Aufdeckung und Korrektur von Biases in publizierter Evidenz

Frank Renkewitz & Melanie Keiner

Universität Erfurt

Die Replikationskrise

- Zahlreiche psychologische Befunde sind nicht replizierbar – bis zu 64% (OSC, 2015)

- Dies betrifft auch „prominente“ und vielfach untersuchte Effekte:
  - ego depletion, behavioral priming, unconscious thought effect, facial feedback…

Welchen Befunden sollten wir trauen?

- Reevaluation publizierter Evidenz

Gründe für Replikationsprobleme

- Publikationsbias
  - Negative (nicht-signifikante) Befunde werden selten publiziert.
  - Anteil positiver (signifikanter) Befunde zu zentralen Hypothesen:
    - Sterling (1959): 97,3%
    - Bozarth & Roberts (1972): 94%
    - Sterling et al. (1995): 95,6%
    - Fanelli (2010): 92%
- Fragwürdige Forschungspraktiken / *p*-Hacking
  - "Produktion" signifikanter Ergebnisse (durch wiederholtes Testen).
  - Hinweise, dass dies weit verbreitet ist (John et al., 2012; Head et al., 2015).

Benötigt: Methoden, die diese Probleme identifizieren (und korrigieren)

Methoden zur Aufdeckung von Biases

Verteilung von Effektstärken (Funnel Plot)

- Trim-and-Fill (Duval & Tweedie, 2000)
- Regressionsmethoden (e.g. PET-PEESE, Stanley & Doucouliagos, 2014)

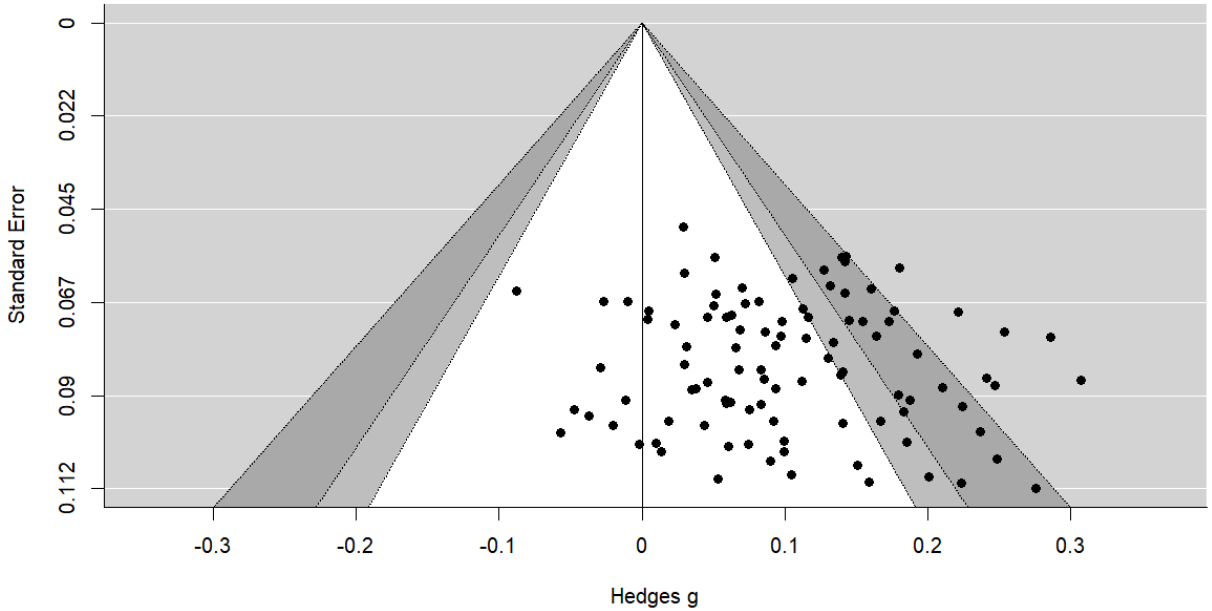
Verteilung (signifikanter) *p*-Werte

- p-curve (Simonsohn et al., 2014)
- p-uniform (van Assen et al., 2014)

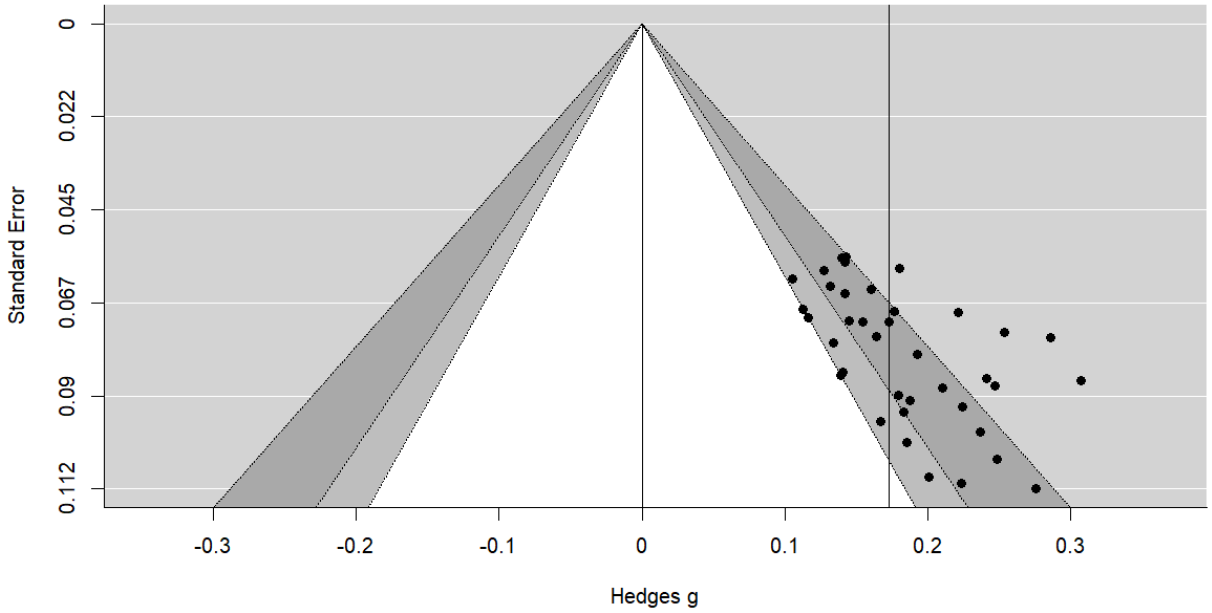
Weiteres

- Test of excess significance (TES - Francis, 2013)
- Test of insufficient variance (TIVA – Schimmack, 2014)
- …

Funnel Plot

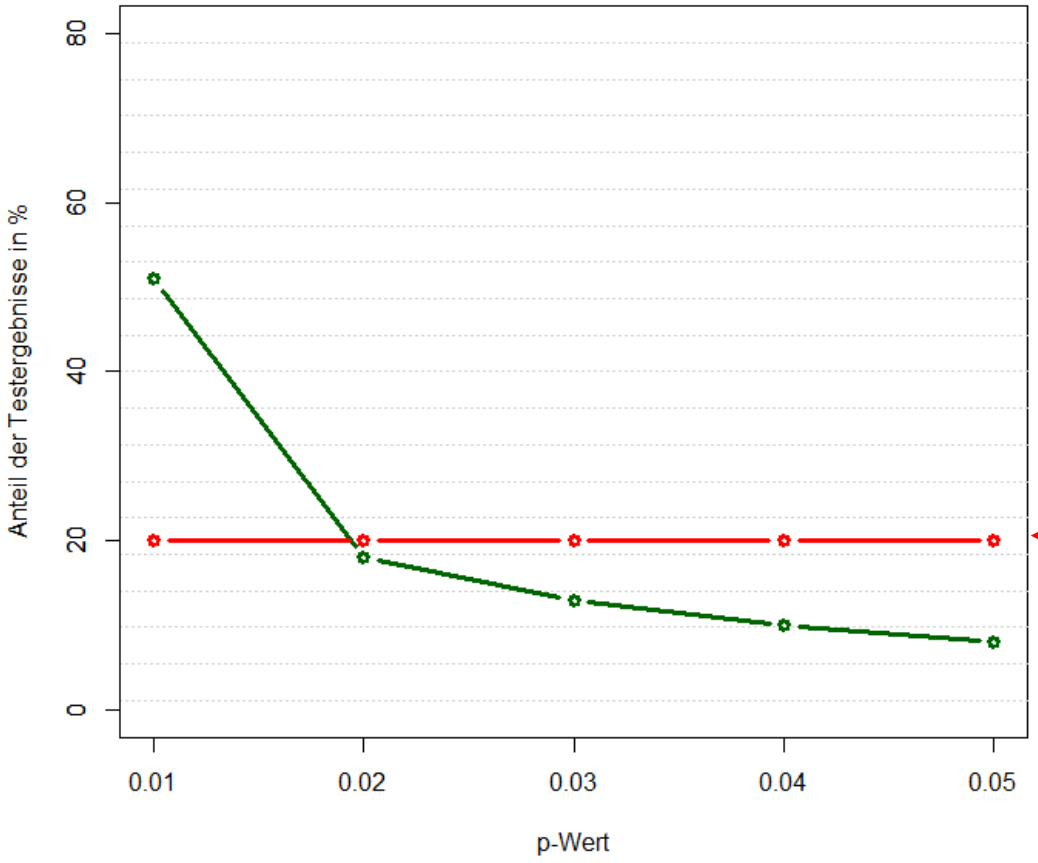

Standard Error
0
0.022
0.045
0.067
0.09
0.112
-0.3
-0.2
-0.1
0
0.1
0.2
0.3
Hedges g

Funnel Plot



- Regressionsmethoden (Stanley & Doucouliagos, 2014; Egger et al., 1997)
  - PET: E[*ES*] = $b_0 + b_1 \cdot SE$ Gewicht: $1/SE^2$
  - PEESE: E[*ES*] = $b_0 + b_1 \cdot SE^2$ Gewicht: $1/SE^2$
- Trim-and-Fill (Duval & Tweedie, 2000)
  - Non-parametrische Maße der Asymmetrie
  - **Bias-Modell: Ausschluss der *k* kleinsten *ES***
- Begg‘s Rangkorrelation (1994)

p-curve und p-uniform



p-Verteilung bei *ES* = 0

p-Verteilung bei *ES* > 0
(Schiefe steigt mit Power der Studien)

- Bias-Modell: Ausschluss nicht-signifikanter Studien

Weiteres…

TES (Francis, 2013)

- Vergleich der beobachteten Power mit dem Anteil signifikanter Studien

TIVA (Schimmack, 2014)

- Transformation von *p*-Werten in *z*-Werte (der non-zentralen Verteilung).
- Erwartete Varianz der *z*-Werte: $s^2 = 1$
- Varianzeinschränkung ist Indiz für Publikationsbias (oder p-Hacking)

- …

Foschungsfrage

- „Relative“ und „absolute“ Performanz dieser Methoden?
  - Bei der Aufdeckung von Biases?
  - Bei der Korrektur von Biases?

- Monte-Carlo-Simulation

Design I (Bias-Modelle)

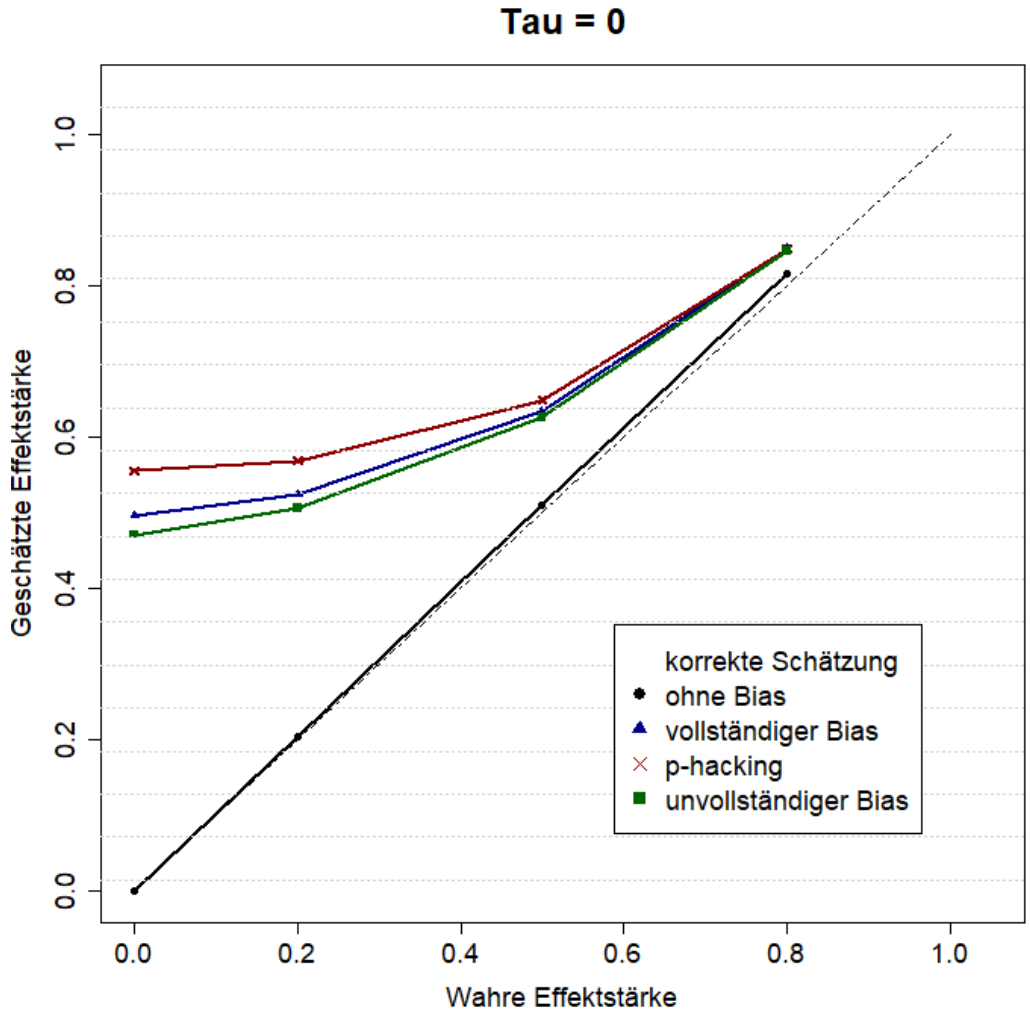
- „Selektionsbedingungen“
  - Kein Bias (alle Studien inkludiert)
  - Vollständiger Publikationsbias (Ausschluss aller nicht-signifikanter Studien)
  - „Unvollständiger“ Publikationsbias
    - Studien mit $0.05 < p < 0.10$ werden mit der Wahrscheinlichkeit .2 inkludiert
    - 92% aller „publizierten“ Studies sind signifikant
  - *p*-Hacking
    - Start mit $n = 20$ (pro Zelle)
    - Test wird wiederholt mit 1 **oder** 5 zusätzlichen Teilnehmer(n) pro Bedingung.
    - Stopp bei $n = 40$, wenn kein signifikantes Ergebnis erzielt wurde.

Design II (Daten-Modelle)

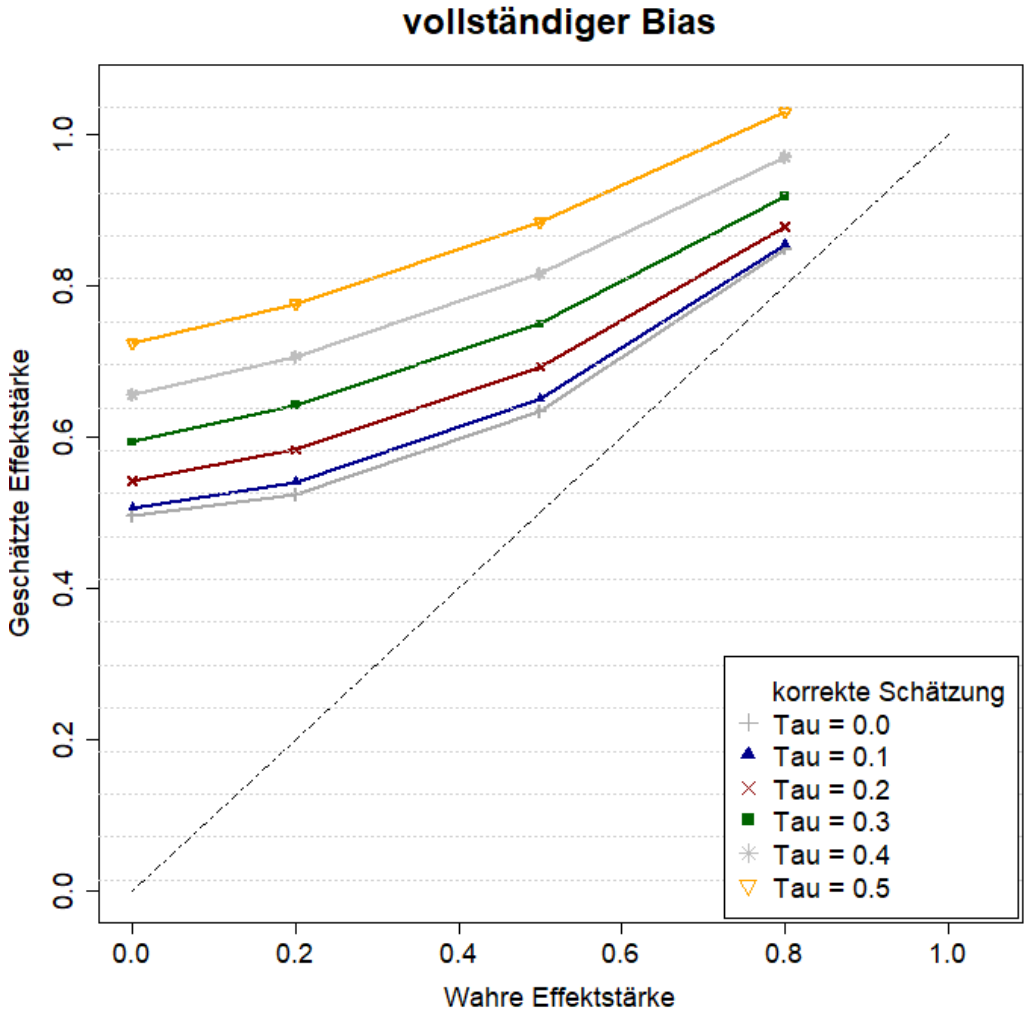
- Effektstärken: δ = 0, 0.2, 0.5, 0.8
- Studien pro Metaanalyse: 5, 7, 10, 30, 50
- Stichprobengrößen: Gleichverteilungen
  - $U(20, 30)$, $U(10, 40)$, $U(45, 55)$, $U(35, 65)$ and $U(20, 80)$
- Heterogenität: τ = 0, 0.1, 0.2, 0.3, 0.4, 0.5

- 1.000 Iterationen pro Bedingung
- Anzahl simulierter Metaanalysen: 2.04 Mill.

Metaanalytische Schätzungen: Simulierte Biases

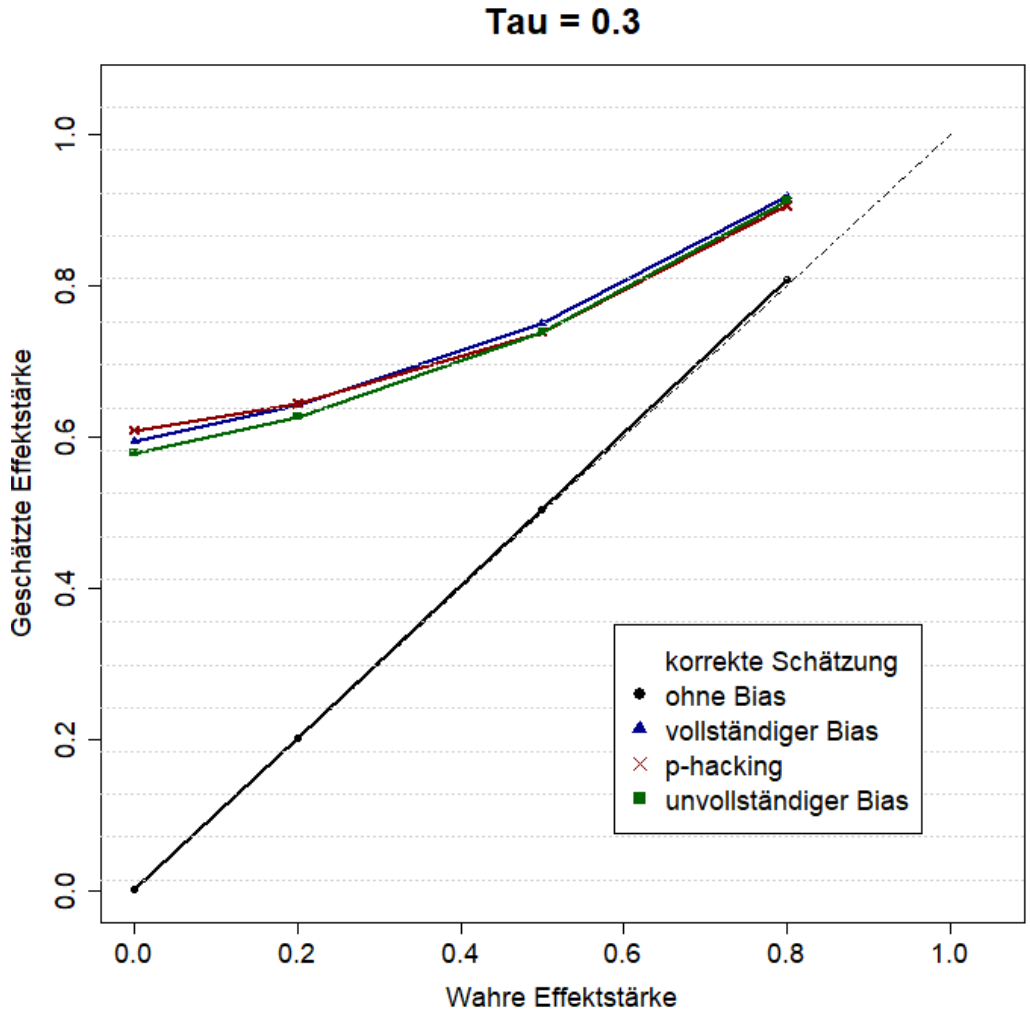


Metaanalytische Schätzungen: Simulierte Biases



Heterogenität verstärkt Biases!

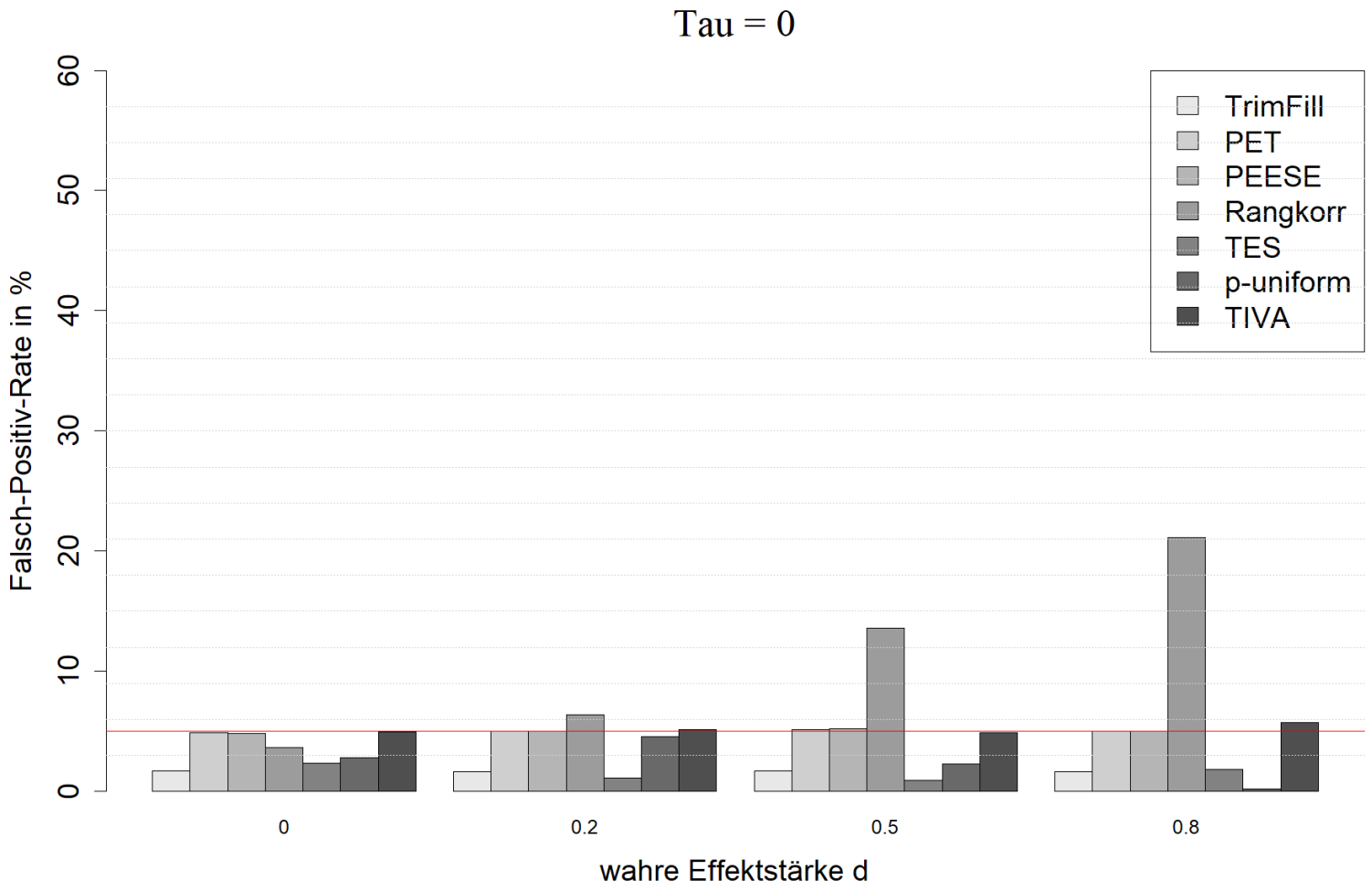
Metaanalytische Schätzungen: Simulierte Biases



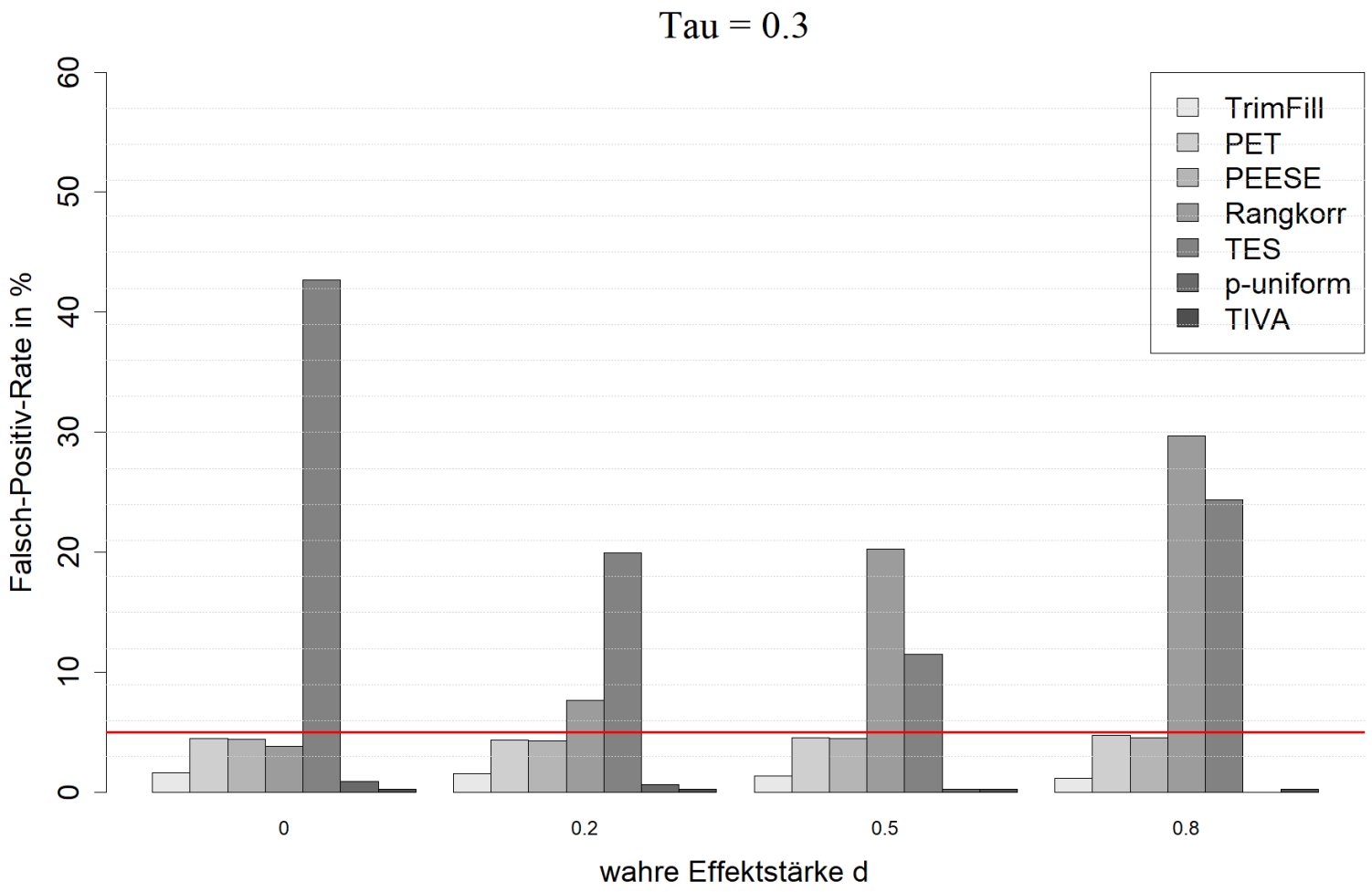
Heterogenität nivelliert Biases!

Können wir Biases aufdecken?

Falsch-positiv-Raten (Homogenität)



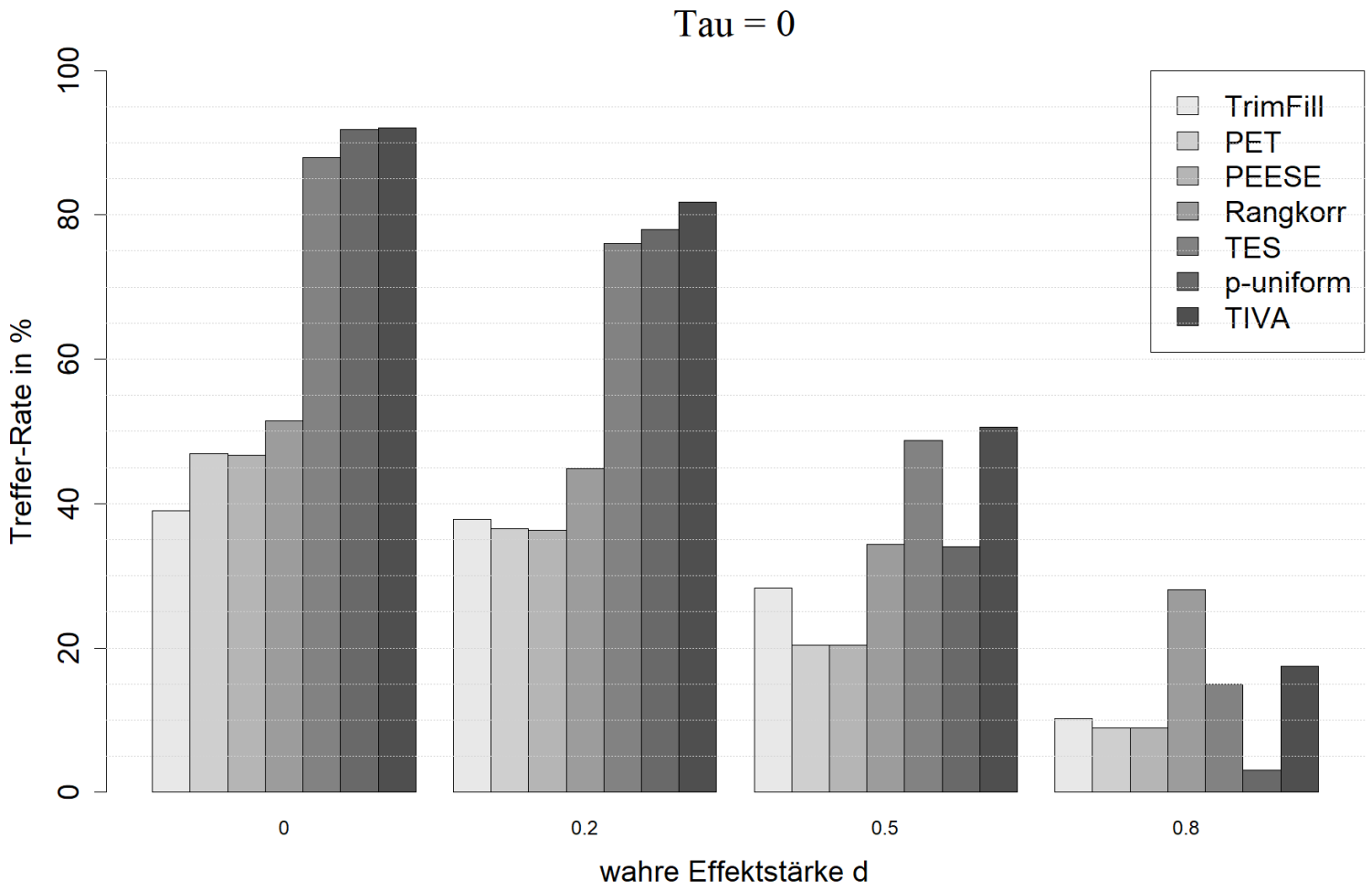
Falsch-positiv-Raten (Heterogenität)



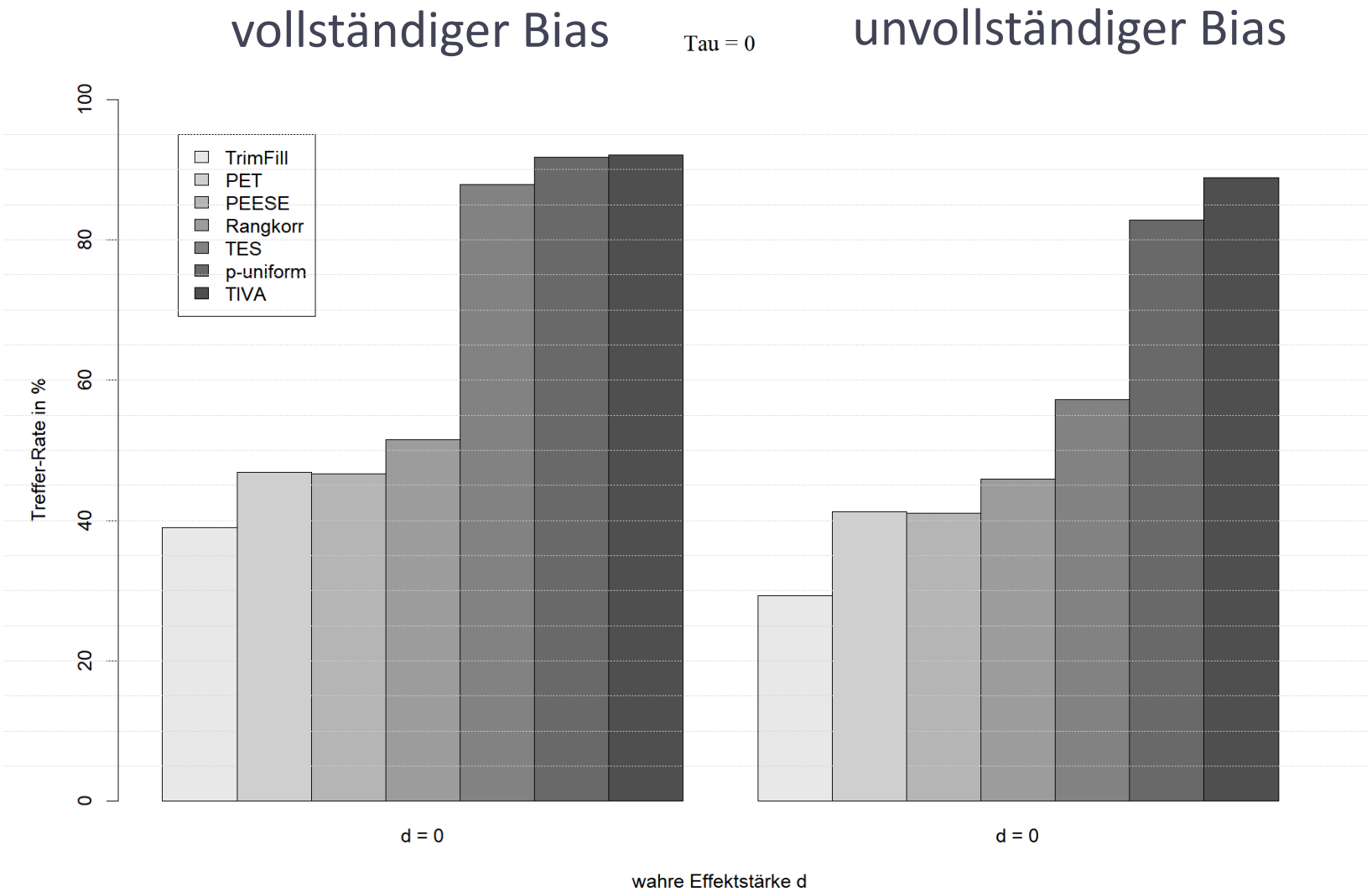
Falsch-positiv-Raten

- Begg‘s Rangkorrelation ist raus.
- Bei (nennenswerter) Heterogenität ist auch TES raus.

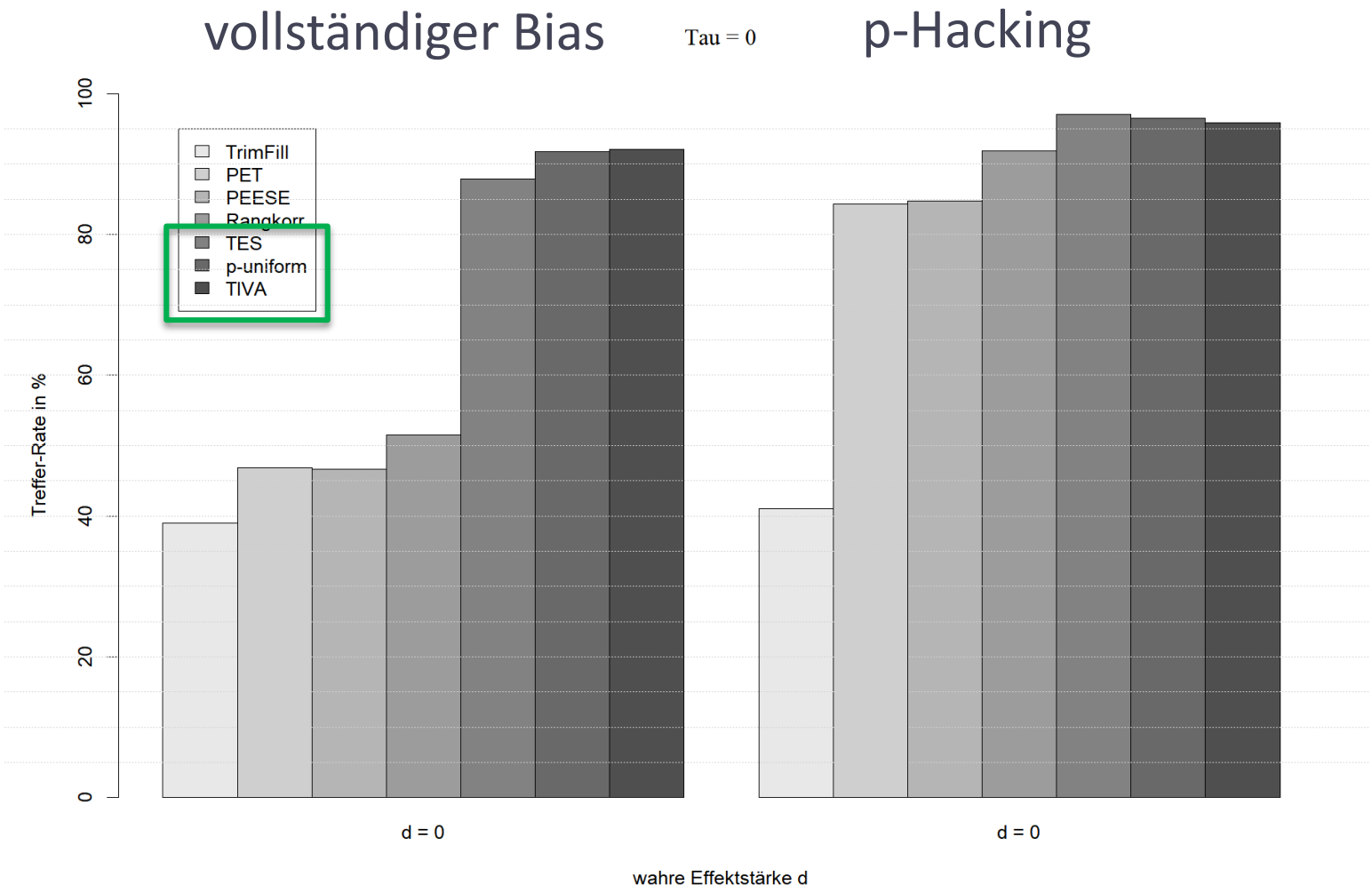
Power (bei vollständigem Publikationsbias)



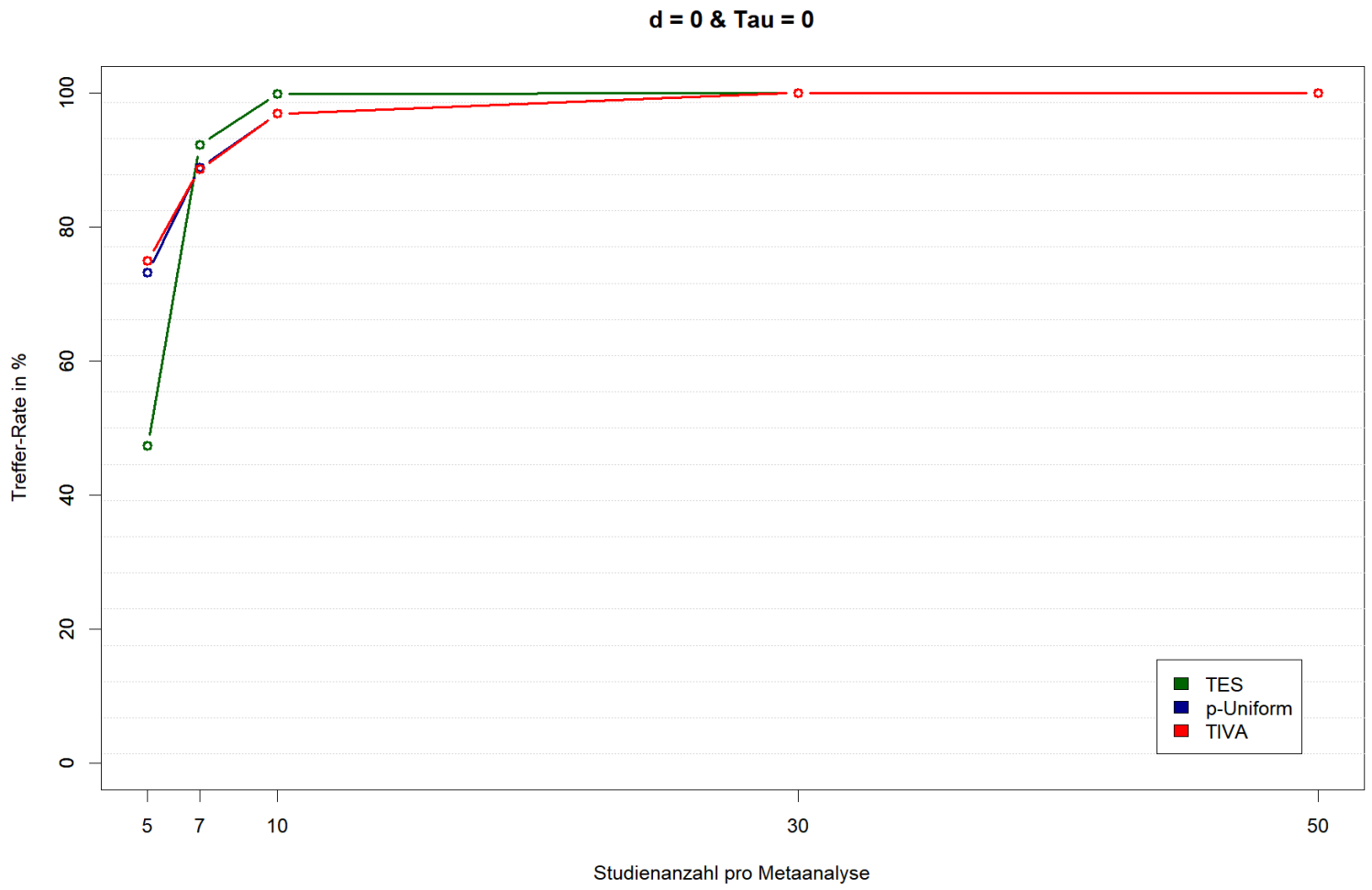
Power


vollständiger Bias
Tau = 0
unvollständiger Bias
TrimFill
PET
PEESE
Rangkorr
TES
p-uniform
TIVA
Treffer-Rate in %
0
20
40
60
80
100
d = 0
d = 0
wahre Effektstärke d

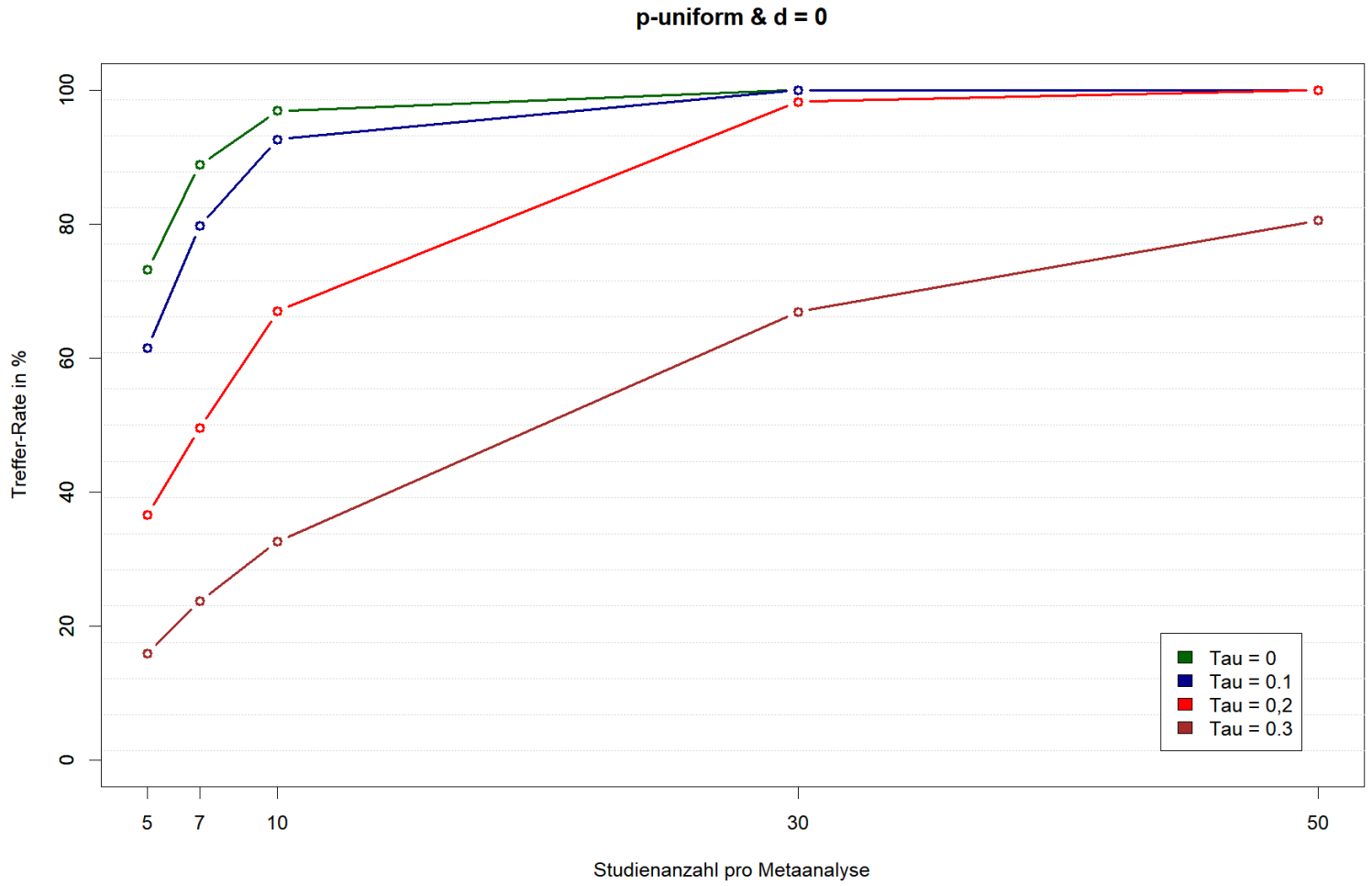
Power


vollständiger Bias
Tau = 0
p-Hacking
TrimFill
PET
PEESE
Rangkorr
TES
p-uniform
TIVA
Treffer-Rate in %
0
20
40
60
80
100
d = 0
d = 0
wahre Effektstärke d

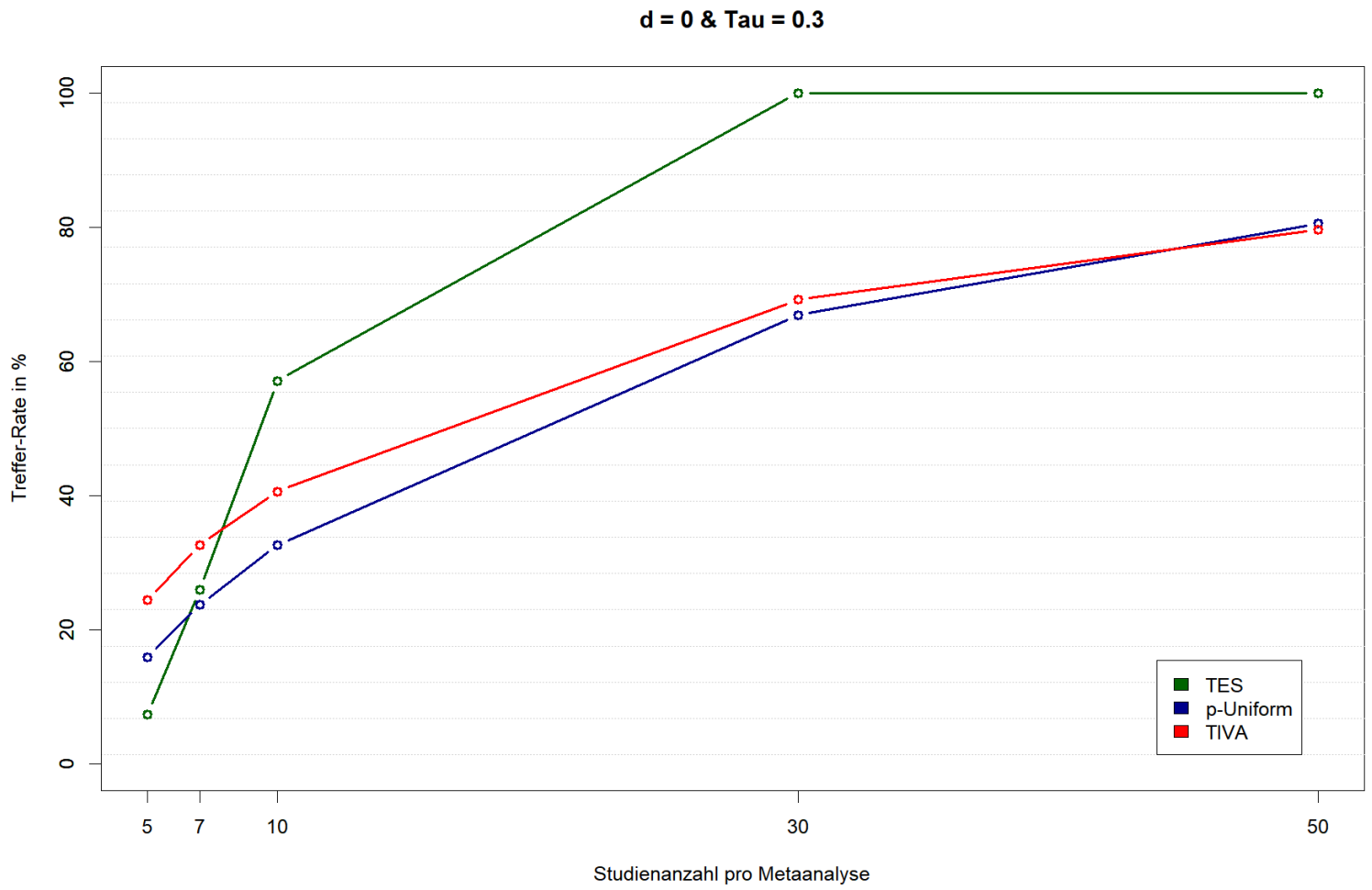
Power bei unterschiedlichen Studienzahlen (vollständiger Bias)



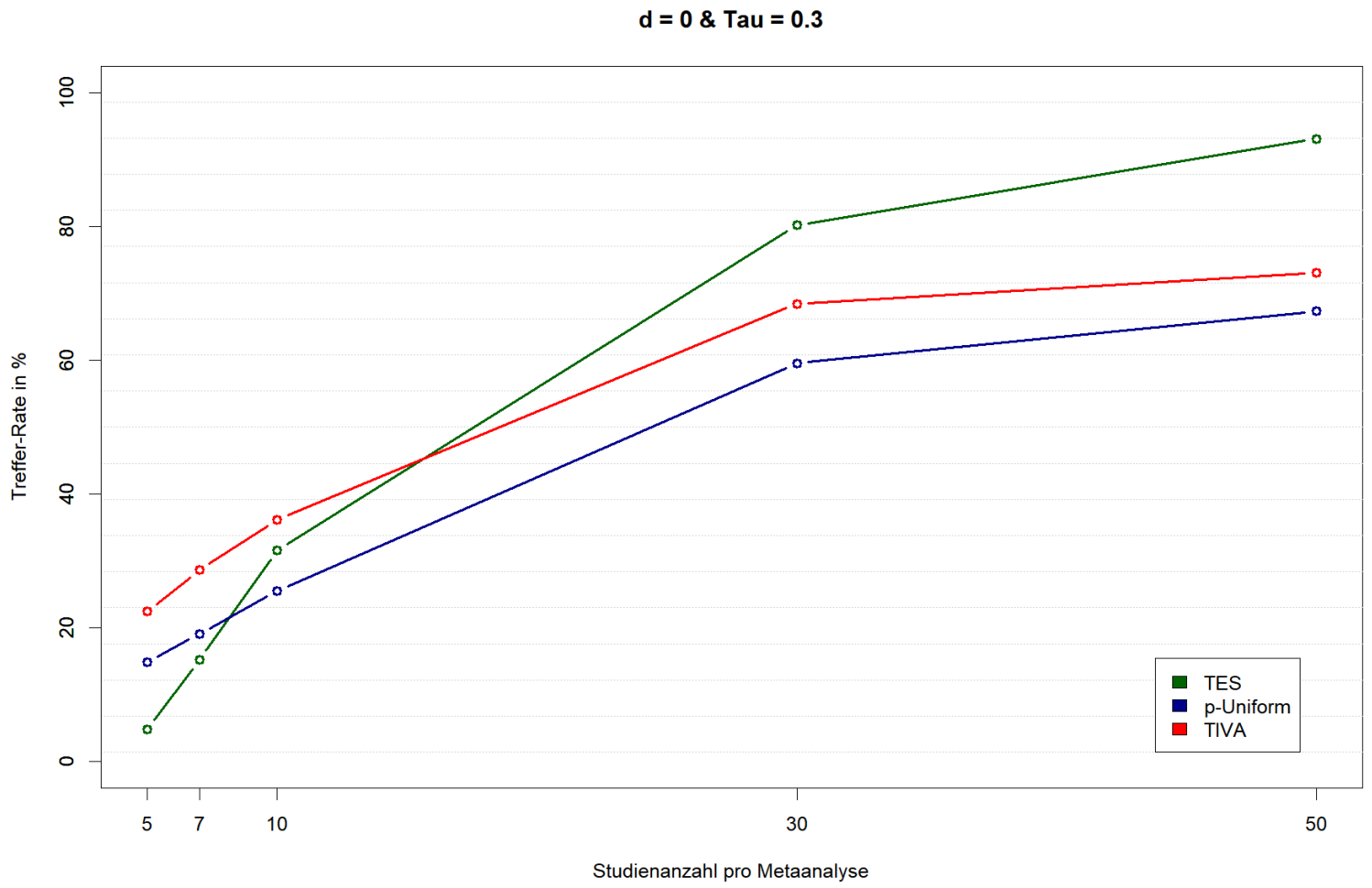
p-uniform: Power bei Heterogenität (vollständiger Bias)



Power bei Heterogenität (vollständiger Bias)



Power bei Heterogenität (unvollständiger Bias)

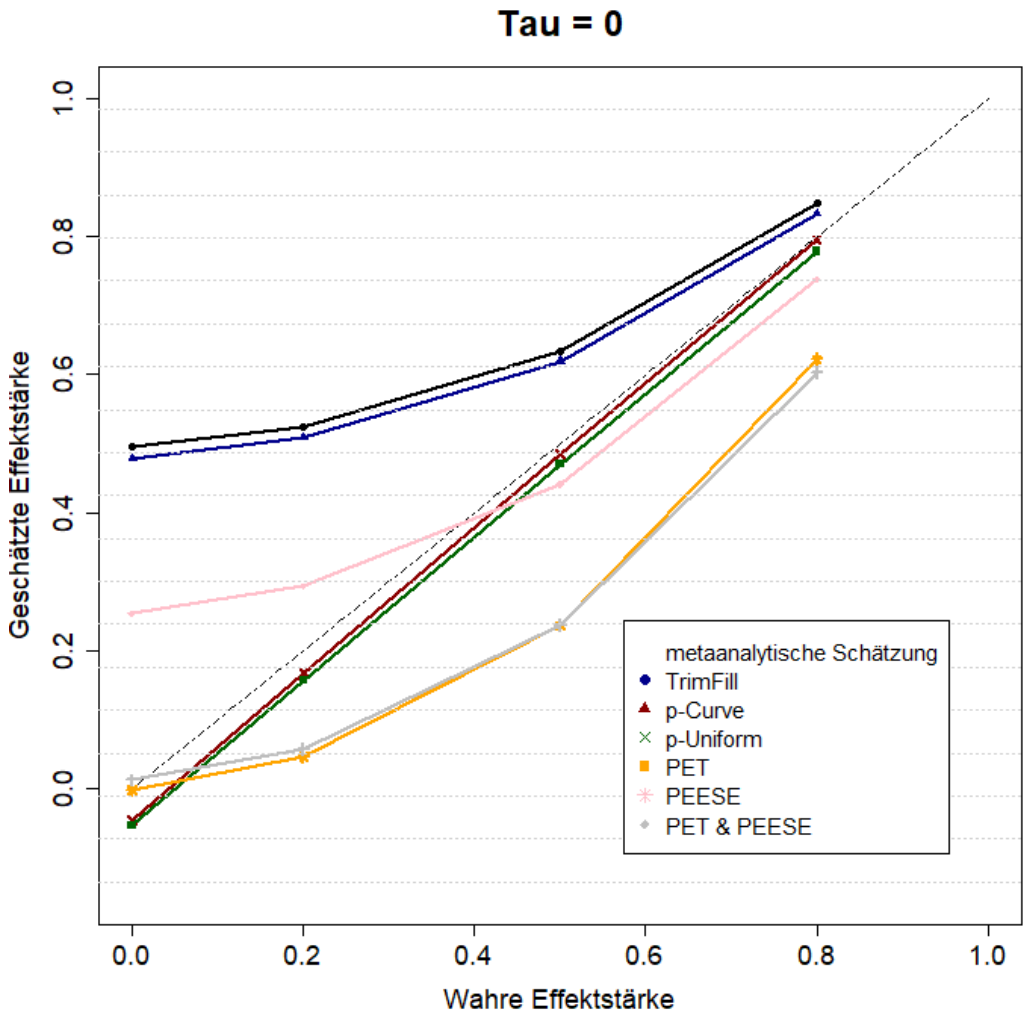


Können wir Biases aufdecken?

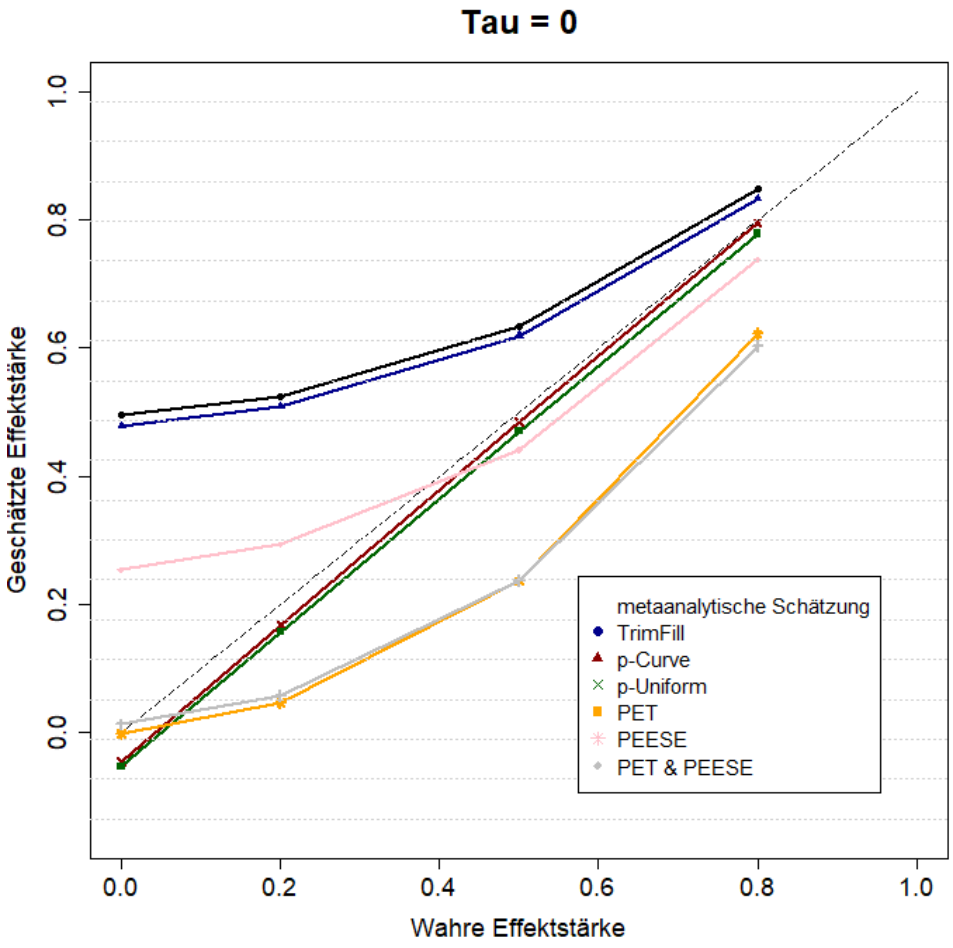
- JA! – Wenn einige Bedingungen erfüllt sind:
  - Geeignete Methoden – vorzugsweise *p*-uniform oder TIVA
    - Trim-and-Fill und Begg‘s Rangkorrelation sind generell ungeeignet, TES ist es bei Heterogenität.
  - Unter Homogenität können Biases selbst bei einer kleinen Studienzahl mit einer „vernünftigen“ Wahrscheinlichkeit entdeckt werden.
  - Heterogenität vermindert die Sensitivität (aber verstärkt Biases)!
    - Hinreichende Power erst mit deutlich mehr als 10 Studien.
  - Starke (und relevante) Biases

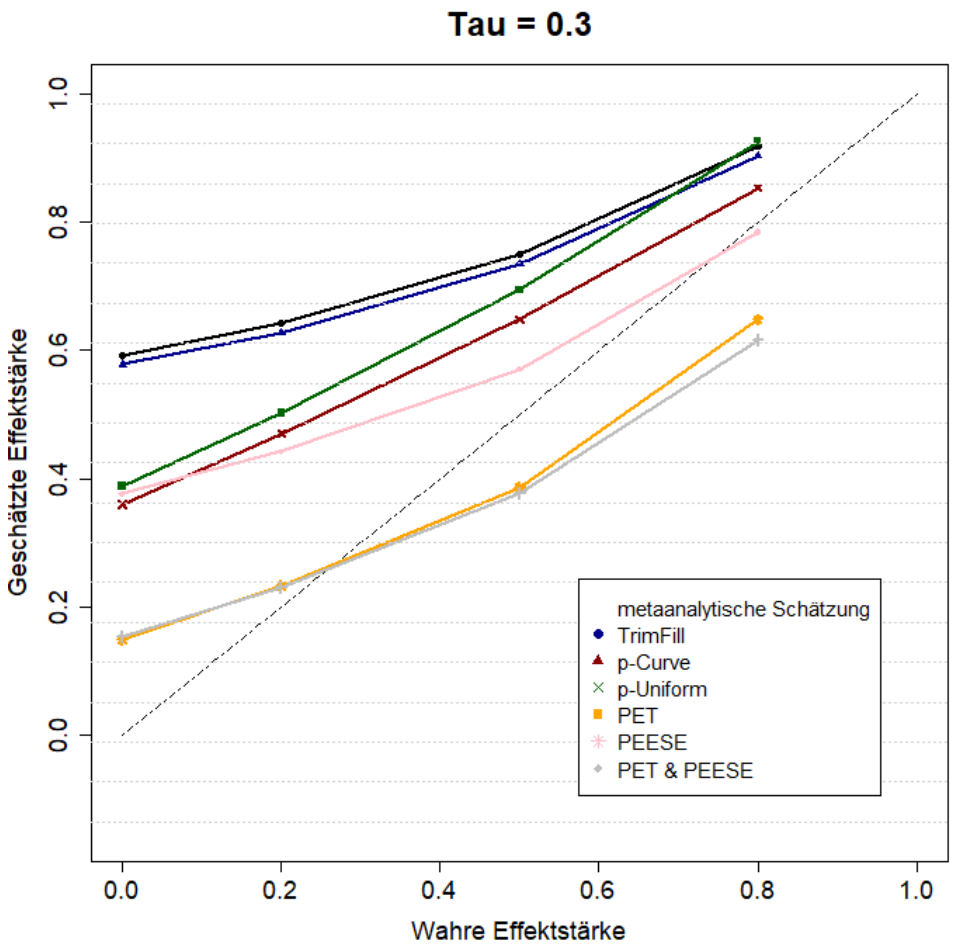
Können wir Biases korrigieren?

Publikationsbias: Korrigierte Schätzungen

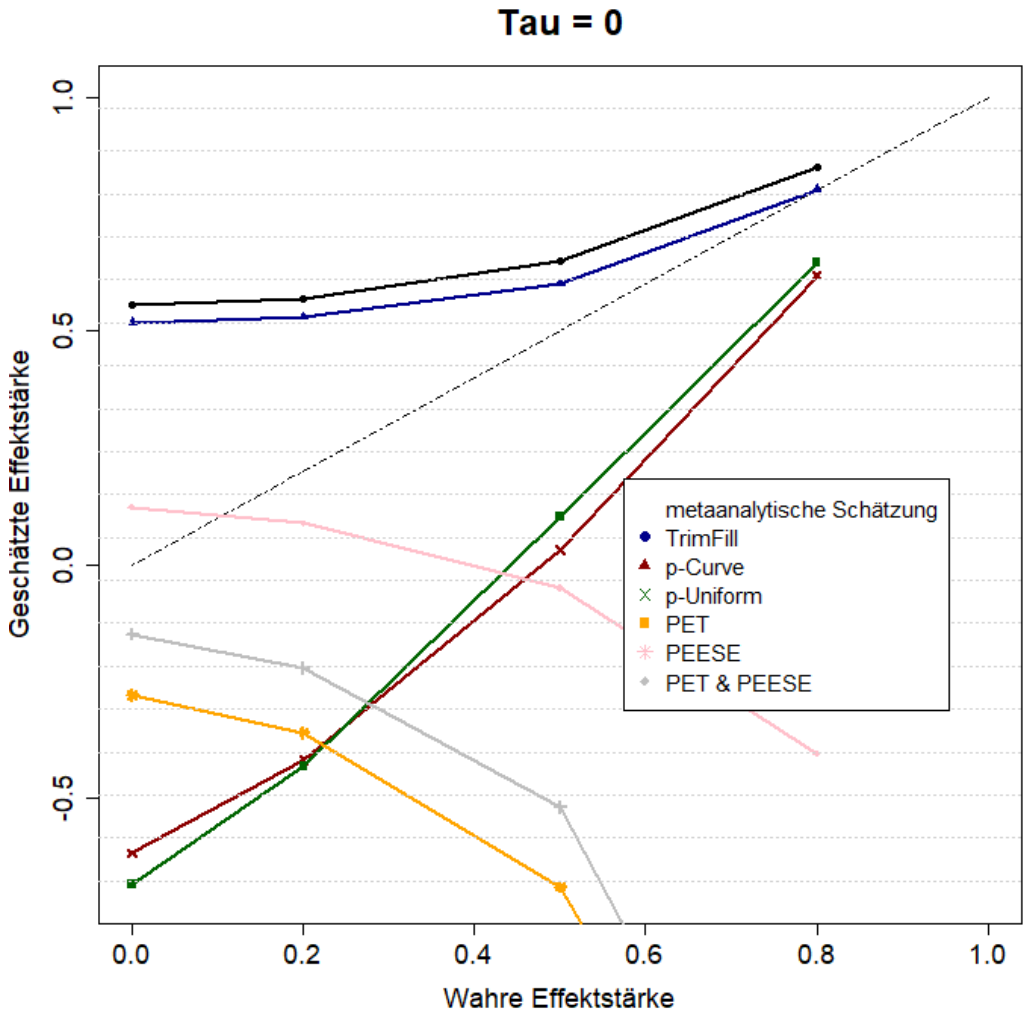


Publikationsbias: Korrigierte Schätzungen

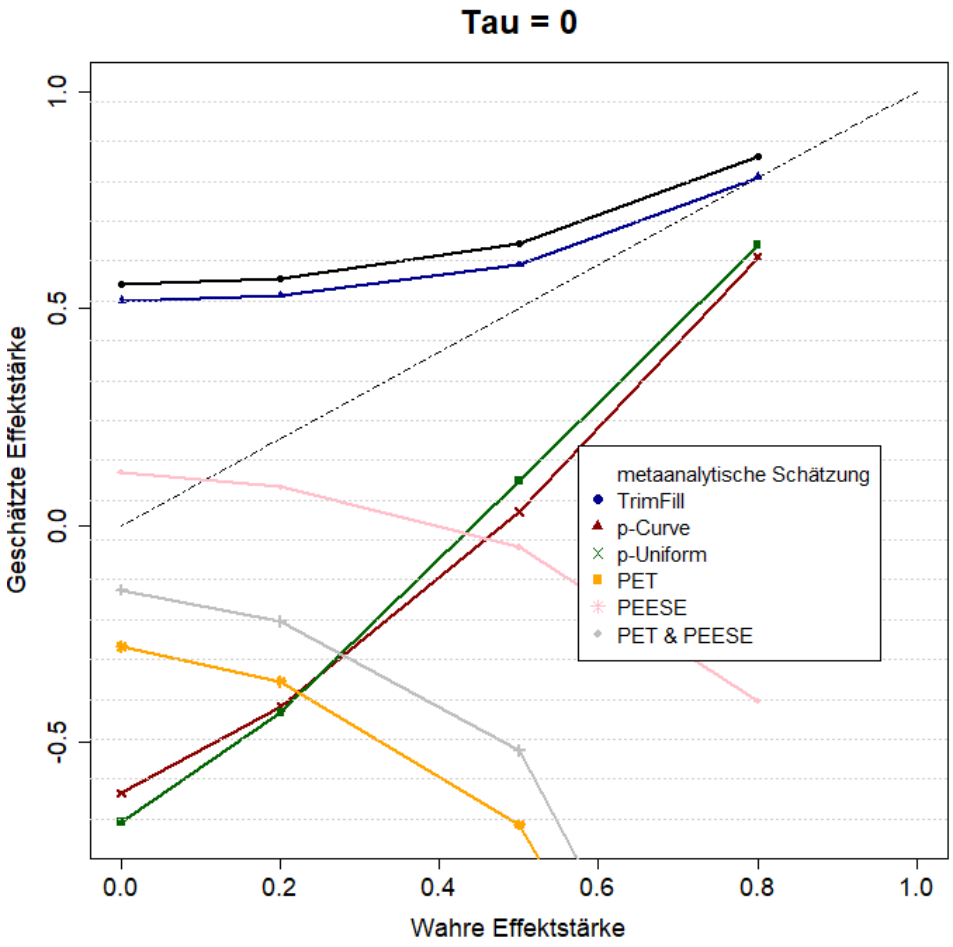


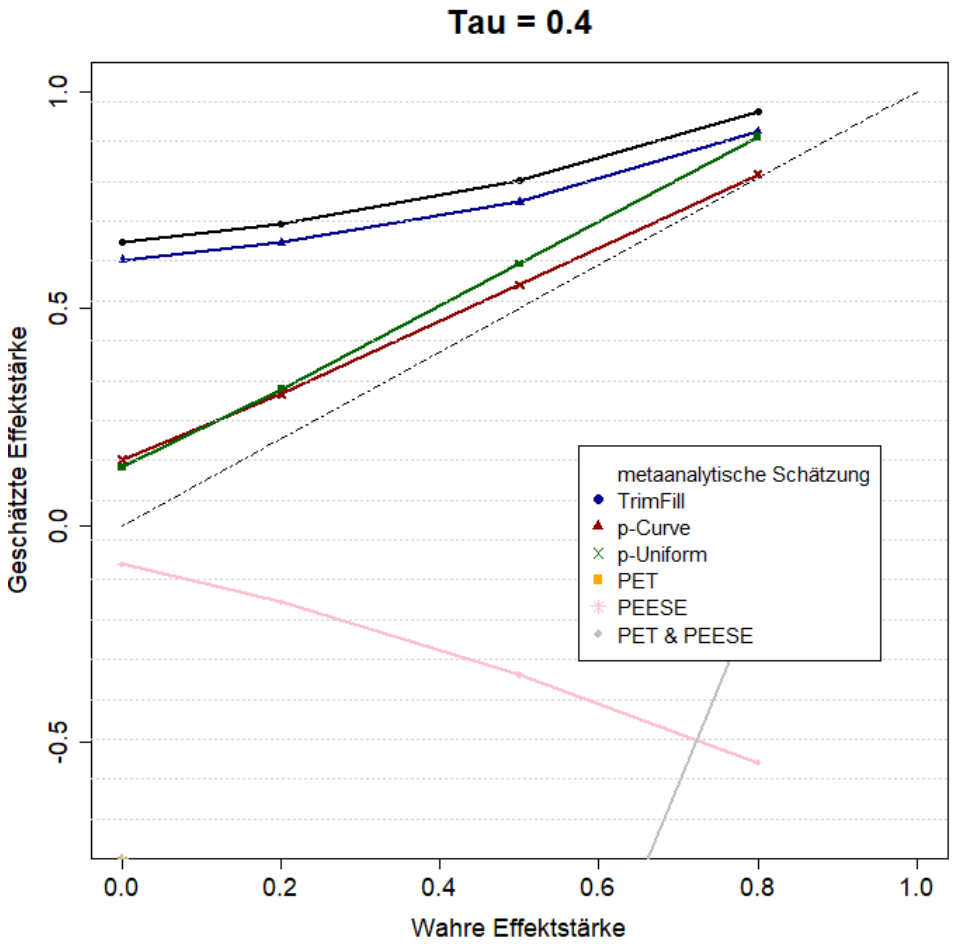


p-Hacking: Korrigierte Schätzungen



p-Hacking: Korrigierte Schätzungen





Können wir Biases korrigieren?

- NEIN!
- Ausnahme: Bei Homogenität und ohne p-Hacking liefern p-curve und p-uniform annähernd erwartungstreue Schätzungen. Diese Schätzungen sind allerdings unpräzise.
- Bei Heterogenität und mit p-Hacking sind auch diese Schätzungen stark verzerrt.
- Alle anderen (evaluierten) Verfahren liefern generell ungeeignete Schätzungen.

Diskussion

- Die Detektion von Biases kann gelingen.
  - Geeignete Verfahren: p-uniform, TIVA
  - Heterogenität erschwert die Detektion.
    - Ausreichende Studienzahl
- Die Korrektur von Biases ist ein aussichtsloses Unterfangen.
  - Homogenität und *kein* p-Hacking sind in metaanalytischen Datensätzen sicher die glückliche Ausnahme.

- In vielen Metaanalysen wird die einzige formal korrekte Schlussfolgerung lauten: „Dies ist verzerrte Evidenz“.

- Wir sollten bessere Forschung nach besseren Kriterien publizieren.
  - Eine zentrale Zutat: Replikationen, Replikationen, Replikationen…

Diskussion

- Die Verfahren liefern dort fehlerhafte Ergebnisse, wo die Annahmen ihres Bias- und Datenmodells verletzt sind.
  - Selektion nach Signifkanz statt ES (Trim-and-Fill).
  - Heterogenität
  - p-Hacking
- Bessere Bias- und Datenmodell könnten helfen.
- Selektionsmodelle (z.B. Hedges, 1984; Ulrich, 2018).
  - Theoretisch überlegen und positiv evaluiert (McShane et al., 2016; Carter et al., 2017)
  - Aber:
    - Große Anzahl an Studien erforderlich
    - Keine Modellierung von *p*-Hacking
    - Auch diese Modelle „scheitern“ an für sie ungünstigen Bedingungen.

Denn der radikalste Zweifel ist der Vater der Erkenntnis.

(Max Weber)

Danke!