Correcting for publication bias in a meta-analysis

Robbie C.M. van Aert & Marcel A.L.M. van Assen

Tilburg University

September 24, 2018

Message

- Consequences of publication bias are horrible for science
- Publication bias $\rightarrow$ overestimation of effect size in meta-analysis
- The publication bias method *p*-uniform overestimates effect size in case of between-study variance in true effect size
- The improved and extended method *p*-uniform*:
  1. eliminates overestimation due to between-study variance
  2. is a more efficient estimator than *p*-uniform's estimator
  3. enables estimating and testing of the between-study variance

Overview

1. Publication bias
2. From *p*-uniform to *p*-uniform*
3. Selection model approach
4. Analytical study
5. Monte-Carlo simulation study
6. Conclusion and discussion

Publication bias

- Publication bias is “the selective publication of studies with a significant outcome”
- Longer history in dealing with publication bias in medical research than social sciences
- Nowadays, increased attention for publication bias in various fields
- Evidence for publication bias in various research fields

Publication bias: Evidence

- Coursol and Wagner (1986) surveyed researchers on the effects of positive findings

Table 1

Relation Between Outcome (Positive vs. Neutral or Negative) and Decision to Submit Research for Publication

Direction of outcome	Submission decision: Yes	Submission decision: No	Total
Positive (Client improved)	106	23	129
Neutral or negative (Client did not improve)	28	37	65
Total	134	60	194

Publication bias: Evidence

- Coursol and Wagner (1986) surveyed researchers on the effects of positive findings

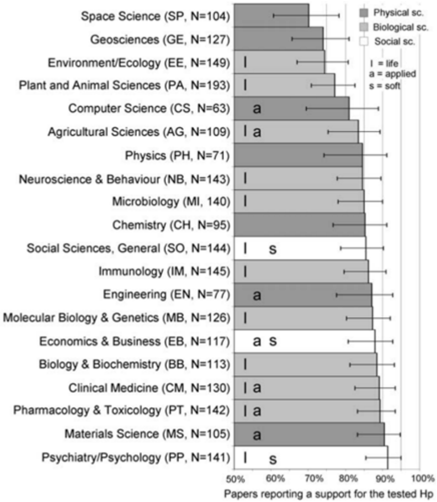
Table 2

Relation Between Outcome (Positive vs. Neutral or Negative) and Acceptance of Research Submitted for Publication

Direction of outcome	Accepted	Not accepted	Total
Positive (Client improved)	85	21	106
Neutral or negative (Client did not improve)	14	14	28
Total	99	35	134

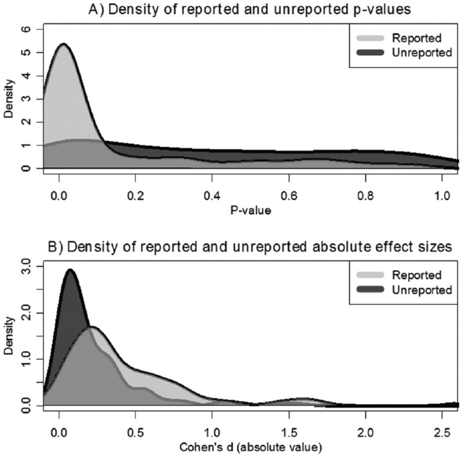
Publication bias: Evidence

- Fanelli (2010) scored for published articles whether there was positive or negative support for studied hypothesis
- 90% of hypotheses are significant in psychology
- However, this is not in line with average statistical power (about 20-50%)



Publication bias: Evidence

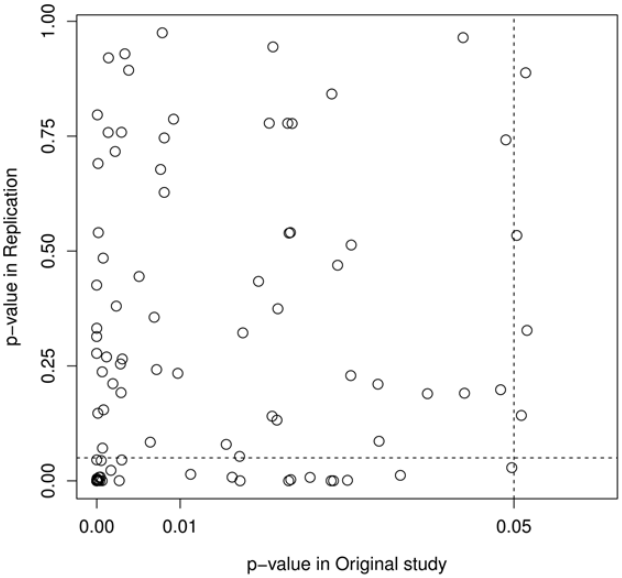
- Franco et al. (2016) studied publication bias by redoing analyses planned in grant proposals
- Comparing reported results in articles with unreported results
- Difference between reported and unreported *p*-values and effect size



Publication bias: Evidence

- Open Science Collaboration initiated Reproducibility Project which was a large-scale replication attempt of psychological research
- 100 studies were replicated from three flagship journals: JPSP, Psychological Science, and Journal of Experimental Psychology
- Results shocked many people inside and outside academia:
  - 97% of original studies were significant and only 36% of replications
  - Effect size estimates decreased from r=0.4 to 0.2

Publication bias: Evidence



Publication bias: Evidence

- Experimental economics: 89% of original studies were significant and 69% of replications
- Hematology and oncology: 11% of studies were deemed to be successfully replicated

Publication bias: Evidence

- Experimental economics: 89% of original studies were significant and 69% of replications
- Hematology and oncology: 11% of studies were deemed to be successfully replicated
- Substantial amount of critique on these projects

Publication bias: Evidence

- Experimental economics: 89% of original studies were significant and 69% of replications
- Hematology and oncology: 11% of studies were deemed to be successfully replicated
- Substantial amount of critique on these projects
- Two plausible causes of this low replicability:
  - Publication bias
  - Questionable research practices

Publication bias: Consequences

- Effects of publication bias are horrible:
  - False impression that effect exists (false positives)
  - Overestimation of effect size
  - Questionable research practices

From p-uniform to p-uniform*: p-uniform

- Only considers significant effect sizes and discards others
- Distribution of p-values at the true effect size is uniform
- Only significant effect sizes, so conditional probabilities:

$$q_i = \frac{1 - \Phi\left(\frac{y_i - \mu}{\sigma_i}\right)}{1 - \Phi\left(\frac{y_{cv} - \mu}{\sigma_i}\right)}$$

- Tests for uniformity are used to evaluate whether q_i are uniformly distributed

From *p*-uniform to *p*-uniform*: *p*-uniform

- Only considers significant effect sizes and discards others
- Distribution of *p*-values at the true effect size is uniform
- Only significant effect sizes, so conditional probabilities:

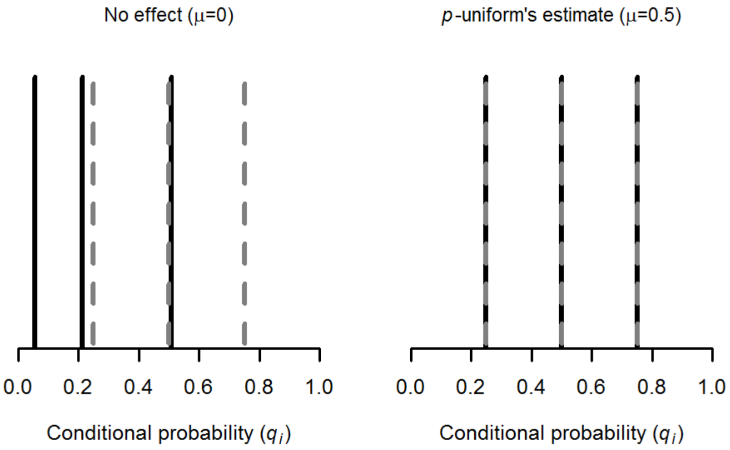
$$q_i = \frac{1 - \Phi\left(\frac{y_i - \mu}{\sigma_i}\right)}{1 - \Phi\left(\frac{y_{cv} - \mu}{\sigma_i}\right)}$$

- Tests for uniformity are used to evaluate whether q_i are uniformly distributed
- Assumptions:
  - Homogeneous true effect size
  - All significant effect sizes have an equal probability of getting included in a meta-analysis

From *p*-uniform to *p*-uniform*: *p*-uniform

- Example with three observed effect sizes ($\mu = 0.5$):

$t(48)=3.133$, $p=.0029$; $t(48)=2.646$, $p=.011$; $t(48)=2.302$, $p=.025$



From *p*-uniform to *p*-uniform*: *p*-uniform

- McShane et al. (2016) criticized *p*-uniform for three reasons:
  1. Assumption of homogeneous true effect size

From *p*-uniform to *p*-uniform*: *p*-uniform

- *P*-uniform is positively biased if true effect size is heterogeneous (van Aert et al., 2016)

- Simulation with extreme publication bias and $\mu = 0.397$

	No	Moderate	Large	Larger	Very large
p-uniform	0.387	0.522	0.679	0.776	0.903
FE MA	0.553	0.616	0.738	0.875	1.104
RE MA	0.553	0.616	0.743	0.897	1.185

From *p*-uniform to *p*-uniform*: *p*-uniform

- *P*-uniform is positively biased if true effect size is heterogeneous (van Aert et al., 2016)

- Simulation with extreme publication bias and $\mu = 0.397$

	No	Moderate	Large	Larger	Very large
p-uniform	0.387	0.522	0.679	0.776	0.903
FE MA	0.553	0.616	0.738	0.875	1.104
RE MA	0.553	0.616	0.743	0.897	1.185

From *p*-uniform to *p*-uniform*: *p*-uniform

- *P*-uniform is positively biased if true effect size is heterogeneous (van Aert et al., 2016)

- Simulation with extreme publication bias and $\mu = 0.397$

	No	Moderate	Large	Larger	Very large
p-uniform	0.387	0.522	0.679	0.776	0.903
FE MA	0.553	0.616	0.738	0.875	1.104
RE MA	0.553	0.616	0.743	0.897	1.185

- Recommendations:
  - At most moderate: interpret as average *true* effect size
  - More than moderate: interpret as estimate of only significant effect sizes included in meta-analysis
  - If possible, create homogeneous subgroups of effect sizes

From *p*-uniform to *p*-uniform*: *p*-uniform

- McShane et al. (2016) criticized *p*-uniform for three reasons:

  1. Assumption of homogeneous true effect size
  2. Not an efficient estimator

From *p*-uniform to *p*-uniform*: *p*-uniform

- McShane et al. (2016) criticized *p*-uniform for three reasons:

  1. Assumption of homogeneous true effect size
  2. Not an efficient estimator
  3. *P*-uniform uses method of moments rather than maximum likelihood estimation

From p-uniform to p-uniform*: p-uniform

- McShane et al. (2016) criticized p-uniform for three reasons:

  1. Assumption of homogeneous true effect size
  2. Not an efficient estimator
  3. P-uniform uses method of moments rather than maximum likelihood estimation

- Hence, we improved p-uniform (called p-uniform*) such that:

  1. True effect size can be hetergeneous and overestimation caused by it is eliminated
  2. Nonsignificant effect sizes are incorporated $\rightarrow$ more efficient estimator
  3. Maximum likelihood estimation is implemented

From p-uniform to p-uniform*: p-uniform*

- *P*-uniform* considers the significant **and** nonsignificant effect sizes
- Now effect sizes not only conditional on significance but also on nonsignificance
- Maximum likelihood estimation is used $\rightarrow$ truncated densities

Significant	Nonsignificant
$q_i^* = \dfrac{\phi\left(\frac{y_i-\mu}{\sqrt{\sigma_i^2+\tau^2}}\right)}{1-\Phi\left(\frac{y_{cv}-\mu}{\sqrt{\sigma_i^2+\tau^2}}\right)}$	$q_i^* = \dfrac{\phi\left(\frac{y_i-\mu}{\sqrt{\sigma_i^2+\tau^2}}\right)}{\Phi\left(\frac{y_{cv}-\mu}{\sqrt{\sigma_i^2+\tau^2}}\right)}$

- Likelihood function: $L(\mu, \tau^2) = \prod q_i^*$

From *p*-uniform to *p*-uniform*: *p*-uniform*

- Profile likelihood confidence intervals around estimates of average effect size and between-study variance

- Likelihood-ratio test for testing null hypotheses of no effect and homogeneity

- We also implemented several method of moments estimators

From p-uniform to p-uniform*: p-uniform*

- Profile likelihood confidence intervals around estimates of average effect size and between-study variance

- Likelihood-ratio test for testing null hypotheses of no effect and homogeneity

- We also implemented several method of moments estimators

- Important assumption:
  - Probability of a significant and nonsignificant effect size being included in a meta-analysis is assumed to be constant (but may differ from each other)

Selection model approach

- Selection model approaches are now seen as the state-of-the-art methods to correct of publication bias
- Many selection model approaches have been proposed
- Selection model approaches combine an effect size model with a selection model
  - Effect size model: Fixed-effect or random-effects model
  - Selection model: Function determining likelihood of a study to get published
- Issues:
  - Convergence problems for less than 100 studies
  - Selection model can often not be accurately estimated

Selection model approach

- Selection model approaches are now seen as the state-of-the-art methods to correct of publication bias
- Many selection model approaches have been proposed
- Selection model approaches combine an effect size model with a selection model
  - Effect size model: Fixed-effect or random-effects model
  - Selection model: Function determining likelihood of a study to get published
- Issues:
  - Convergence problems for less than 100 studies
  - Selection model can often not be accurately estimated
- *Note.* p-uniform* is actually also a selection model approach

Analytical study: Method

- **Goal:** Evaluate statistical properties of methods for one significant and one nonsignificant effect size
- Standardized mean difference was used as effect size measure with a sample size of 50 per group
- 1,000 equally spaced cumulative probabilities given significance/nonsignificance with $\alpha = .05$
- Converting probabilities to effect sizes: 1,000 x 1,000 $=$ 1,000,000

Analytical study: Method

- Conditions:
  - $\mu = 0$; 0.5
  - $\tau = 0$; 0.346 $\rightarrow I^2 = 0\%$; 75%
- Included methods:
  - *P*-uniform* using maximum likelihood estimation
  - Selection model approach by Hedges (1992) $\rightarrow$ cut-off at α=.05
- Outcome variables for both μ and τ:
  - Average, median, and standard deviation of estimates
  - Root mean square error (RMSE)
  - Coverage probability and width of 95% confidence interval

Analytical study: Results

- *P*-uniform always converged and Hedges1992 convergence was high (99.98%)

Analytical study: Results

- *P*-uniform always converged and Hedges1992 convergence was high (99.98%)

- Estimating μ for $\tau = 0$:

		$\mu = 0$	$\mu = 0.5$
Average (SD)	*p*-uniform*	0.014 (0.214)	0.486 (0.213)
	Hedges1992	0.029 (0.193)	0.486 (0.213)
RMSE	*p*-uniform*	214.5	213.1
	Hedges1992	195.1	213
Coverage	*p*-uniform*	0.958	0.959
	Hedges1992	0.971	0.949

Analytical study: Results

- Estimating μ for $\tau = 0.346$:

		$\mu = 0$	$\mu = 0.5$
Average (SD)	*p*-uniform*	0.043 (0.404)	0.475 (0.4)
	Hedges1992	0.062 (0.378)	0.477 (0.393)
RMSE	*p*-uniform*	406	400.3
	Hedges1992	383.5	393.8
Coverage	*p*-uniform*	0.818	0.821
	Hedges1992	0.84	0.81

Analytical study: Results

- Estimating μ for $\tau = 0.346$:

		$\mu = 0$	$\mu = 0.5$
Average (SD)	*p*-uniform*	0.043 (0.404)	0.475 (0.4)
	Hedges1992	0.062 (0.378)	0.477 (0.393)
RMSE	*p*-uniform*	406	400.3
	Hedges1992	383.5	393.8
Coverage	*p*-uniform*	0.818	0.821
	Hedges1992	0.84	0.81

- **Conclusions:**
  - Hardly any convergence problems
  - Performance of methods was comparable with small bias
  - Undercoverage in case of heterogeneity

Analytical study: Results

- Estimating τ for $\mu = 0$:

		$\tau = 0$	$\tau = 0.346$
Average (SD)	*p*-uniform*	0.031 (0.073)	0.167 (0.192)
	Hedges1992	0.037 (0.076)	0.185 (0.189)
RMSE	*p*-uniform*	78.8	262.5
	Hedges1992	84.9	248.3
Coverage	*p*-uniform*	0.996	0.995
	Hedges1992	-	-

Analytical study: Results

- Estimating τ for $\mu = 0$:

		$\tau = 0$	$\tau = 0.346$
Average (SD)	*p*-uniform*	0.031 (0.073)	0.167 (0.192)
	Hedges1992	0.037 (0.076)	0.185 (0.189)
RMSE	*p*-uniform*	78.8	262.5
	Hedges1992	84.9	248.3
Coverage	*p*-uniform*	0.996	0.995
	Hedges1992	-	-

- **Conclusions:**
  - Negative bias for estimating τ (also for $\mu = 0.5$)
  - Performance of methods was comparable
  - Severe overcoverage of *p*-uniform*'s confidence interval

Simulation study: Method

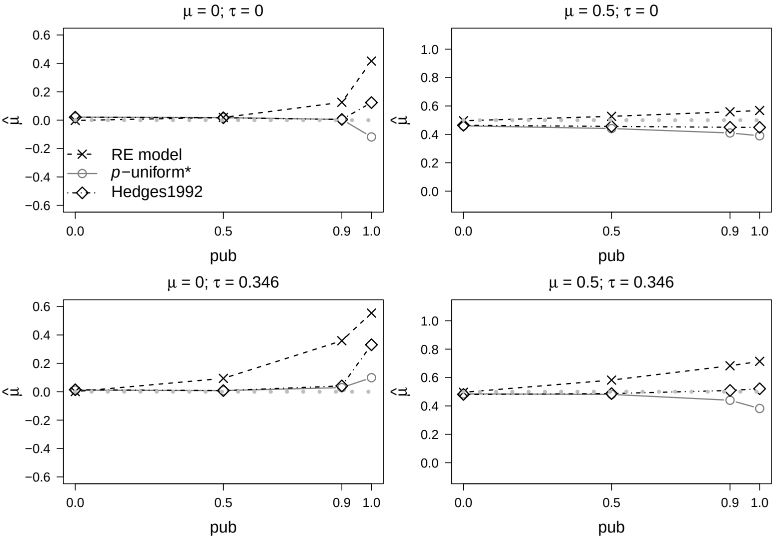
- **Goal:** Evaluate performance of *p*-uniform* and compare to other methods under realistic conditions

- Effect size measure is standardized mean difference with 50 as sample size per group

- Conditions:
  - $\mu = 0; 0.2; 0.5$
  - $\tau = 0; 0.163; 0.346 \rightarrow I^2 = 0\%; 40\%; 75\%$
  - Number of studies (k) = 10; 30; 60; 120
  - Extent of publication bias (*pub*) = 0; 0.5; 0.9; 1

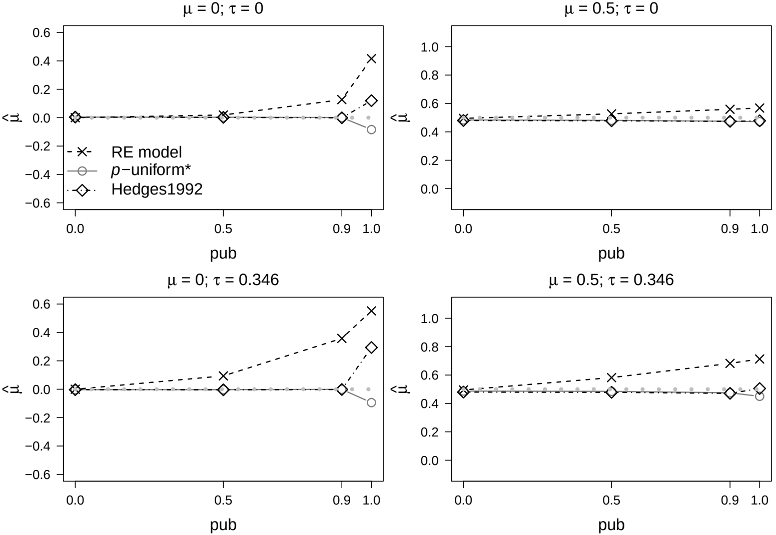
- Included methods:
  - *P*-uniform* using maximum likelihood estimation
  - Random-effects model $\rightarrow$ Paule-Mandel estimator for τ^2
  - Selection model approach by Hedges (1992) $\rightarrow$ cut-off at α=.05

Simulation study: Estimating μ



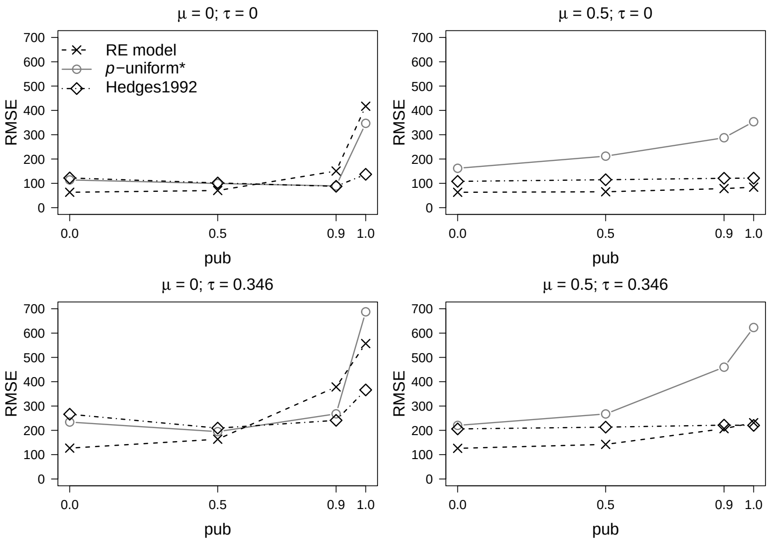
- Random-effects model overestimates μ if $pub > 0$
- Bias of p-uniform* and Hedges1992 is largest if $pub = 1$

Simulation study: Estimating μ ($k = 120$)



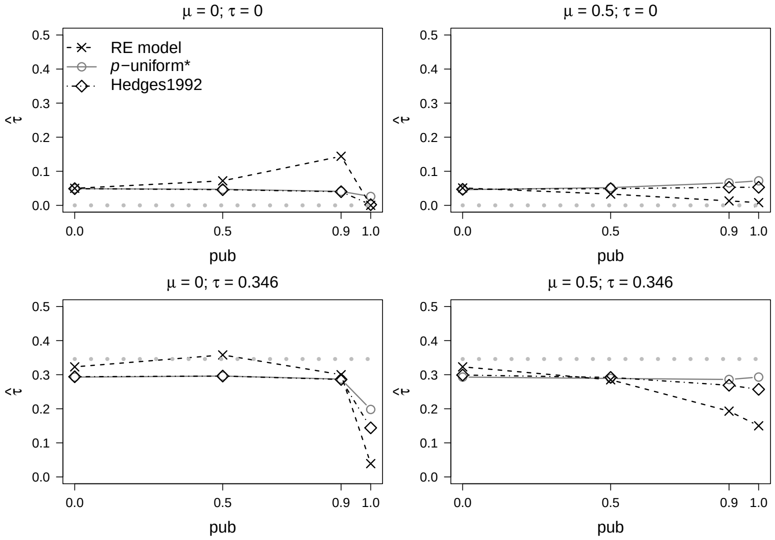
- Bias decreased for p-uniform* but hardly for Hedges1992

Simulation study: RMSE Estimating μ



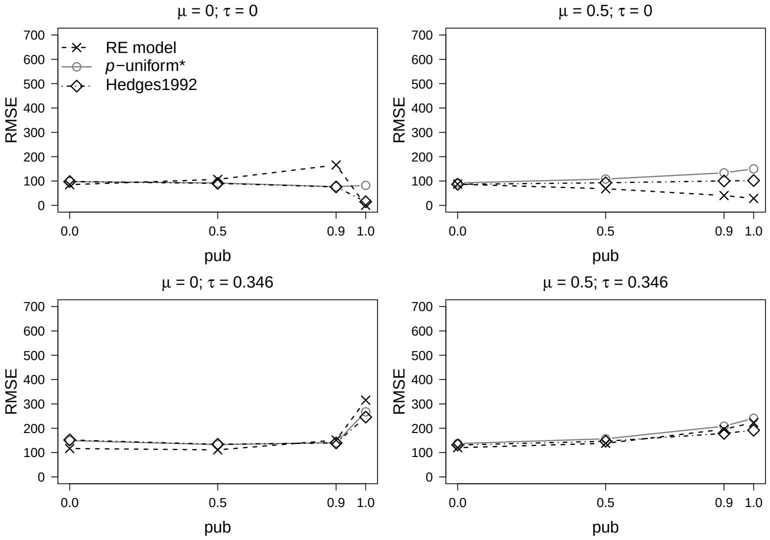
- RMSE of all methods increased as a function of τ and *pub*
- RMSE of *p*-uniform* generally larger than Hedges1992

Simulation study: Estimating τ



- RE model overestimates τ if $\tau = 0$ and underestimates if $\tau > 0$
- *P*-uniform* less negatively biased than Hedges1992 if $\tau > 0$

Simulation study: RMSE Estimating τ



- RMSE of all methods increased as a function of *pub* if $\tau > 0$
- RMSE of *p*-uniform* generally slightly larger than Hedges1992

Simulation study: Conclusions

- Random-effects model had the best properties in the absence of publication bias
- *P*-uniform*'s and Hedges1992's performance was comparable and outperformed random-effects model if $pub > 0$
- Non-convergence rates were at most 12.6% for *p*-uniform* and 15.8% for Hedges1992
- Worst statistical properties of all methods if $pub = 1$
- A systematic positive bias in estimating μ was apparent for Hedges1992

Conclusion and discussion

- *P*-uniform* is an improvement over *p*-uniform, because
  1. eliminates overestimation due to between-study variance
  2. is a more efficient estimator than *p*-uniform's estimator
  3. enables estimating and testing of the between-study variance

- Statistical properties of *p*-uniform* and the selection model approach by Hedges (1992) were comparable

- Non-convergence was not as severe as suggested in the literature

- Recommendations:
  - Report results of *p*-uniform* and Hedges1992 in any meta-analysis
  - Do not put too much trust in estimates if you expect extreme publication bias with only significant effect sizes

Conclusion and discussion

- Software:
  - *p*-uniform*: R package `puniform` and web application https://rvanaert.shinyapps.io/p-uniformstar
  - Hedges' selection model approach: R package `weightr` and web application https://vevealab.shinyapps.io/WeightFunctionModel

- Future research:
  - Violation of the assumption of equal probabilities of significant and nonsignificant effect sizes for being included in a meta-analysis
  - *P*-uniform*'s publication bias test
  - Consequences of questionable research practices

Thank you for your attention

For these slides see: www.robbievanaert.com