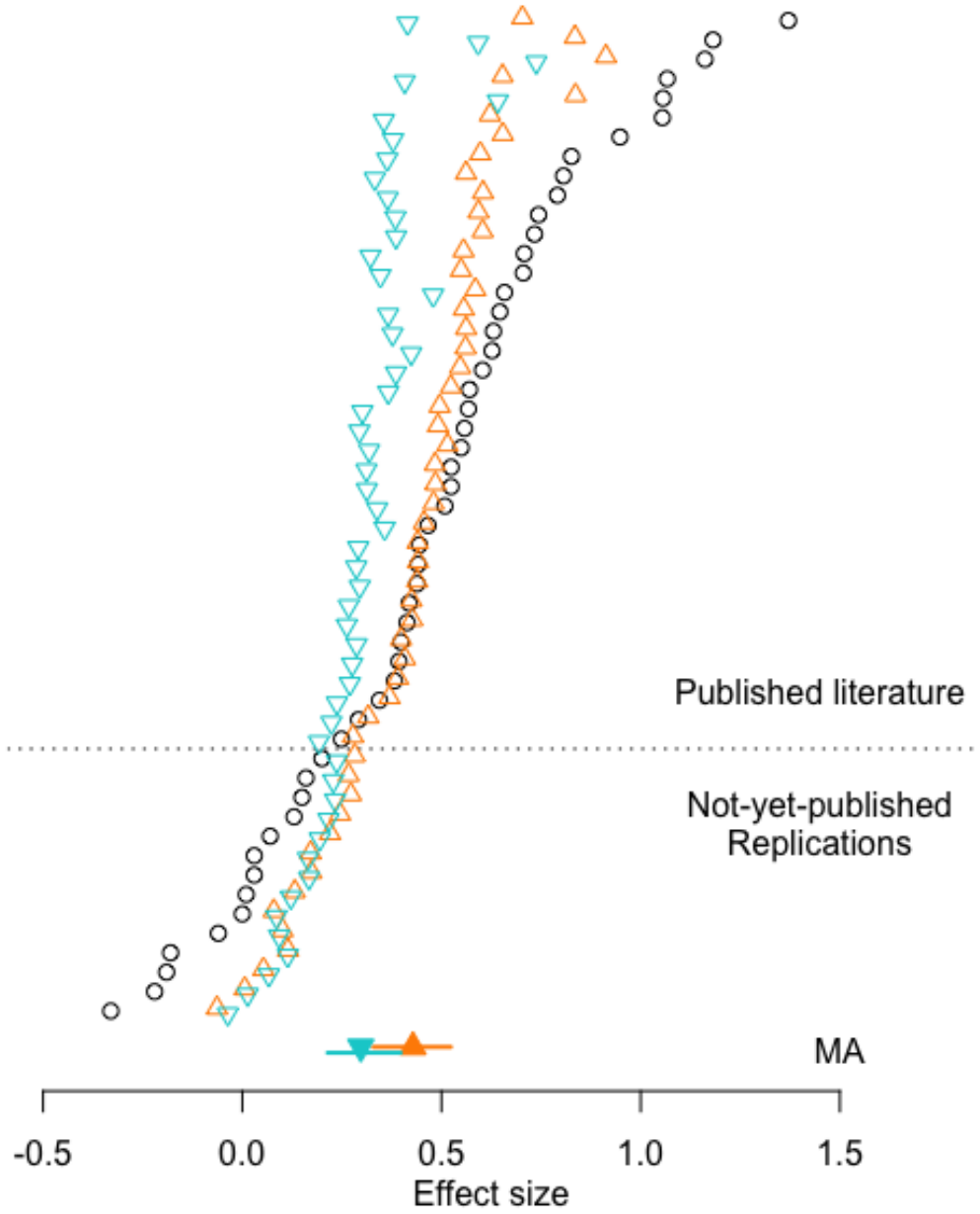
Naive hierarchical model



Debiasing hierarchical model



Comparing the model estimates



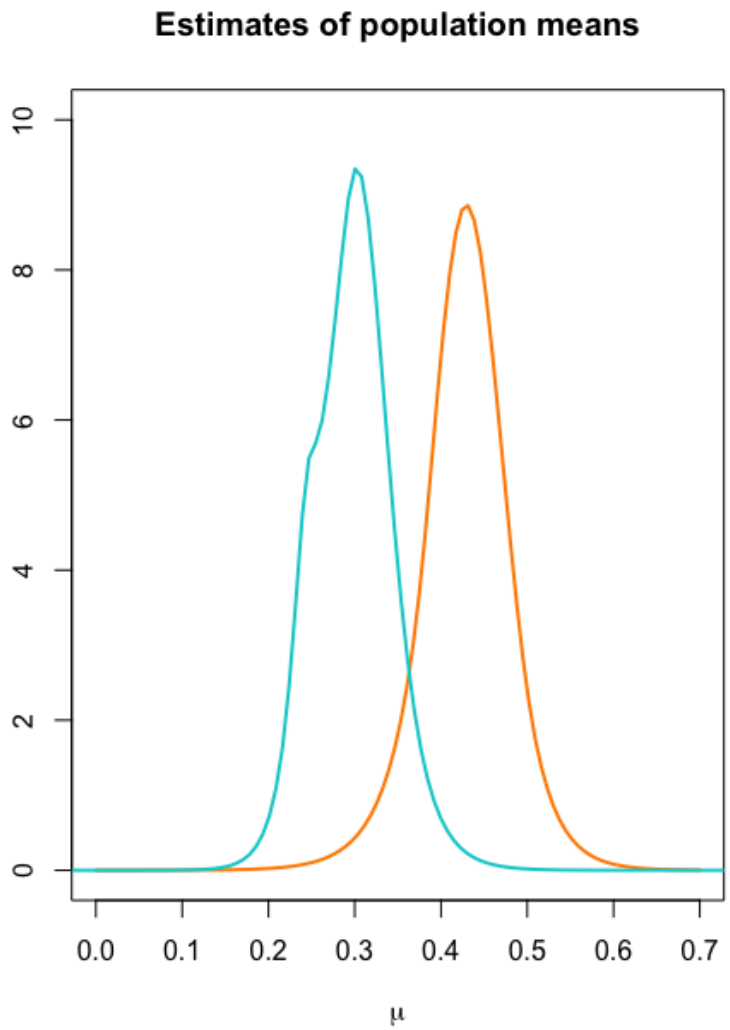
Summary

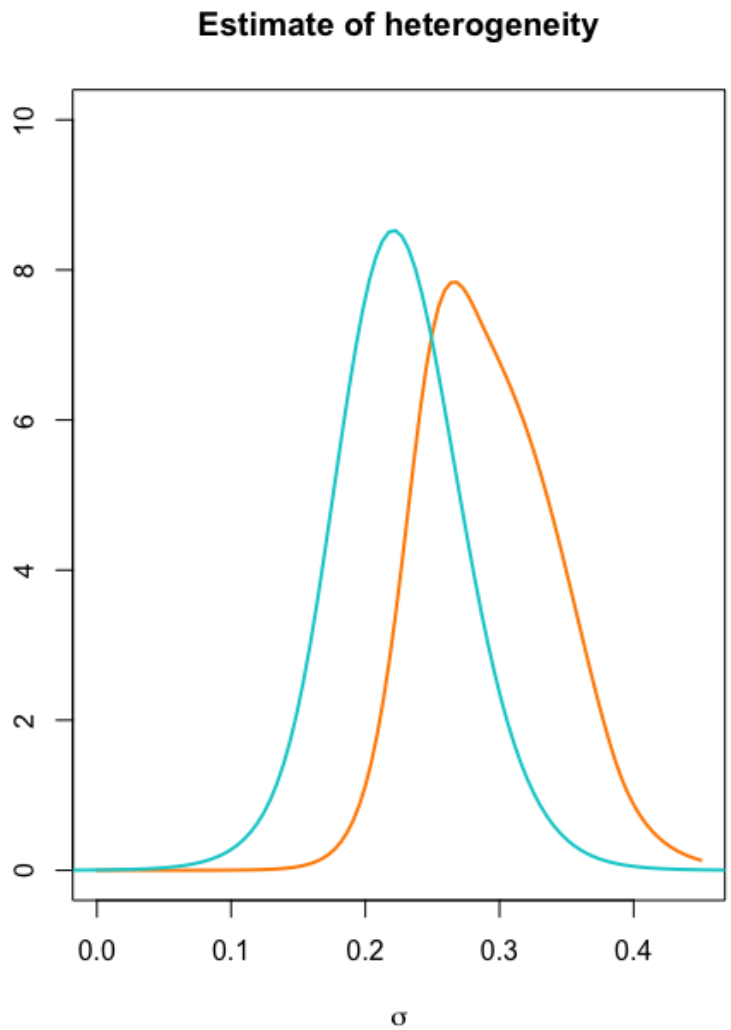
In both models we see the shrinkage typical of Bayesian methods

- Large effects are pulled down, small effects are pulled up, both "shrinking" toward the overall mean
- The main difference between the naive hierarchical model and our debiasing model: Extra shrinkage for published effects!

More plots

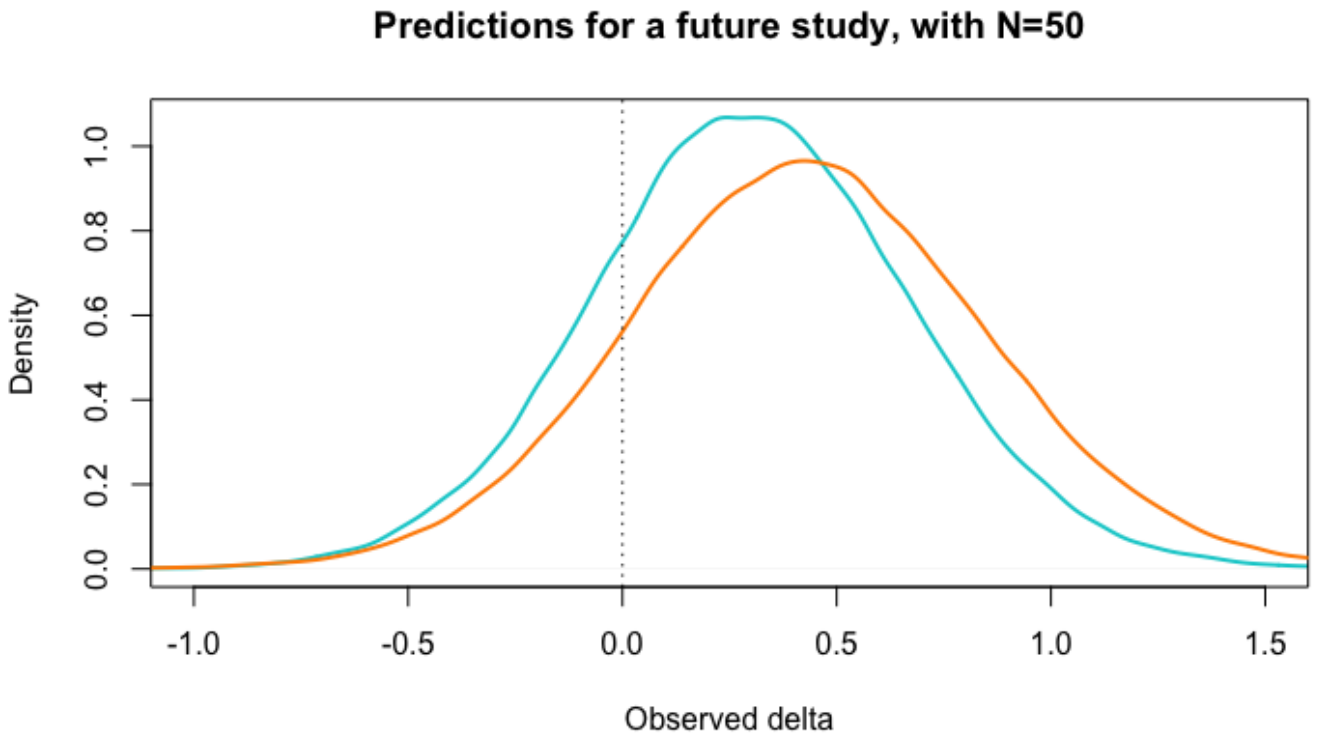
We summarize our results as follows, showing our best estimates of the mean and sd for the distribution from which we draw each study's δ





Model predictions

We can generate predictions from the models (i.e., posterior predictives) for what we should expect for a new romance priming study with N=50



For the naive model, the predictive has a mean of .43 and sd of .42. The bias correction model predictive has mean of .30 and sd of .38.

Acknowledgements

Funding for this project is provided by the NSF Graduate Research Fellowship Program, as well as a grant from the NSF's Methods, Measurements, and Statistics panel.