How to detect publication bias in psychological research?

A comparative evaluation of six statistical methods

Frank Renkewitz & Melanie Keiner

University of Erfurt

The replicability crisis

- Many psychological findings are not replicable – up to 64% (OSC, 2015)
- Concerns even very prominent effects:
  - ego depletion, behavioral priming, unconscious thought effect, facial feedback…

What to trust?

- Re-evaluation of published evidence

Reasons for replicability problems

- Publication bias
- Questionable research practices / *p*-hacking

Needed: Methods to detect these problems

Methods to detect publication biases and p-hacking

Distribution of effect sizes (funnel plot)	Distribution of *p*-values
• Regression-based methods (e.g. Stanley & Doucouliagos, 2014) ▫ PET: $E[ES] = b_0 + b_1 \cdot SE$ (weight: $1/SE^2$) • Begg's Rank Correlation (1994) • Trim-and-Fill (Duval & Tweedie, 2000) ▫ Selection model: exclusion of *k* smallest ES	• *p*-curve (Simonsohn et al., 2014) • *p*-uniform (van Assen et al., 2014) • Test of insufficient variance (TIVA; Schimmack, 2014) ▫ Transformation of *p*-values into *z*-scores ▫ Expected variance of *z*-scores: $s^2 = 1$

- Test of excess significance (TES; Ioannidis & Trikalinos, 2007; Francis, 2013)
  - Comparison of post-hoc power and proportion of significant results.

Research question

- Relative and absolute performance of these methods
  - in detecting biases?
  - in correcting for biases?

  - False positive rate
  - Power

- Monte Carlo simulation

Design I (selection models)

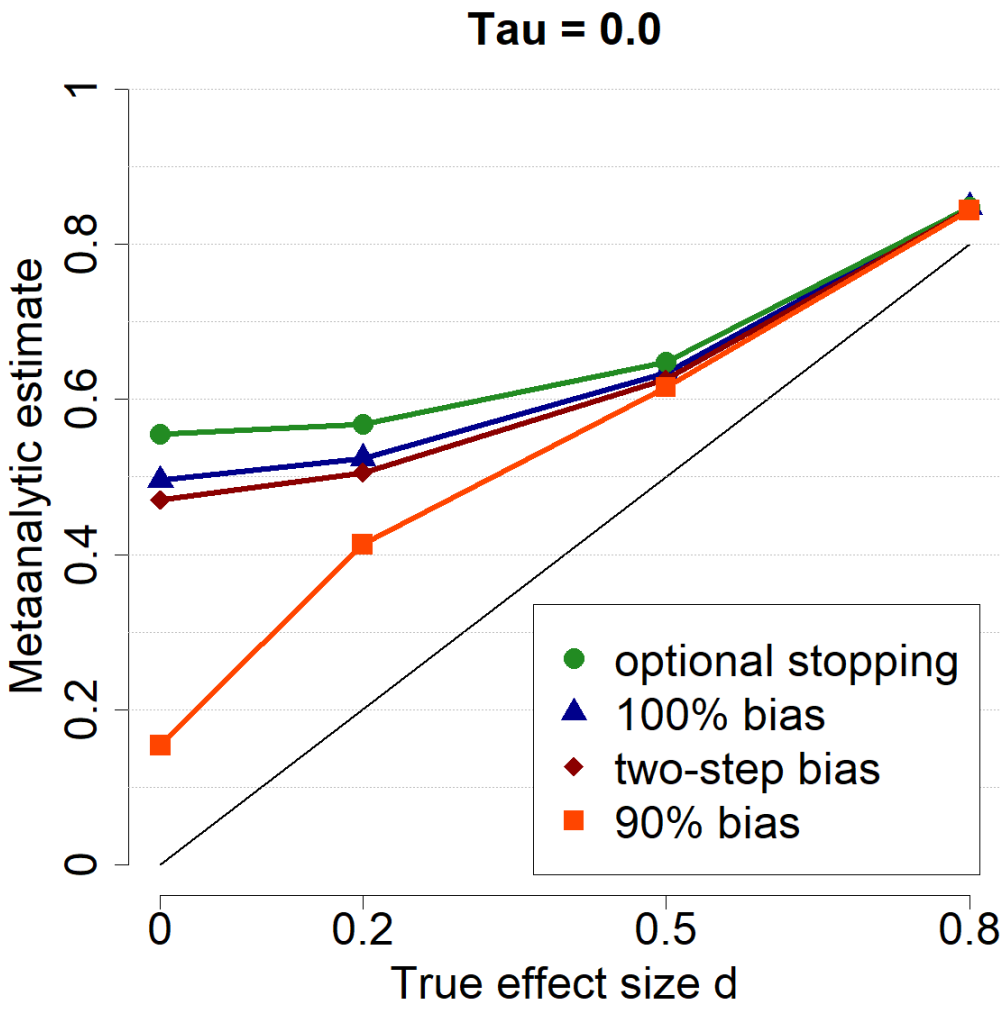
- Bias conditions
  - No bias (all studies included)
  - 100% bias (exclusion of all non-significant studies; p > .05, one-tailed)
  - Two-step bias
    - Studies with $0.05 < p < 0.10$ are included with a probability of 0.2.
  - 90% bias
    - Non-significant studies are excluded with a probability of 0.9.
  - *p*-hacking (optional stopping)
    - Start with n = 20 (per cell)
    - repeat test with 1 **or** 5 additional participants per condition.
    - Stop testing at $n = 40$ if no significant result was obtained.

Design II (data models)

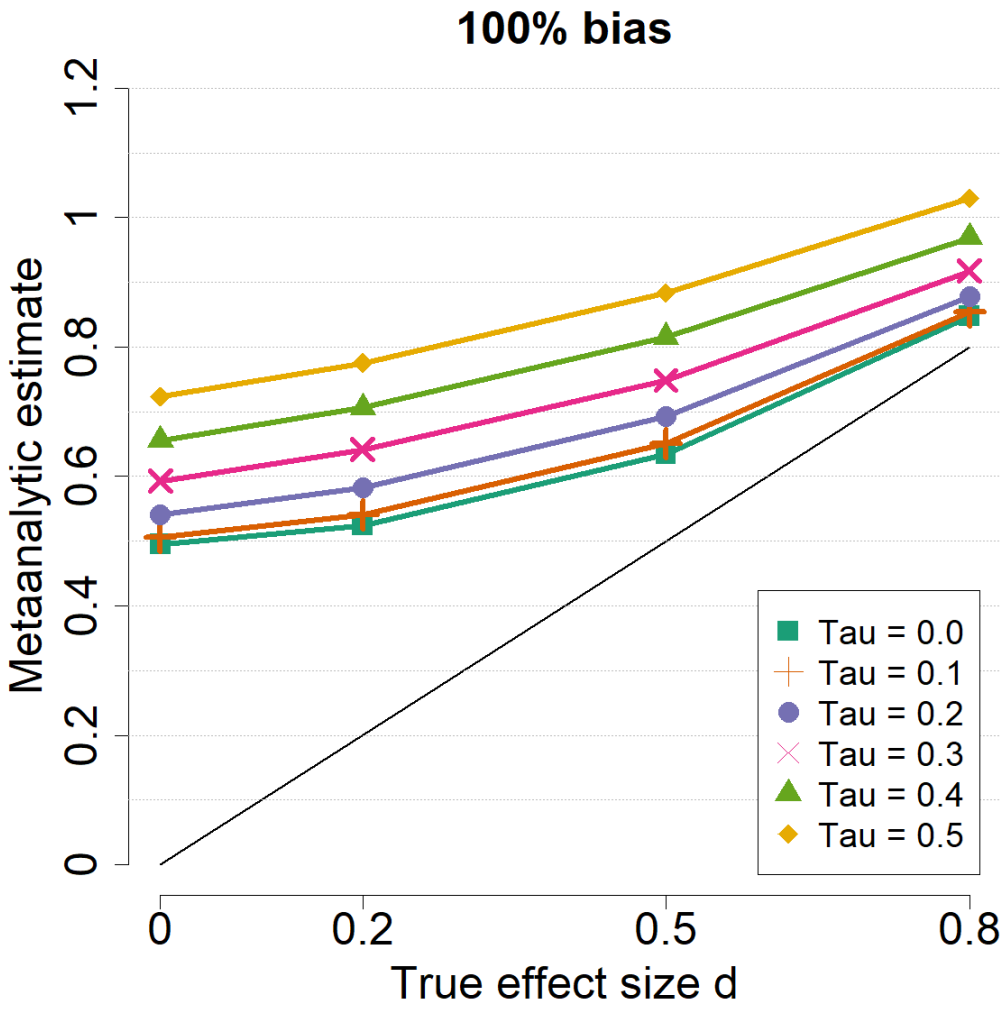
- Effect sizes: *d* = 0, 0.2, 0.5, 0.8
- Heterogeneity: τ = 0, 0.1, 0.2, 0.3, 0.4, 0.5
- Sample sizes per study: Drawn from uniform distributions
  - $U(20, 30)$, $U(10, 40)$, $U(45, 55)$, $U(35, 65)$ and $U(20, 80)$
- Studies per meta-analysis: 5, 7, 10, 30, 50

- 2,640 conditions in total
- 1,000 iterations per condition

Biases in meta-analytic estimates



Biases in meta-analytic estimates

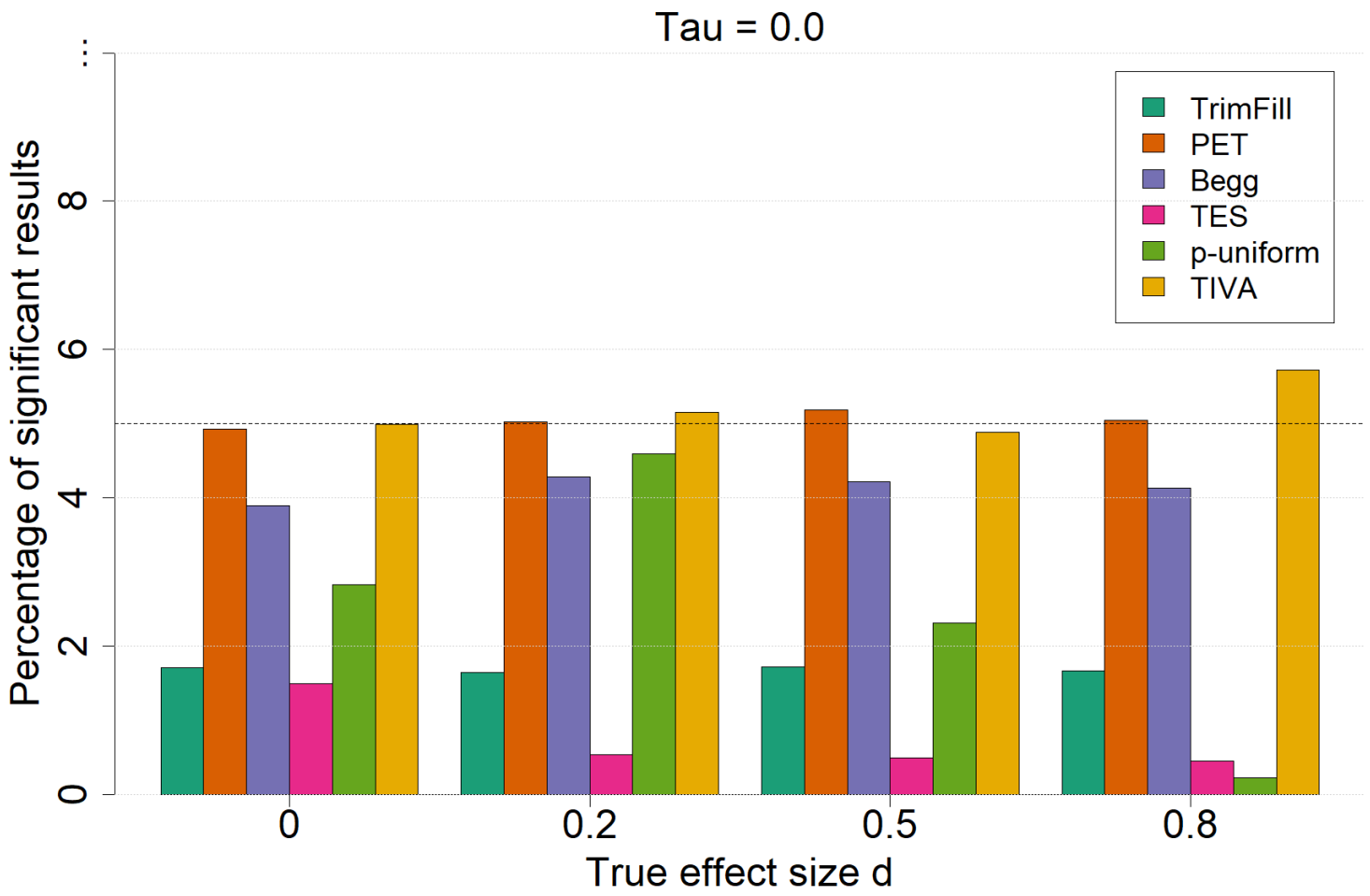


Heterogeneity boosts the amount of bias.

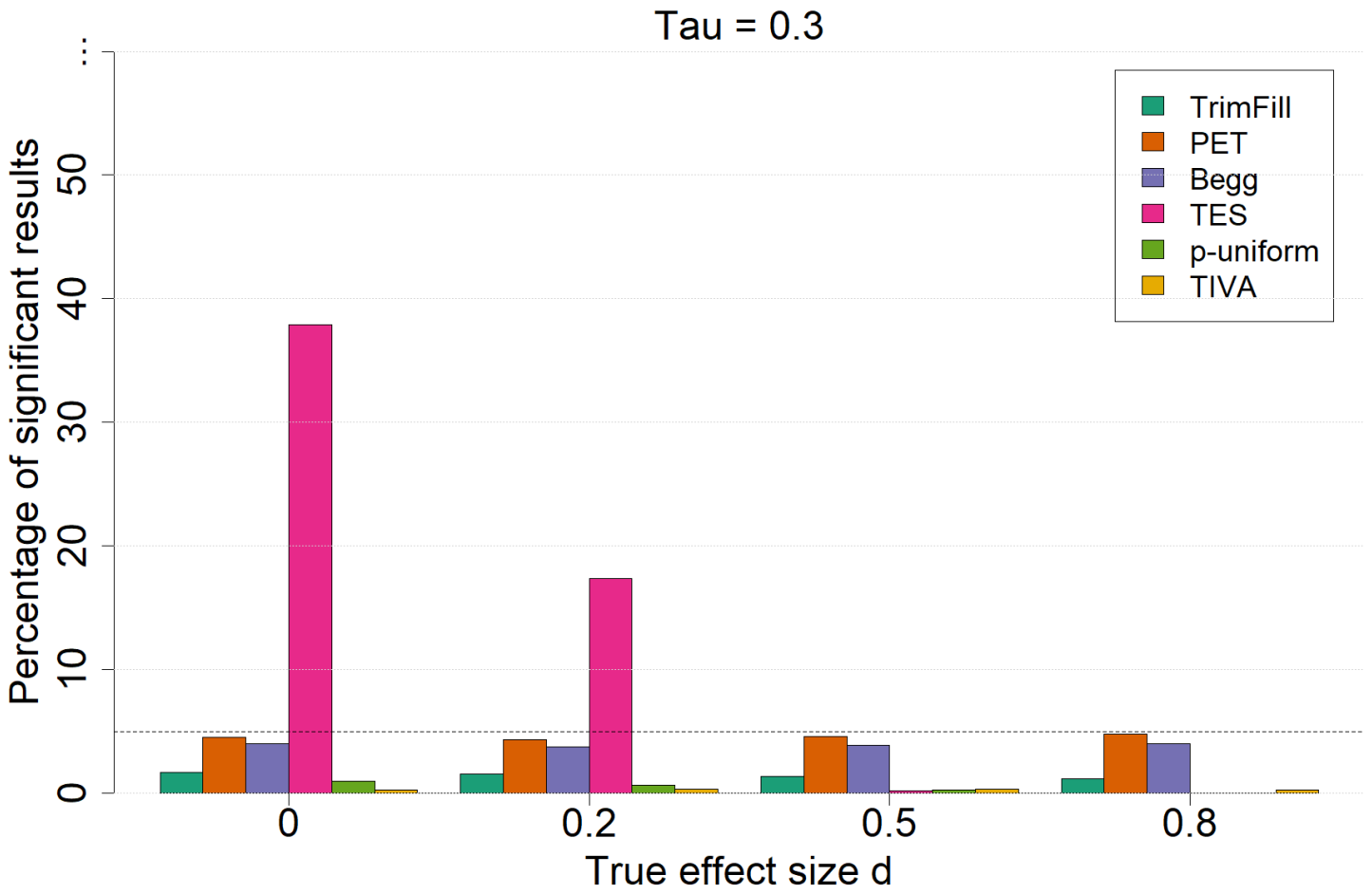
Results I

False positive rates

False positive rates (under homogeneity)



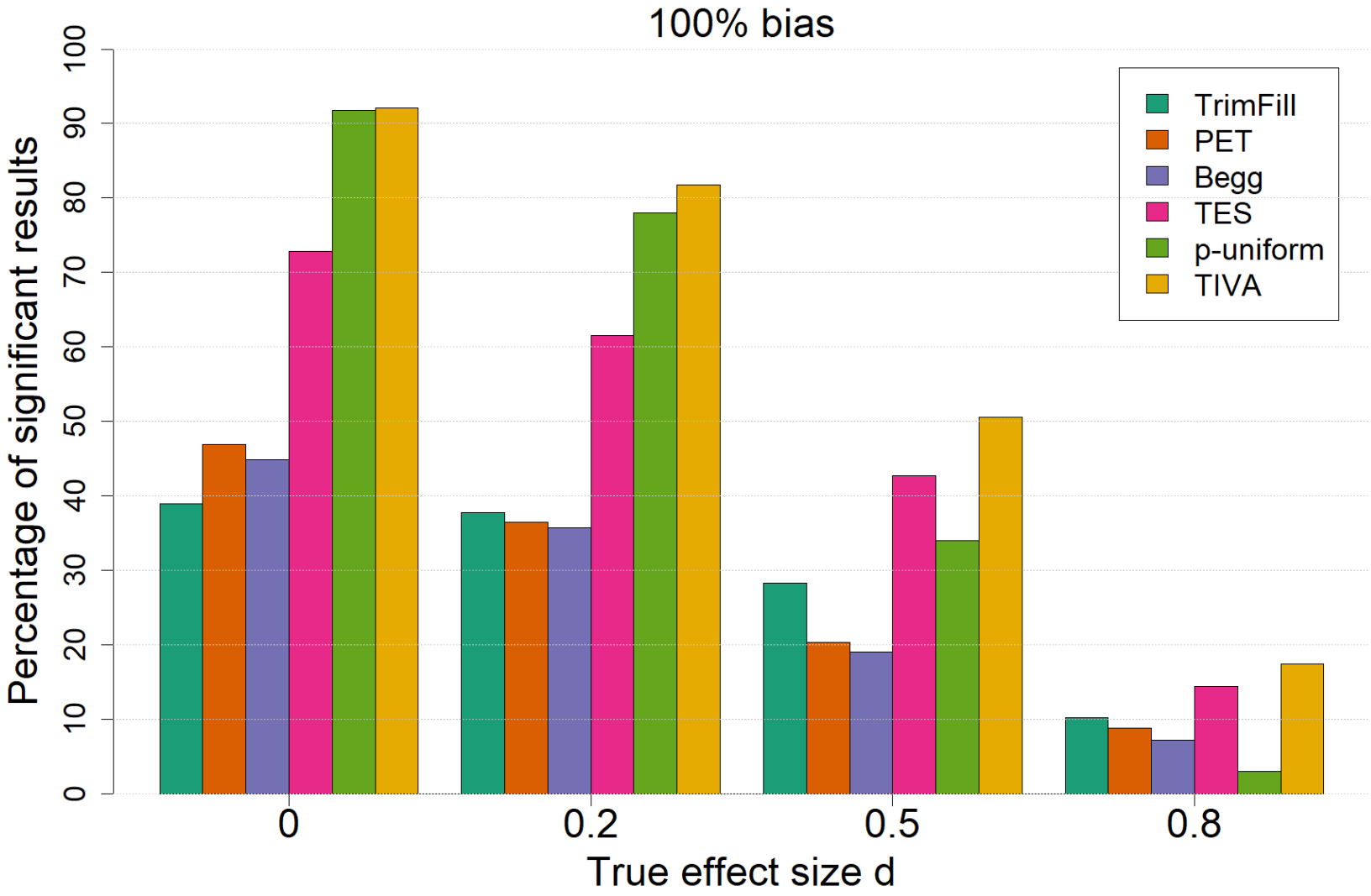
False positive rates (under heterogeneity)



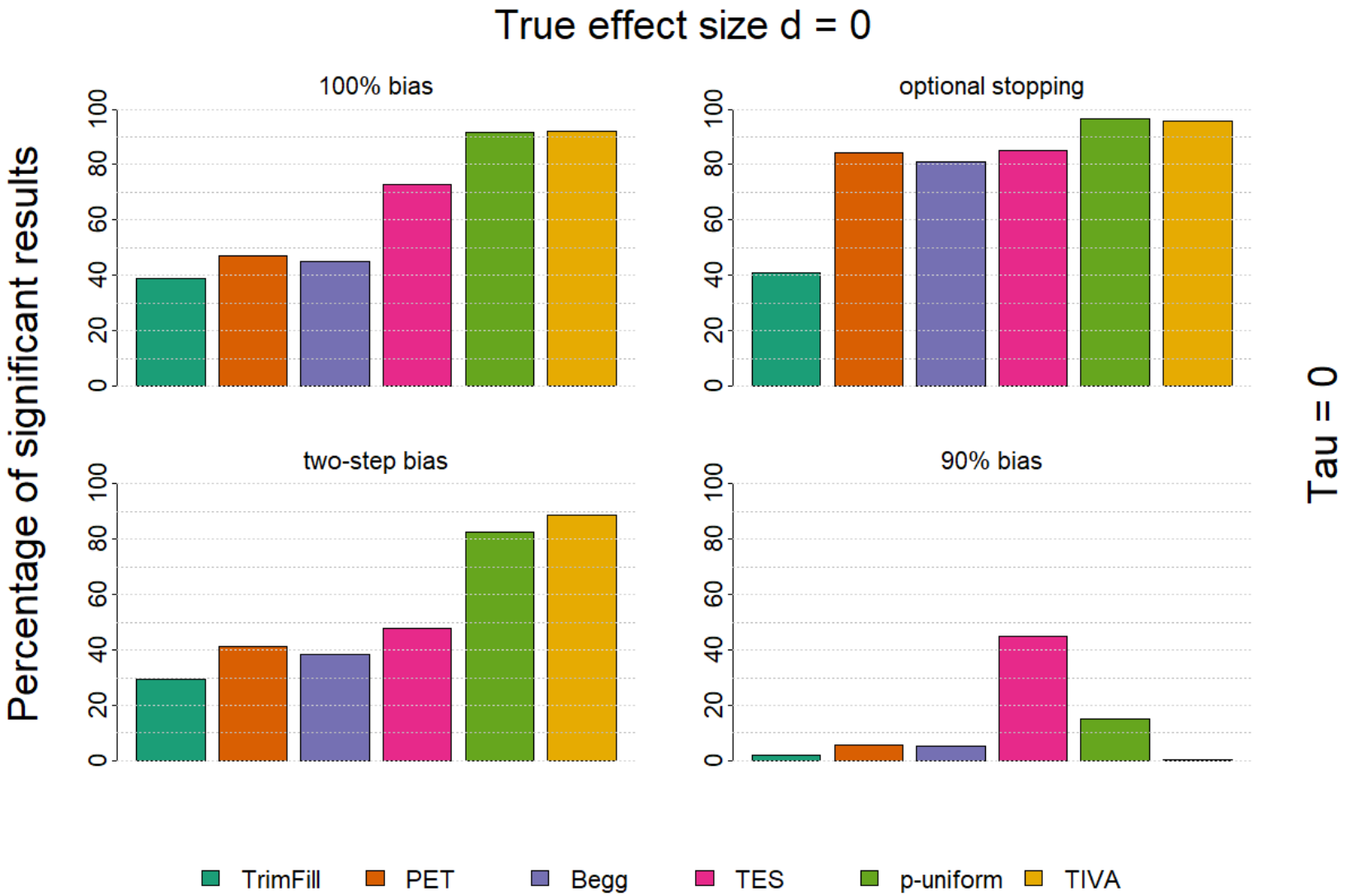
Results II

Power

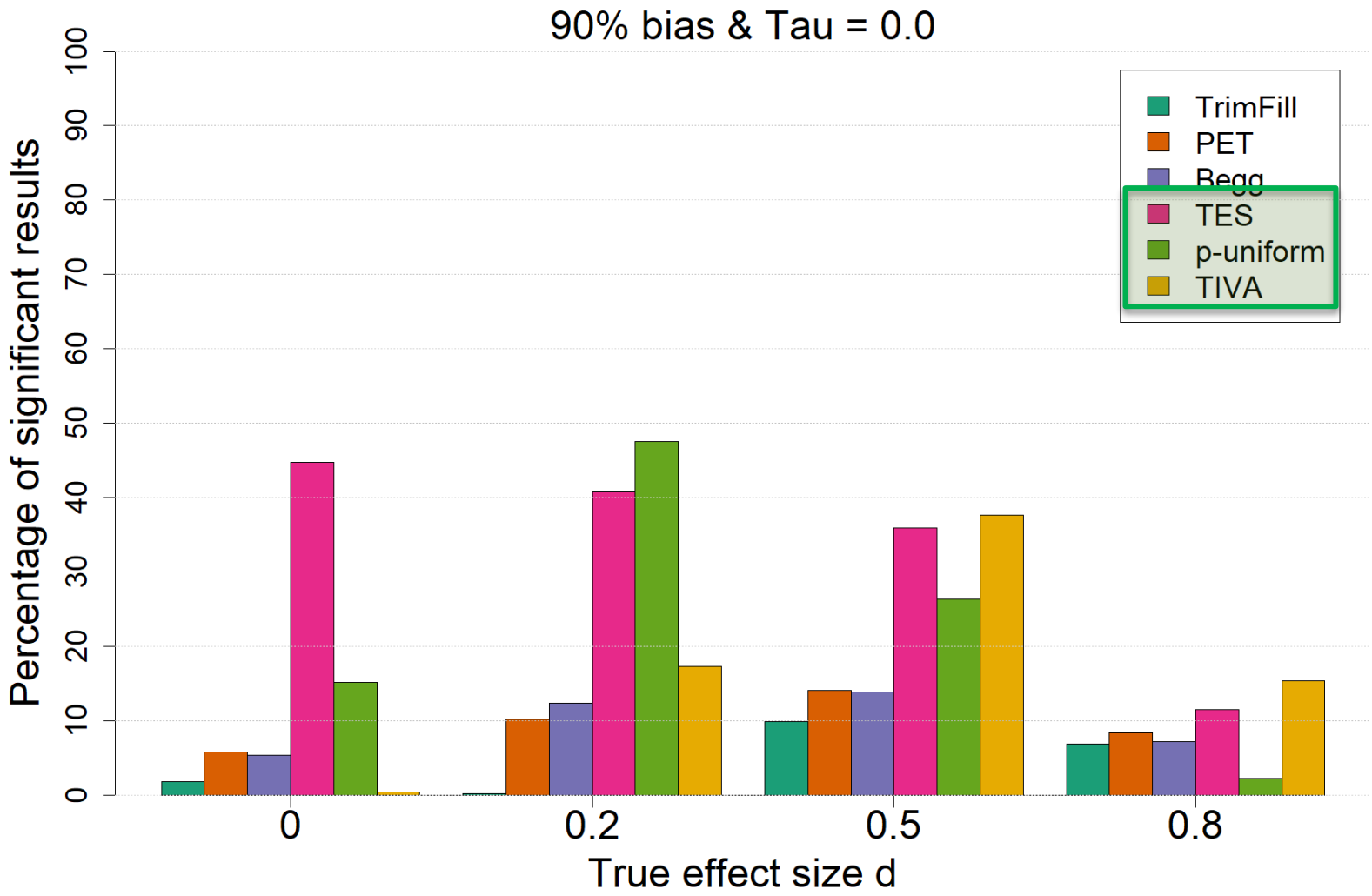
Power (under homogeneity)



Power (under homogeneity)

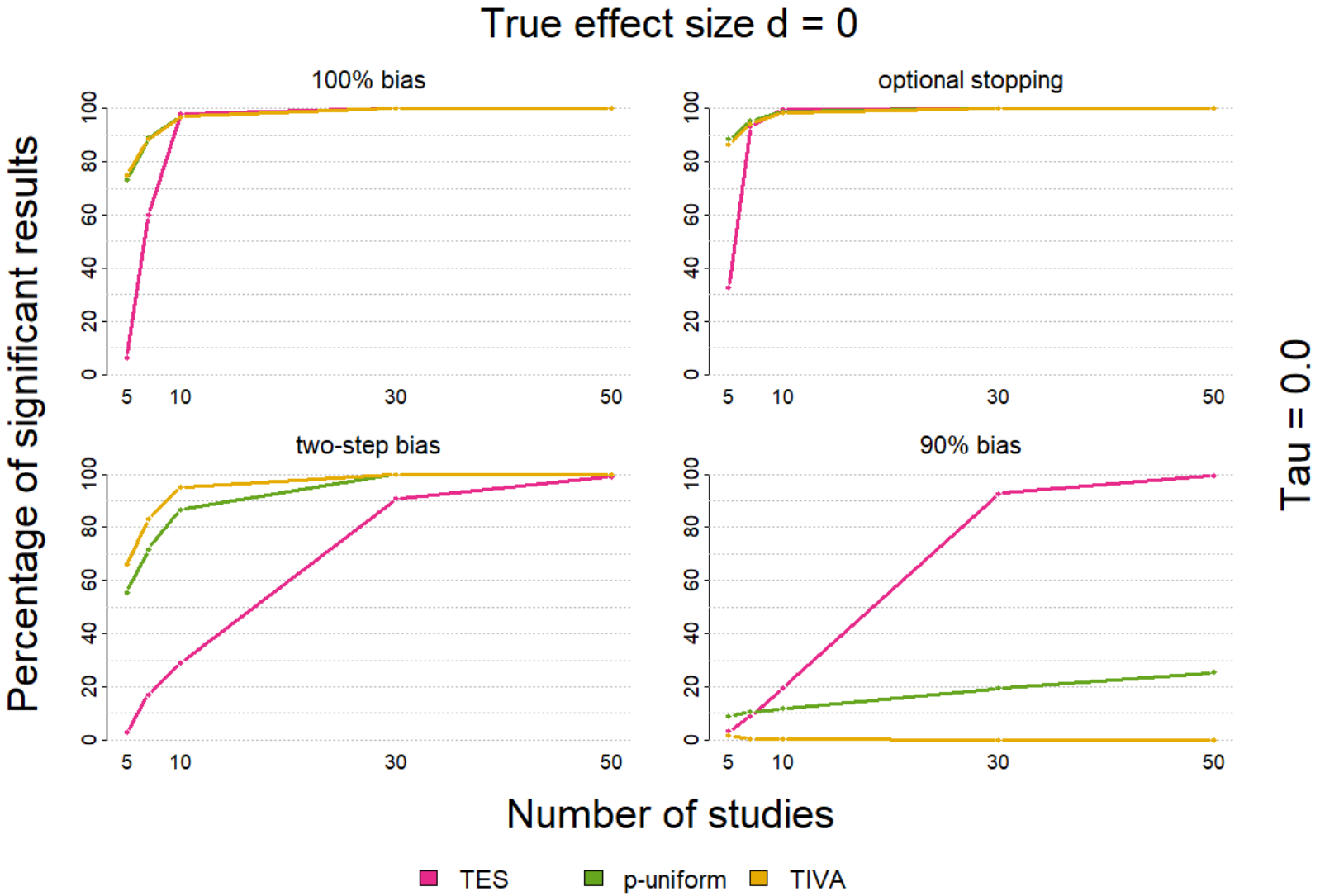


Power (under homogeneity, 90% bias)



Power for different numbers of primary studies

(under homogeneity)



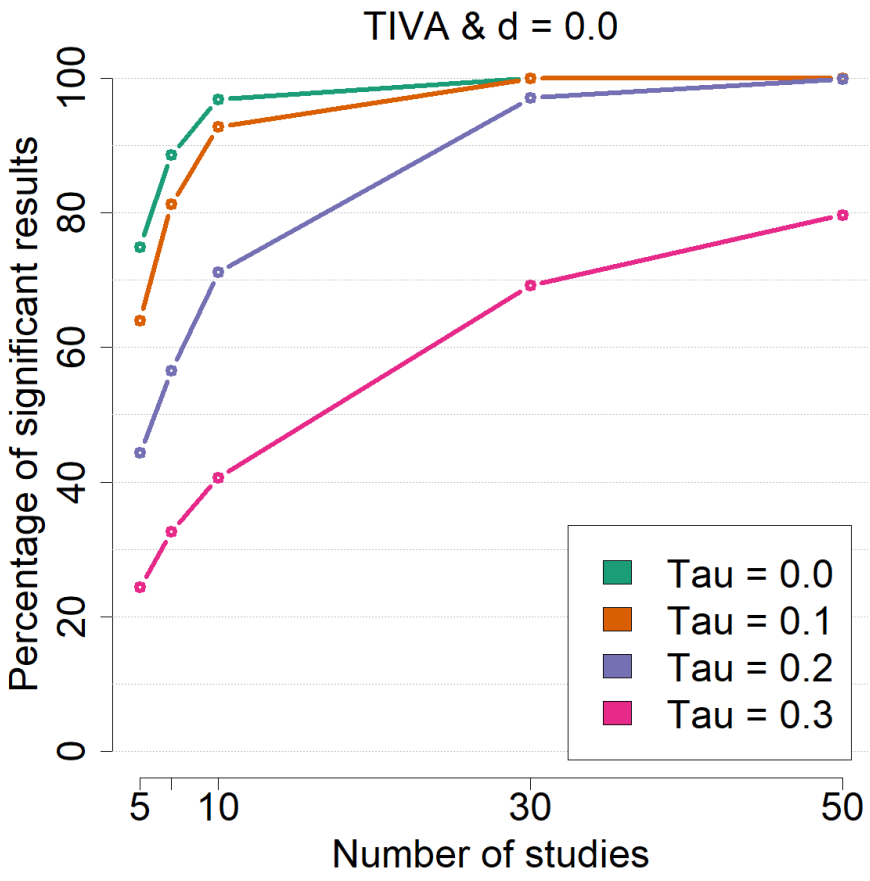
Summary I- Bias detection under homogeneity

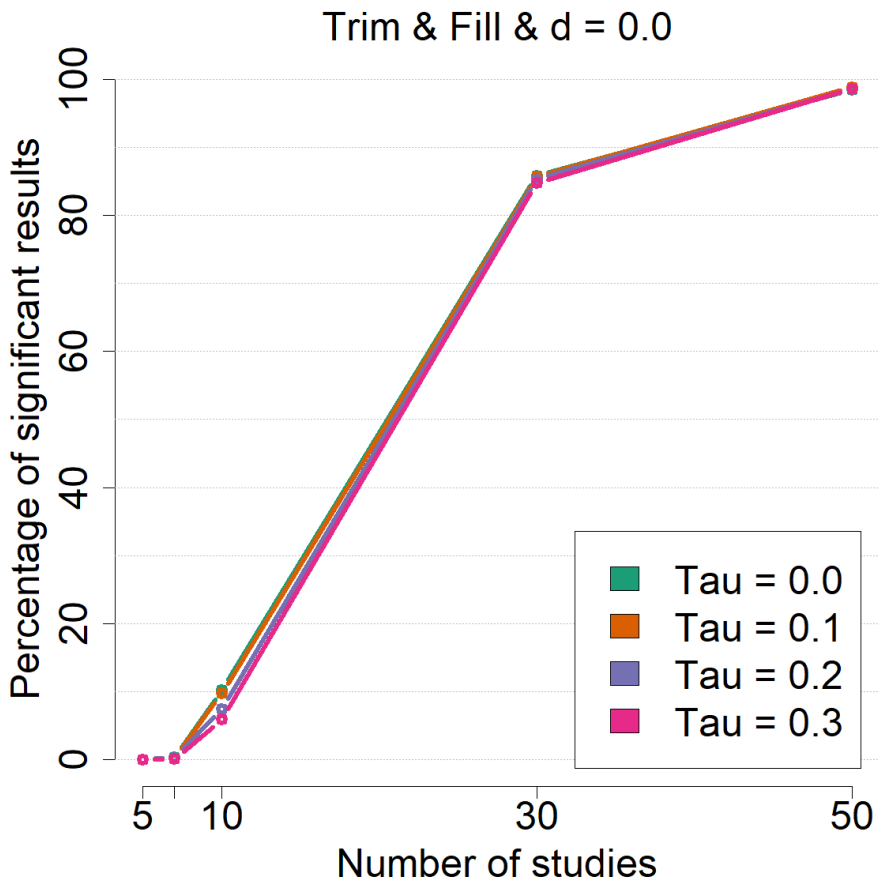
- Methods based on the distribution of *p*-values perform very well even in small study sets ($k = 5$) when…
  - non-significant results are severely censored **and**
  - the true effect size is small.

- When censorship is less severe…
  - TIVA and *p*-uniform fail.
  - Bias detection in small study sets ($k \leq 10$) is futile.
  - TES performs excellently in larger sets ($k \geq 30$)
    - But: Inflated Type 1 error rate under heterogeneity

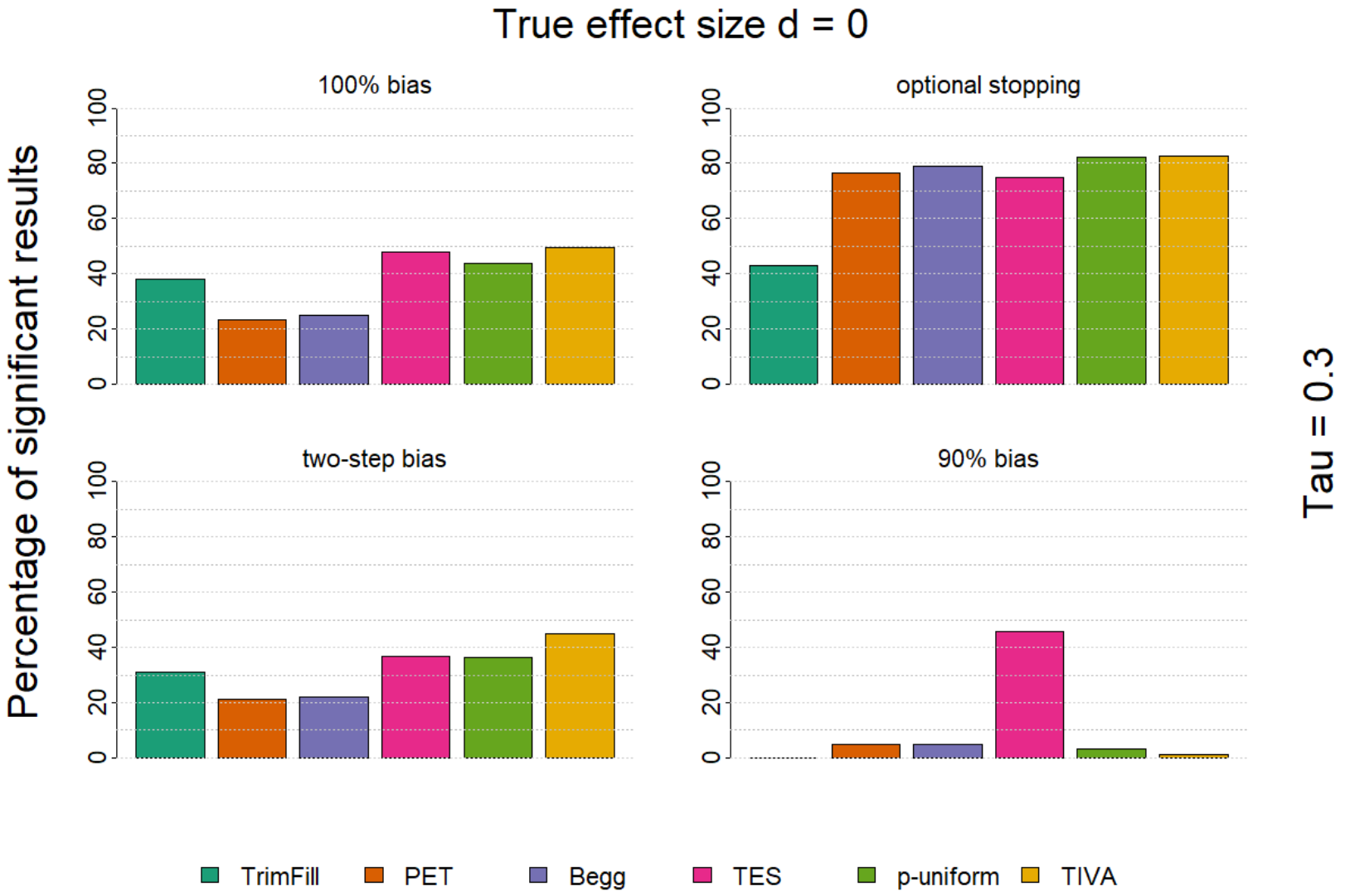
- No generally superior method

Power (under heterogeneity)



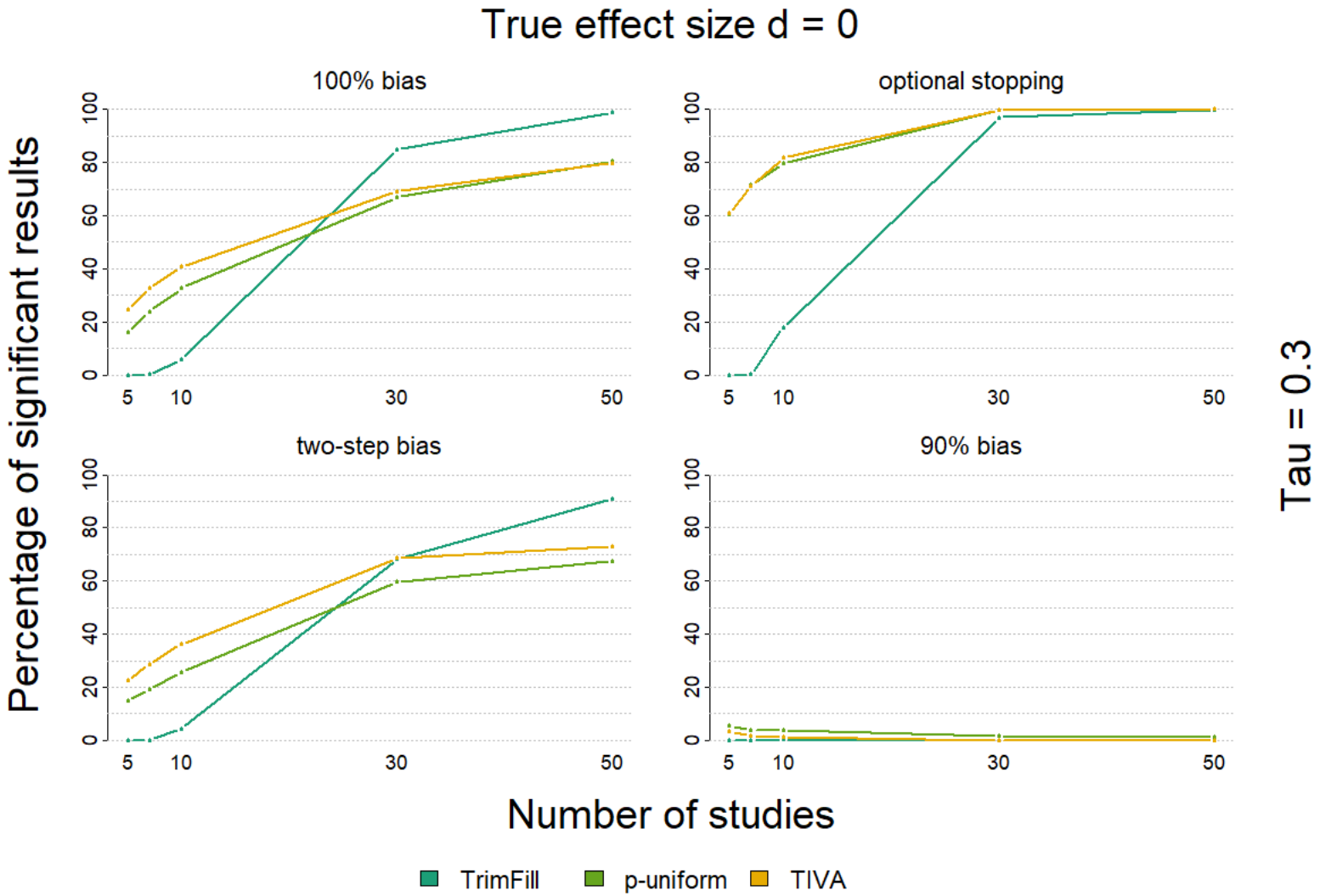


Power (under heterogeneity)



Power for different numbers of primary studies

(under heterogeneity)



Summary II - Bias detection under heterogeneity

- Heterogeneity hampers the detection of publication biases
  - even though it increases the bias in meta-analytic estimates.
  - **Meta-analysts should try to reduce heterogeneity.**
- In small study sets ($k \leq 10$) power is (almost) always below 50%.
- In larger study sets and with severe censorship TIVA and *p*-uniform still perform relatively well
  - But they are outperformed by trim-and-fill.
  - No generally superior method.

- When censorship is less severe and there is a medium degree of heterogeneity bias detection is impossible.
  - In many psychological meta-analyses biases will remain undetected.

Discussion

- Methods fare well when assumptions of their data and selection model are met.
- But they react sensitively to violations of these assumptions.
  - selection functions, heterogeneity, *p*-hacking…

- Too many unknowns in the complete publication process
  - Another plea for open science…

Denn der radikalste Zweifel ist der Vater der Erkenntnis.

(Max Weber)

Danke!