3. Vorlesung Statistik II Publikationsbias, Researcher Degrees of Freedom und Open Science

Wiederholung Metaanalyse

- In der letzten Vorlesung haben wir mit der Metaanalyse eine Methode kennen gelernt, die es uns erlaubt, aus den Ergebnissen mehrerer (kleiner) Studien einen Hypothesentest mit höherer Power bzw. ein Konfidenzintervall mit geringerer erwarteter Länge als in den einzelnen Studien zu konstruieren.

- Annahme der Metaanalyse mit festen Effekten: $\hat{\theta}_j \overset{ind}{\sim} N\left(\theta, \sigma_j^2\right)$

- Punktschätzung: $\hat{\theta} = \frac{1}{\sum_{j=1}^{N} W_j} \sum_{j=1}^{N} W_j \cdot \hat{\theta}_j$

- Intervallschätzung: $\left[\hat{\theta} - z_{1-\frac{\alpha}{2}} \cdot \frac{1}{\sqrt{\sum_{j=1}^{N} W_j}}, \hat{\theta} + z_{1-\frac{\alpha}{2}} \cdot \frac{1}{\sqrt{\sum_{j=1}^{N} W_j}}\right]$

- Metaanalysen gelten eigentlich als Königsweg wissenschaftlicher Evidenz: Wenn Konfidenzintervalle oder Hypothesentests aus Metaanalysen mit einer großen Anzahl an Studien sowie einer hohen Gesamtstichprobengröße auf einen Effekt in der Population hindeuten, sollten wir uns eigentlich sehr sicher sein können, dass der Effekt tatsächlich existiert… **Oder?**

Datenqualität bei Metanalysen: „Garbage in - Garbage out“

- Wenn die Schätzungen $\hat{\theta}_j$ der einzelnen Studien aus methodischen Gründen verzerrt sind, kann dies auch durch eine Metaanalyse nicht so einfach ausgeglichen werden.

- Die der Metaanalyse mit festen Effekten zugrundeliegende Annahme $E(\hat{\theta}_j) = \theta$ für alle j ist in diesem Fall nicht erfüllt und die metaanalytischen Verfahren sind verzerrt.

- Daher ist es sehr wichtig, die einzelnen Studien kritisch zu betrachten, bevor man die Ergebnisse einer Metaanalyse interpretiert.

- Folge: Eine methodisch hochwertige Studie mit einer großen repräsentativen Stichprobe hat unter Umständen eine höhere Aussagekraft als eine Metaanalyse von vielen kleinen methodisch fragwürdigen Studien.

Plan heute

- Thema heute: Probleme, die das Vertrauen in metaanalytische Befunde in der Praxis stark einschränken können, sowie mögliche Lösungsansätze.

- Probleme von Metaanalysen:
  - (Heterogenität der Effekte)
  - Publikationsbias
  - Researcher Degrees of Freedom und P-Hacking

- Folgen: Replikationskrise in der Psychologie

- Lösungsansatz: Open Science
  - Präregistrierung
  - Veröffentlichung der Rohdaten, Versuchsmaterialien und Analyseskripten

Plan heute

- Thema heute: Probleme, die das Vertrauen in metaanalytische Befunde in der Praxis stark einschränken können, sowie mögliche Lösungsansätze.

- **Probleme von Metaanalysen:**
  - **(Heterogenität der Effekte)**
  - **Publikationsbias**
  - **Researcher Degrees of Freedom und P-Hacking**

- Folgen: Replikationskrise in der Psychologie

- Lösungsansatz: Open Science
  - Präregistrierung
  - Veröffentlichung der Rohdaten, Versuchsmaterialien und Analyseskripten

Heterogenität der Effekte I

- Die in der letzten Vorlesung besprochene Metaanalyse mit festen Effekten nimmt an, dass in jeder der N Studien der gleiche Effekt untersucht (und geschätzt) wird:

  $$E(\hat{\theta}_j) = \theta$$

  für alle Studien j.

- Dazu haben wir im Übungsblatt 2 gesehen, dass die Metaanalyse mit festen Effekten zum gleichen Ergebnis kommt, wie wenn man die Rohdaten aller Einzelstudien kombinieren und ein Konfidenzintervall bzw. Hypothesentests mit den kombinierten Daten berechnen würde.

- In der Praxis wird jedoch häufig in jeder Studie ein etwas anderer Effekt untersucht:

  $$E(\hat{\theta}_j) = \theta_j$$

- In solchen Fällen spricht man auch von **heterogenen Effekten**.

Heterogenität der Effekte II

- Mögliche Gründe für Heterogenität am Beispiel aus der letzten Vorlesung: Unterschied zwischen Verhaltenstherapie und Gestalttherapie bei Zwangsstörungen
- Die Studien untersuchen verschiedene Populationen, in denen sich die Wirksamkeit der beiden Therapien leicht unterscheidet.
  - Kulturelle oder regionale Unterschiede
  - Sozioökonomischer Status, Bildungsniveau
  - Schweregrad der Zwangsstörung, Vorliegen zusätzlicher Störungen
- Unterschiede in der Durchführung der beiden Therapieformen:
  - Ausbildung der Therapeutinnen, therapeutisches Setting
  - „Verhaltenstherapie ist nicht gleich Verhaltenstherapie“
- Unterschiede in der untersuchten abhängigen Variable:
  - Verschiedene Messinstrumente zur Erhebung der Zwangssymptomatik
  - Messung leicht unterschiedlicher psychologischer Variablen

Heterogenität der Effekte III

- In allen Szenarien haben die wahren Effekte θ_j der N Studien zwar viel gemeinsam, sind aber nicht identisch: Eventuell ist Verhaltenstherapie im Durchschnitt effektiver als Gestalttherapie, aber manchmal ist der Unterschied größer und manchmal kleiner.
- Heterogene Effekte sind nicht unbedingt ein Problem, eher im Gegenteil:
  Der Befund eines beobachteten Vorteils der Verhaltenstherapie gegenüber der Gestalttherapie wäre viel überzeugender, wenn er in verschiedenen Populationen und unter einer Vielzahl an unterschiedlichen Bedingungen nachgewiesen werden kann.

- Aber: Der „durchschnittliche Unterschied“ zwischen Verhaltenstherapie und Gestalttherapie bei Zwangsstörungen im Falle heterogener Effekte muss mit einer **Metaanalyse mit zufälligen Effekten** untersucht werden, die methodisch komplizierter ist und deren Grundlagen erst im Masterstudium gelehrt werden.
- In der Praxis ist meist eine Metaanalyse mit zufälligen Effekten notwendig.
  Die Interpretation der KIs bzw. Hypothesentests funktioniert aber sehr ähnlich.

Ausgangssituation in der Psychologie

- Zur Erinnerung: In psychologischen Fachzeitschriften werden fast ausschließlich Studien mit signifikanten Ergebnissen veröffentlicht. Im Fachbereich Psychologie und Psychiatrie berichten mehr als 90% der publizierten Artikel signifikante Ergebnisse (Fanelli, 2011).
- Falls alle diese Studien ein geringes Signifikanzniveau und eine hohe Power aufweisen würden, wäre dies prinzipiell nicht so schlimm (vorausgesetzt natürlich, dass die Studien methodisch sauber sind und dass H_1-Entscheidungen tatsächlich relevanter als H_0-Entscheidungen sind).
- Die durchschnittliche Power in psychologischen Studien wird jedoch nur auf ungefähr 35% geschätzt (Bakker et al., 2012). Wie lassen sich die vielen signifikanten Ergebnisse erklären? Wir müssen davon ausgehen, dass der Anteil der falschen Ergebnisse unter den veröffentlichten Ergebnissen sehr hoch ist.

- Können metaanalytische Methoden dieses Problem lösen?

Publikationsbias

- Antwort: Nein.

- Übertriebenes Beispiel: Falls ausschließlich signifikante Studien publiziert werden, führt dies zu einer Verzerrung der Ergebnisse der Metaanalysen.

- Intuitive Erklärung: Falls wir aus allen Studien diejenigen entfernen, die nicht signifikant geworden sind, bleiben nur noch die Studien mit großen Schätzwerten für θ übrig. Falls wir auf der Basis dieser Schätzwerte einen kombinierten Schätzwert berechnen, ist dieser nach oben verzerrt.

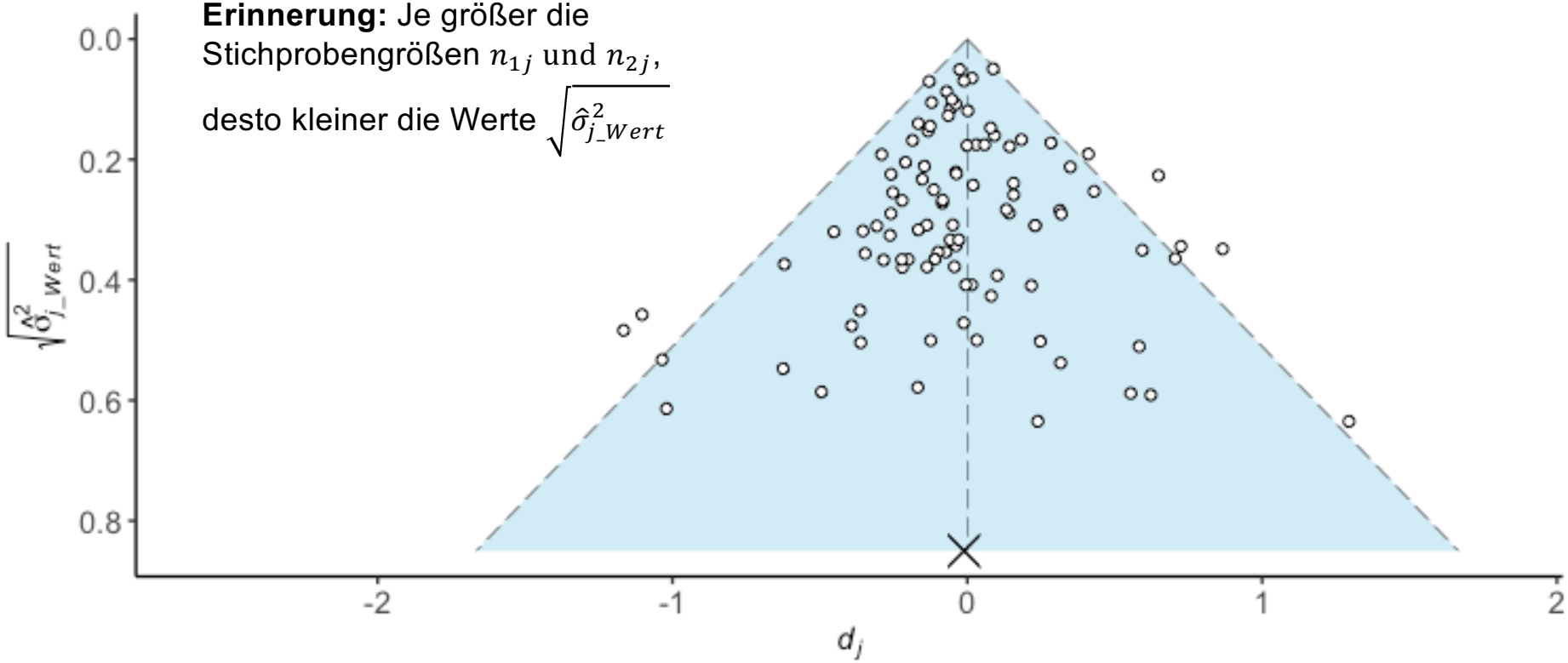
- Dieses Problem wird **Publikationsbias** genannt.

Mögliche Gründe für Publikationsbias

- Studien mit signifikanten Ergebnissen haben generell eine höhere Chance, in wissenschaftlichen Zeitschriften veröffentlich zu werden. Nicht in wissenschaftlichen Zeitschriften veröffentlichte Studien, erscheinen häufig gar nicht in der zugänglichen wissenschaftlichen Literatur. Sie landen im sogenannten „Filedrawer“ (deutsch: „in der Schublade“).
- Studien mit signifikanten Ergebnissen werden häufiger von den Editorinnen und den externen Gutachterinnen der Zeitschriften zur Veröffentlichung freigegeben. Gründe:
  - Nicht signifikante Studien werden als weniger „interessant“ eingestuft.
  - Nicht signifikante Studien widersprechen häufig etablierten Theorien.
  - Nicht signifikante Studien werden seltener zitiert (schlecht für die Zeitschrift).
- Studien mit nicht signifikanten Ergebnissen werden seltener eingereicht. Gründe:
  - Andere Manuskripte haben eine höhere Chance auf Veröffentlichung.
  - Nicht signifikante Studien widersprechen häufig der Theorie der Autor*in und werden aus psychologischen oder strategischen Gründen nicht weiterverfolgt.

Funnelplot ohne Publikationsbias

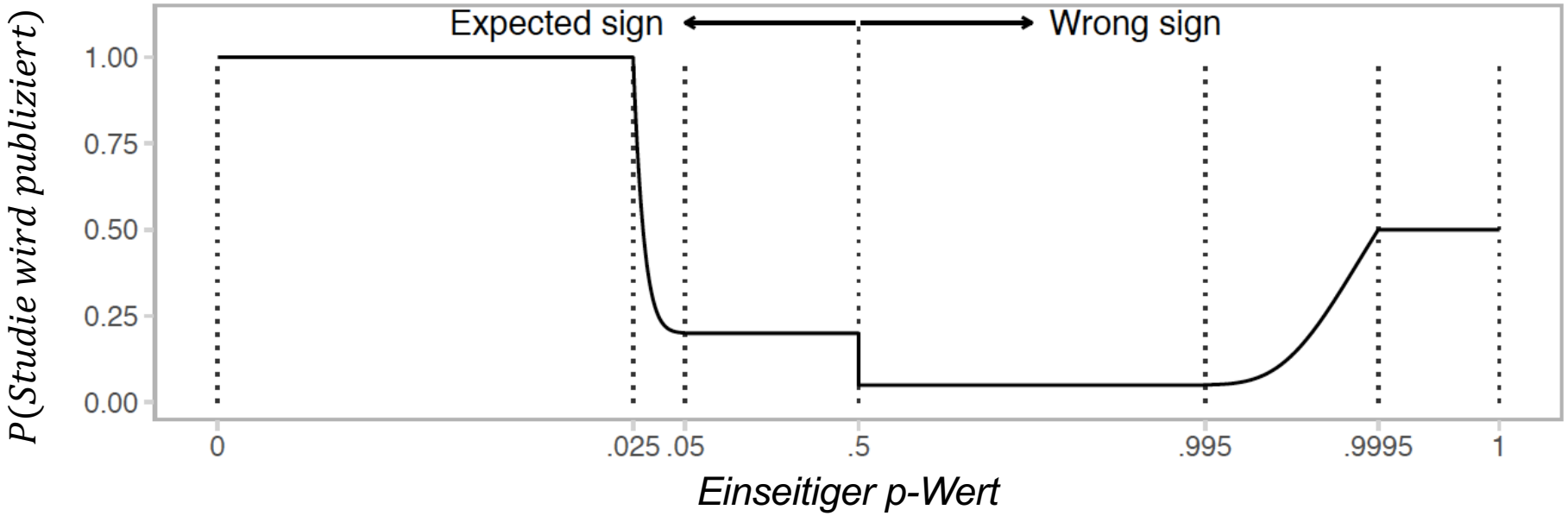
Erinnerung: Je größer die Stichprobengrößen n_{1j} und n_{2j}, desto kleiner die Werte $\sqrt{\hat{\sigma}^2_{j\_Wert}}$



Simulierte Metaanalyse mit $N = 100$ Studien und $\delta = 0$ (ein Punkt ist eine Studie):

- X-Achse zeigt die Schätzwerte für δ aus den N Studien: d_j
- Y-Achse zeigt die geschätzten Standardfehler: $\widehat{SE}(D_j)_{Wert} = \sqrt{\hat{\sigma}^2_{j\_Wert}}$
- Blauer Bereich gilt: $p \geq \alpha = 0.05$ (t-Test unabhängige Stichproben, ungerichtete H_0)

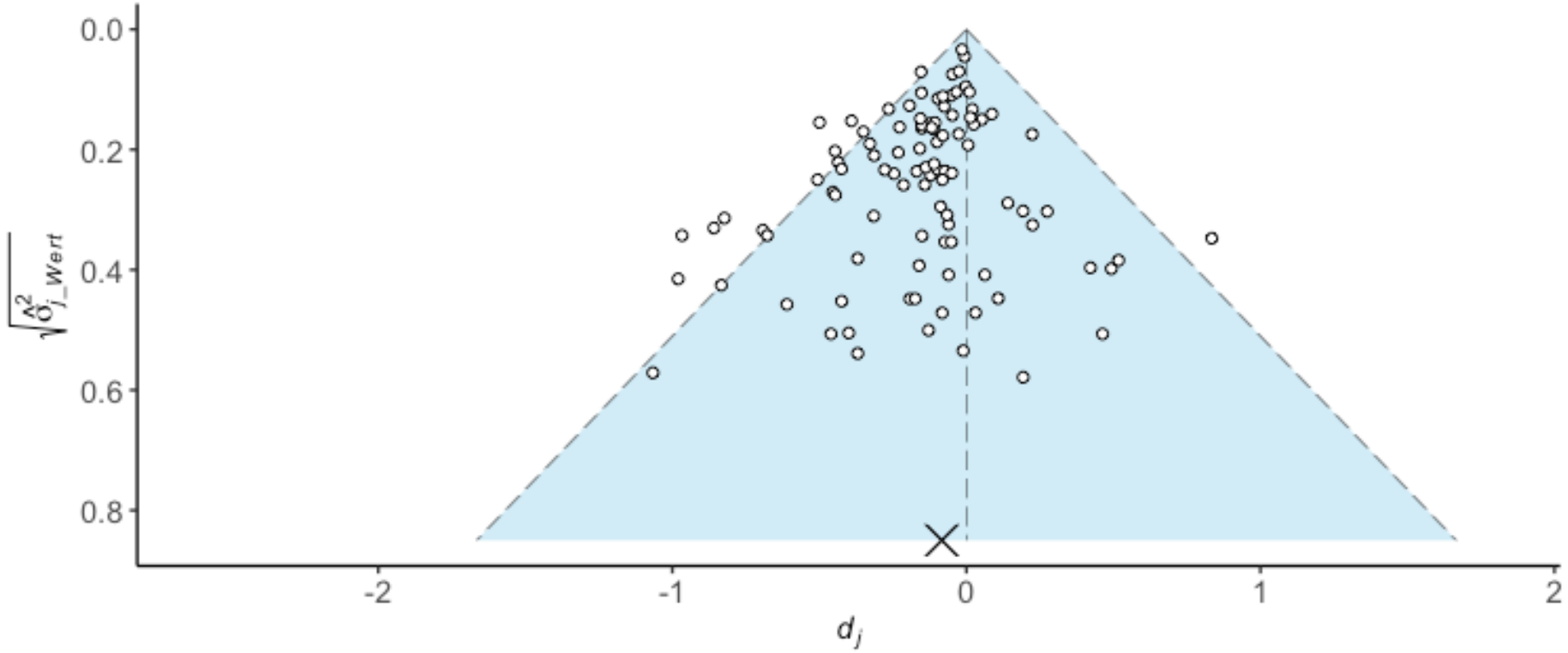
Exkurs: Simulation von Publikationsbias



Annahmen in Carter, Schönbrodt, Gervais & Hilgard, 2019:

- Es gibt eine „Wunschrichtung“ für den Effekt (hier $\delta < 0$), trotzdem wird ein t-Test mit ungerichteten Hypothesen und $\alpha = 0.05$ durchgeführt.
- Signifikante Ergebnisse mit Punktschätzer ($d_j < 0$) in der gewünschten Richtung (entspricht rechtsseitigem p-Wert < 0.025) werden auf jeden Fall publiziert.
- Nicht signifikante Effekte werden mit niedrigerer Wahrscheinlichkeit publiziert.
- Stark signifikante Ergebnisse in die „falsche“ Richtung werden mit mittlerer Wahrscheinlichkeit publiziert.

Funnelplot mit Publikationsbias



Simulierte Metaanalyse mit $N = 100$ Studien, $\delta = 0$ und Publikationsbias:

- Mehr publizierte Studien mit $d_j < 0$ (in „Wunschrichtung“)
- Mehr signifikante Studien (in „Wunschrichtung“)
- Unterschätzung von δ durch die Metaanalyse:
  Verzerrtes KI und erhöhte Wahrscheinlichkeit für Fehler 1. Art in der Metaanalyse

Researcher Degrees of Freedom und P-Hacking

- Es gibt Methoden, um das Ausmaß des Publikationsbias abzuschätzen und im Rahmen der Metaanalyse zu berücksichtigen. Leider funktionieren diese Methoden nicht besonders gut (Carter et al., 2019).
- Das Problem verfälschter Metaanalysen wird häufig noch dadurch verschärft, dass die publizierten signifikanten Effekte in einigen Fällen noch zusätzlich (im Betrag) nach oben verzerrt sind.

- Häufiger Grund: Sogenannte **Researcher Degrees of Freedom**

- Werden Researcher Degrees of Freedom ausgenutzt, um die Chance auf ein signifikantes Testergebnis zu steigern, spricht man auch von **P-Hacking**.
- In Verbindung mit kleinen Stichproben und einem Signifikanzniveau von $\alpha = 0.05$ ist es so sehr leicht, signifikante Ergebnisse zu bekommen, obwohl der Effekt eigentlich gleich Null ist (Simmons et al., 2011).

Researcher Degrees of Freedom (RDoF)

- Als **Researcher Degrees of Freedom** (RDoF) werden Entscheidungen im Forschungsprozess bezeichnet, die notwendig, aber gleichzeitig inhaltlich oder methodisch nicht eindeutig begründbar sind und daher bis zu einem gewissen Punkt willkürlich getroffen werden müssen (Wicherts et al., 2016).

- Welche Entscheidung (aus vielen möglichen) getroffen wird, kann die Ergebnisse statistischer Verfahren und damit die Interpretation von Studien beeinflussen.
- Befragungen legen nahe, dass RDoF häufig bei fragwürdigen Forschungspraktiken eine Rolle spielen (John et al., 2012, Wang et al., 2018). Allerdings ist es nicht leicht abzuschätzen, wie viele Wissenschaftlerinnen RDoF absichtlich ausnutzen.

- Wichtige Erkenntnis: RDoF sind nicht nur ein Problem, falls Sie von Betrügerinnen absichtlich ausgenutzt werden. Auch viele rechtschaffene Wissenschaftlerinnen wissen eventuell nicht, dass einige ihrer Forschungspraktiken problematisch sind und zur hohen FDR in der psychologischen Forschung beitragen können.

RDoF: Planung und Durchführung von Studien

- Untersuchen mehrerer (ähnlicher) experimenteller Bedingungen.
- Erheben mehrerer abhängiger Variablen, oder Messung der abhängigen Variable mit mehreren Messinstrumenten.
- Erheben von Kontrollvariablen, die zur Untersuchung der Forschungsfrage in Subgruppen oder zur Festlegung von Ausschlusskriterien genutzt werden können.
- Keine Festlegung der geplanten Stichprobengröße, der genauen Population, des Erhebungszeitraums, des Erhebungsorts, der Art der Stichprobenziehung, sowie genauer Kriterien, unter welchen Umständen die Datenerhebung beendet ist.
- Keine Verblindung von Probandinnen und Versuchsleiterinnen hinsichtlich der untersuchten Hypothesen sowie der Zugehörigkeit zur Experimental- bzw. Kontrollgruppe.
- Nicht zufällige Zuteilung der Probandinnen zu experimentellen Bedingungen.
- Datengetriebene Veränderung von Daten während Erhebung oder Dateneingabe.
- Datengetriebenes Beenden der Datenerhebung („Optional Stopping“).

RDoF: Auswertung von Studien

Datengetriebene/r…

- Umgang mit fehlenden Werten.
- Umgang mit Ausreißern.
- Ausschluss von Probandinnen.
- Vorverarbeitung der Daten für inferenzstatistische Analysen.
- Auswahl statistischer Modelle, Hypothesentests, Schätzmethoden, Softwarepakete.
- Veränderung der statistischen Hypothesen oder des Signifikanzniveaus.
- Kodierung von Variablen.
- Kombination von experimentellen Bedingungen.
- Auswahl von unabhängigen, abhängigen oder Kontrollvariablen.
- Umgang mit Verletzungen inferenzstatistischer Annahmen.

RDoF: Interpretation und Veröffentlichung von Studien

- Selektive Beschreibung von Entscheidungen bei der Planung, Durchführung, Auswertung, oder Interpretation.
- Nicht berichten von nicht signifikanten Bedingungen, bzw. inkonsistenten Ergebnissen, die nicht zu etablierten Theorien oder zum Narrativ der Studie passen.
- Nicht berichten von bekannten Limitationen der Studie.
- Falsche Darstellung der ursprünglichen Fragestellung oder Hypothesen.
- Präsentieren einer explorativen Fragestellung als konfirmativ.
- Gezieltes zitieren oder nicht zitieren von Forschungsergebnissen aus der Literatur zur Begründung datengetriebener Entscheidungen.
- Bekräftigung der Aussagekraft der Studie mit Post-Hoc Poweranalysen.
- Kausale Interpretation von korrelativen Zusammenhängen.
- Keine Veröffentlichung der Rohdaten, des Studienmaterials oder der Analysecodes, die zur Überprüfung der Ergebnisse oder für eine Replikation notwendig sind.

Exkurs: Datengetriebene Entscheidungen als Multiples Testproblem

- Bei datengetriebenen, in Abhängigkeit von statistischen Analysen getroffenen Entscheidungen im Studienprozess bestehen Ähnlichkeiten zu dem Szenario nicht zusammengesetzter Hypothesentests aus Vorlesung 2 (Jelly Beans Beispiel):
- Angenommen, man analysiert die Daten einer Studie auf mehrere verschiedene Arten (siehe RDoF) und eine oder mehrere (aber nicht alle) der durchgeführten Hypothesentests werden signifikant. Meist gibt es keine Möglichkeit zu entscheiden, ob die signifikante Analyse „richtig“ ist. Je mehr Analysen durchgeführt werden, desto höher ist die Wahrscheinlichkeit mindestens einen Fehler 1. Art zu begehen.

- Naive Idee: Korrektur der FWER mit der Bonferroni Korrektur.
- Problem: Die Gruppe der Hypothesentests, für die korrigiert werden soll, ist nicht klar. Man müsste eigentlich nicht nur für die Tests kontrollieren, die tatsächlich durchgeführt wurden, sondern auch für alle Tests, die man eventuell durchgeführt hätte, falls sich die Daten (per Zufall) in anderen Werten realisiert hätten.
  -> „The garden of forking paths“ (Gelman & Loken, 2013)

Optional Stopping I

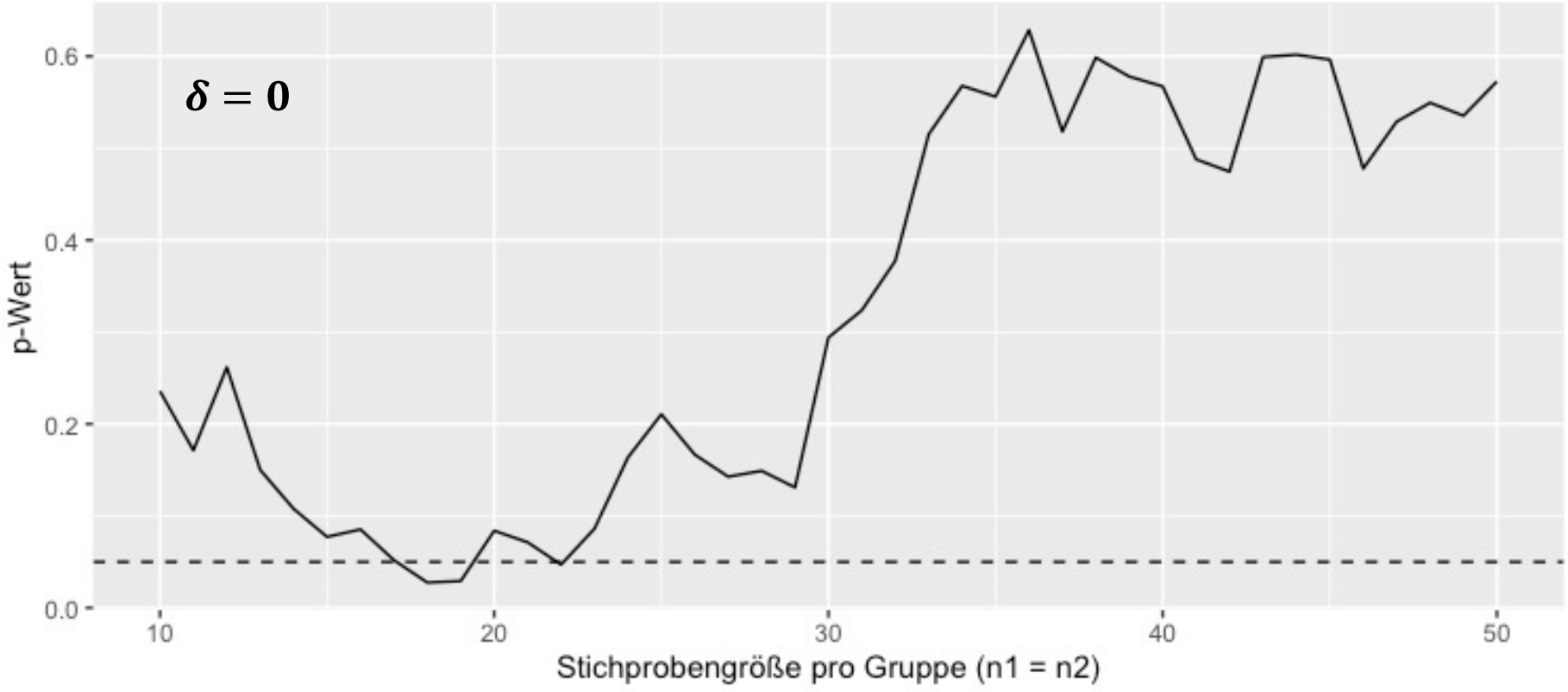
- Ein Beenden der Datenerhebung abhängig vom Ergebnis eines vorzeitigen Hypothesentests (seltener KI) bezeichnet man als **Optional Stopping**. Falls der Test signifikant ist, wird die Datenerhebung gestoppt. Falls der Test nicht signifikant ist, werden zusätzliche Probandinnen erhoben, erneut getestet und so weiter ...

- Verbreiteter Denkfehler: „Ist ein Test (knapp) nicht signifikant, könnte dies daran liegen, dass die Power der Studie zu gering ist. Erhebt man mehr Probandinnen, wird der p-Wert kleiner, falls die H_1 gilt, oder größer, falls die H_0 gilt."

- Problem: Falls die Nullhypothese gilt, sind alle p-Werte zwischen 0 und 1 gleich wahrscheinlich. Mit jeder neuen Proband*in „wandert" der p-Wert völlig zufällig zwischen 0 und 1 umher. Führt man wiederholt Tests nach erneuter Datenerhebung durch, steigt die Chance, zufällig einen Wert im signifikanten Bereich zu beobachten.

  -> Optional Stopping erhöht die Wahrscheinlichkeit für einen Fehler 1. Art!

Optional Stopping II



- Simulation aus Simmons et al., 2011: T-Test (unabhängige Stichproben, ungerichtet) mit n1 = n2 = 10 und $\boldsymbol{\delta = 0}$. Falls nicht signifikant ($\alpha = 0.05$), Erhebung einer weiteren Person pro Gruppe und erneute Testung, bis der Test signifikant ist oder n1 = n2 = 50 erreicht. **-> Wahrscheinlichkeit für einen Fehler 1. Art** $\approx \mathbf{22\%}$
- Grafik zeigt **einen möglichen** Verlauf der p-Werte. Obwohl der p-Wert zwischenzeitlich für einige Tests signifikant ist, steigt er danach wieder an.

Sequentielle Analysen

- In manchen Situationen ist es wünschenswert, vorzeitige Datenanalysen zu erlauben. Beispiel: In einer klinischen Studie kann es ethisch nicht vertretbar sein, weiteren Probandinnen in der Kontrollgruppe die beste Therapie vorzuenthalten oder den Nebenwirkungen der Vergleichstherapie auszusetzen, wenn man sich schon nach einem Teil der Erhebung relativ sicher ist, dass die neue Therapie überlegen ist.

- Für solche Fälle gibt es spezielle statistische Methoden für **sequentielle Analysen**, die wiederholtes Testen erlauben, ohne die Wahrscheinlichkeit für einen Fehler 1. Art zu erhöhen (Lakens, 2014). Diese Verfahren werden hier nicht besprochen (kompliziert).

- Eine einfache (eher schlechte) Alternative wäre auch hier die Bonferroni Korrektur: Sequentielle Tests sind voneinander abhängig, da bereits erhobene Daten auch für alle weiteren Tests verwendet werden und somit die Tests nicht unabhängig sind. Dies ist für die Bonferroni Korrektur kein Problem (siehe Vorlesung 1). Wird allerdings in sehr kurzen Abständen erneut getestet, ist die Abhängigkeit sehr hoch und die Power wird durch die Korrektur stark reduziert.

Sequentielle Analysen: Bonferroni

- Korrigiert man das gewünschte Signifikanzniveau mit der maximal geplanten Anzahl an Tests mithilfe der Bonferroni Methode, wird die FWER eingehalten.
- Sehr wichtig: Es muss vor der Durchführung der Studie genau festgelegt werden, nach wie vielen Probandinnen jeweils getestet wird, wie viele Tests maximal durchgeführt werden und nach wie vielen Probandinnen die Erhebung auf jeden Fall gestoppt wird.

Einfaches Beispiel: Zweistichproben t-Test (unabhängige Stichproben)

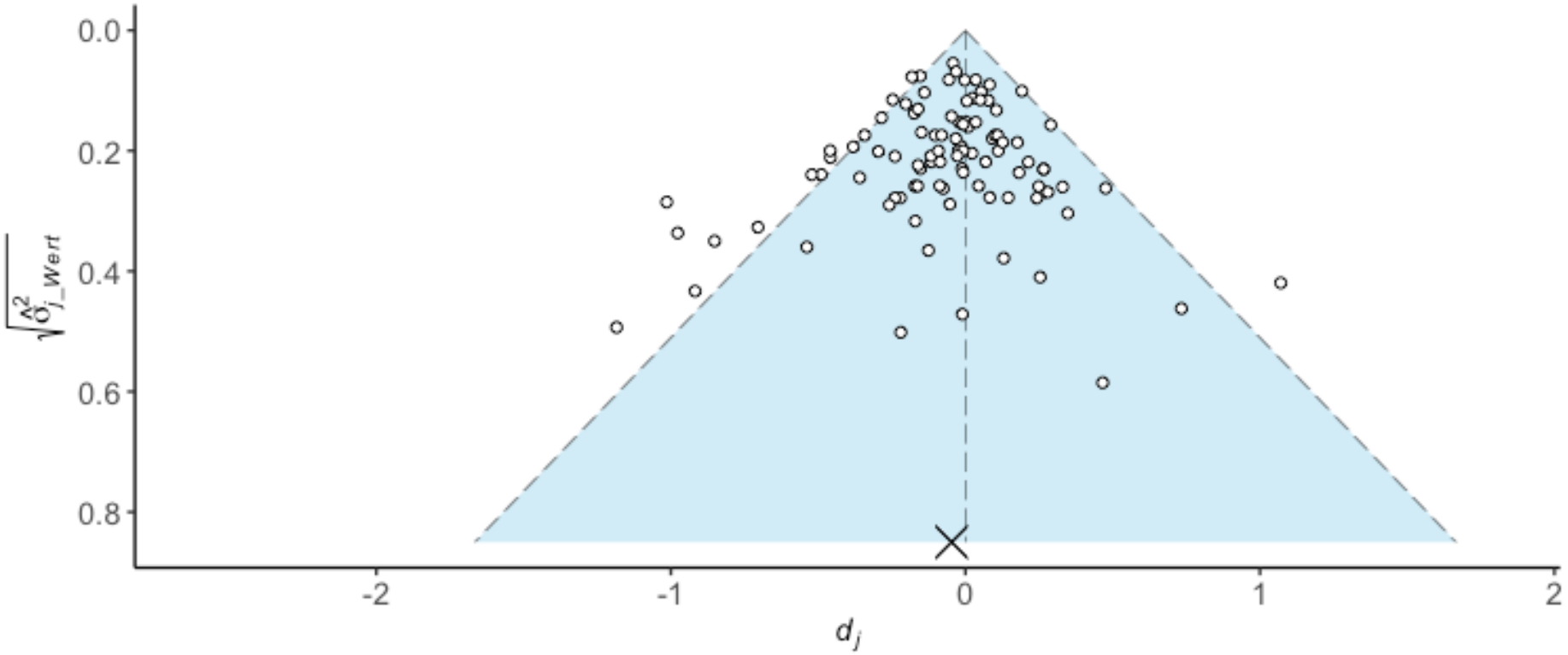
- Maximale Stichprobengröße pro Gruppe: $n_1 = n_2 = 400$
- Signifikanzniveau: $\alpha^* = 0.005$
- Maximal 2 geplante Tests: Test nach $n_1 = n_2 = 200$ sowie nach $n_1 = n_2 = 400$
- Durchführung: Es werden 200 Probandinnen pro Gruppe erhoben und der t-Test mit $\alpha = \frac{0.005}{2} = 0.0025$ durchgeführt. Entscheidet sich der Test für die Alternativhypothese, wird die Datenerhebung vorzeitig gestoppt.
- Ist der Test nicht signifikant erhebt man weiter bis $n_1 = n_2 = 400$ und testet erneut mit $\alpha = 0.0025$. Unabhängig vom Testergebnis ist die Datenerhebung beendet.

Exkurs: Simulation von P-Hacking

Annahmen in Carter, Schönbrodt, Gervais & Hilgard, 2019:

- Kein P-Hacking (in 30% der durchgeführten Studien) -> WK Fehler 1. Art = 5%
- Mittelschweres P-Hacking (in 50% der durchgeführten Studien) -> WK Fehler 1. Art ≈ 9%
  1. Analyse einer 2. abhängigen Variable (AV)
  2. Optional Stopping (bis zu 3 mal 3 weitere Probandinnen pro Gruppe)
     (+ Schritt 1 nach jeder neuen Datenerhebung)
- Schweres P-Hacking (in 20% der durchgeführten Studien) -> WK Fehler 1. Art ≈ 27%
  1. Entfernen von Ausreißern (|z-standardisierte AV| > 2)
  2. Analyse einer 2. AV (+ Ausreißer)
  3. Getrennte Analyse von Männern und Frauen in der 1. AV (+ Ausreißer)
  4. Getrennte Analyse von Männern und Frauen in der 2. AV (+ Ausreißer)
  5. Optional Stopping (bis zu 5 mal 3 weitere Probandinnen pro Gruppe)
     (+ Schritte 1 - 4 nach jeder neuen Datenerhebung)
- Zusatzannahme: Führt das P-Hacking nicht zum Erfolg, wird die ursprüngliche „saubere“ Analyse verwendet.

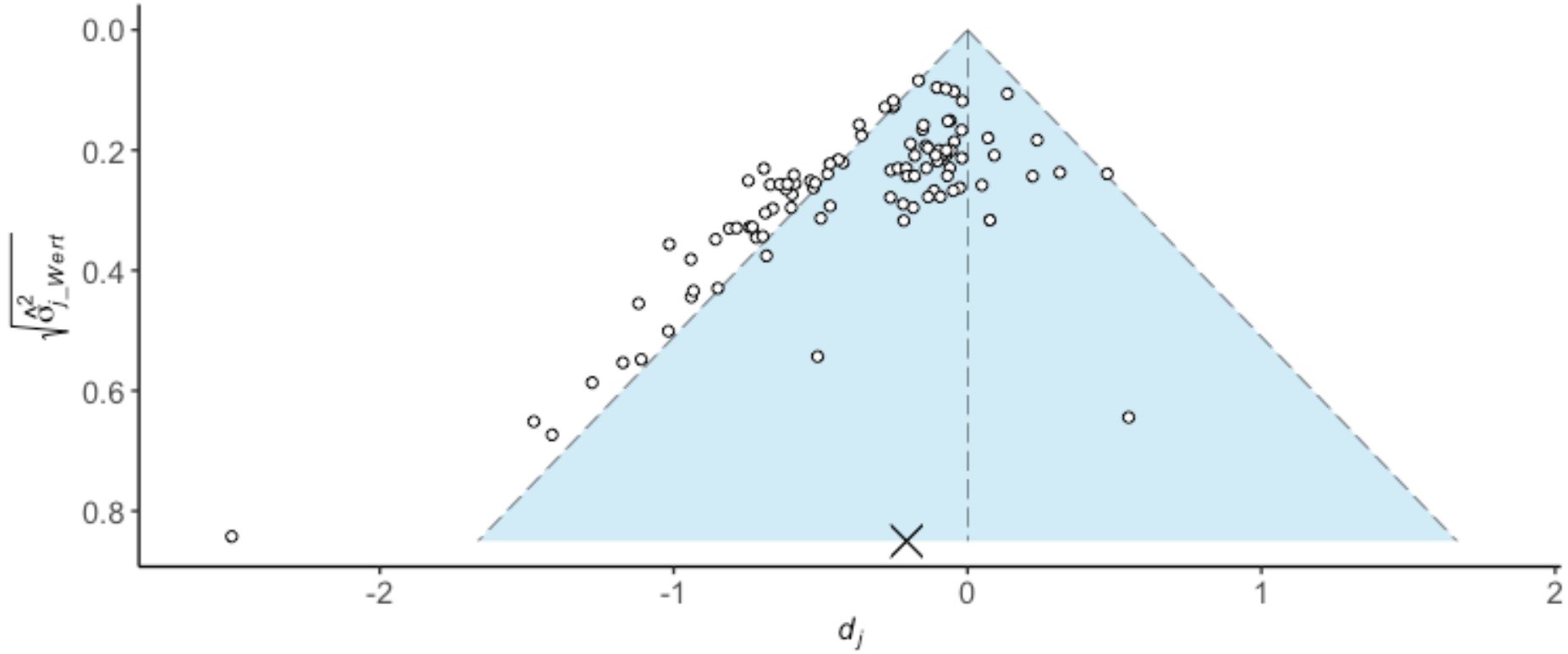
Funnelplot mit P-Hacking



Simulierte Metaanalyse mit $N = 100$ Studien, $\delta = 0$ und P-Hacking:

- Mehr Studien mit knapp signifikantem Ergebnis (falls $d_j < 0$)
- Unterschätzung von δ durch die Metaanalyse:
  Verzerrtes KI und erhöhte Wahrscheinlichkeit für Fehler 1. Art in der Metaanalyse

Funnelplot mit Publikationsbias und P-Hacking



Simulierte Metaanalyse mit $N = 100$ Studien, $\delta = 0$, **Publikationsbias** und **P-Hacking**:

- Mehr publizierte Studien mit $d_j < 0$
- Mehr Studien mit knapp signifikantem Ergebnis
- Deutliche Unterschätzung von δ durch die Metaanalyse:
  Verzerrtes KI und erhöhte Wahrscheinlichkeit für Fehler 1. Art in der Metaanalyse

Plan heute

- Thema heute: Probleme, die das Vertrauen in metaanalytische Befunde in der Praxis stark einschränken können, sowie mögliche Lösungsansätze.

✓ Probleme von Metaanalysen:

- ✓ (Heterogenität der Effekte)
- ✓ Publikationsbias
- ✓ Researcher Degrees of Freedom und P-Hacking

- **Folgen: Replikationskrise in der Psychologie**

- Lösungsansatz: Open Science
  - Präregistrierung
  - Veröffentlichung der Rohdaten, Versuchsmaterialien und Analyseskripten

Replikationskrise in der Psychologie I

- Ausgelöst durch empirische (Bem, 2011) sowie methodische Studien (Simmons et al., 2011), die auf Probleme bei der Verlässlichkeit publizierter psychologischer Effekte hindeuteten, gibt es seit einigen Jahren vermehrte Bestrebungen, in der Literatur gut etablierte und hochrangig publizierte Effekte zu **replizieren**.

- Die Replikation einer Studie ist dabei ein sehr weit gefasster Begriff und bedeutet erst einmal nur, dass ein Effekt an einer unabhängigen Stichprobe erneut untersucht wird.

    - Manche Replikationsstudien versuchen dabei, die Originalstudie so genau wie möglich nachzustellen (gleiche Materialien, gleiche Variablen, gleiche Messinstrumente, gleiche Instruktionen, gleiche Analysen, etc.).

    - Andere Replikationsstudien untersuchen den gleichen Effekt, aber unter etwas anderen Bedingungen (z.B. unterschiedliches Messinstrument, Population, Stimuli).

Replikationskrise in der Psychologie II

- Es wird viel diskutiert, wann eine Replikation als „erfolgreich“ eingestuft werden soll. Aktuelle Forschung beschäftigt sich mit verschiedenen Ansätzen (Held, 2020).

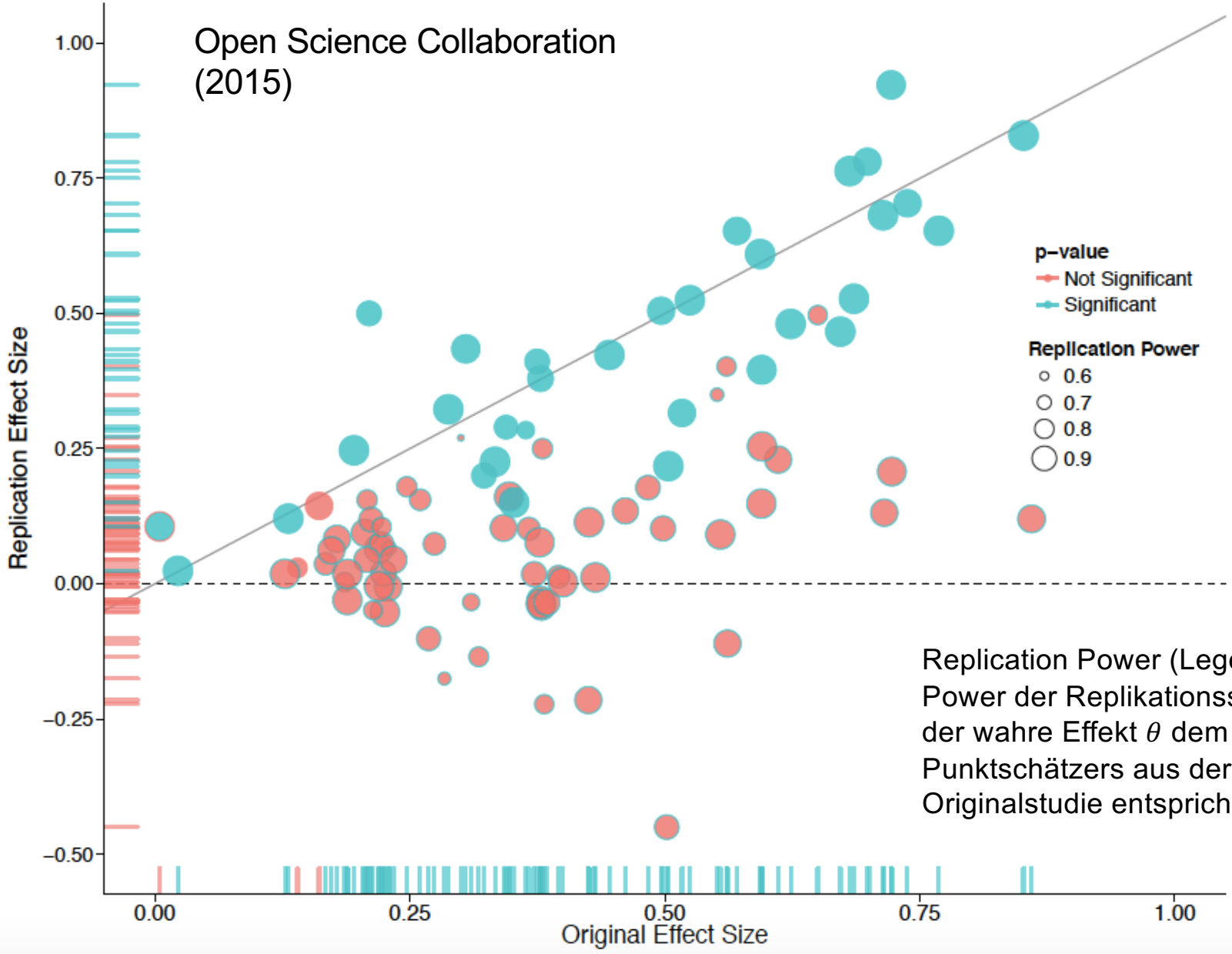
- Der einfachste Ansatz ist, dass eine Replikationsstudie als „erfolgreich“ interpretiert wird, wenn der Hypothesentest in der Replikationsstudie signifikant wird und gleichzeitig der geschätzte Effekt in die gleiche Richtung geht wie in der Originalstudie (z.B. $d_{Originalstudie} > 0$ und $d_{Replikationsstudie} > 0$).

- Der Anteil der Replikationsstudien, in denen das einfache Erfolgskriterium erfüllt war, betrug je nach Untersuchung 36% (Open Science Collaboration, 2015), 54% (Klein et al., 2018) oder 62% (Camerer, 2018).

- Der in den Replikationsstudien geschätzte Effekt im Betrag betrug je nach Untersuchung 49% (Open Science Collaboration, 2015), 25% (Klein et al., 2018) oder 50% (Camerer, 2018) der Größe des Effekts in den Originalstudien.

Exkurs: Reproducibility Project in Psychology

Open Science Collaboration (2015)



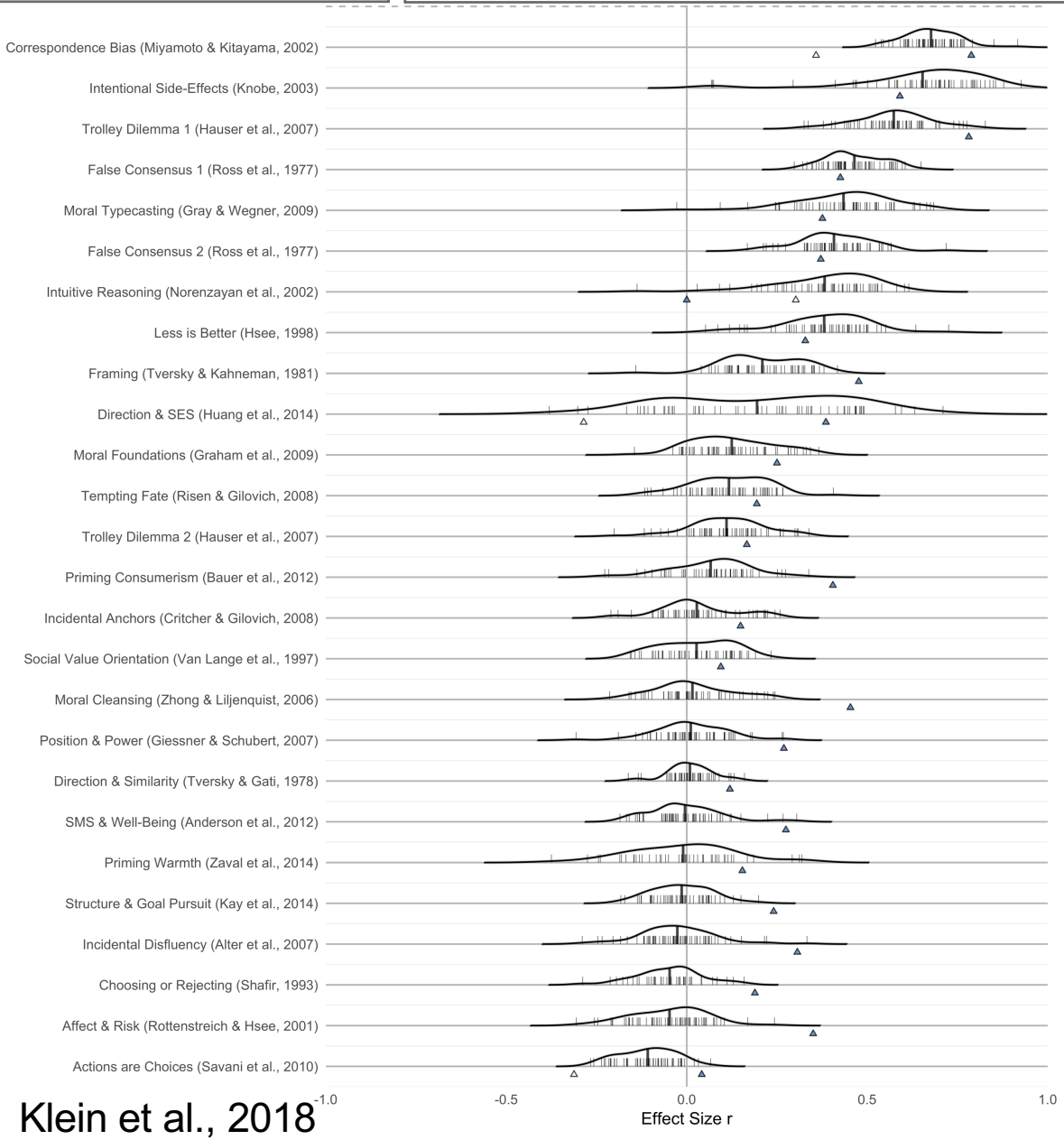
Replication Power (Legende): Power der Replikationsstudie, falls der wahre Effekt θ dem Wert des Punktschätzers aus der Originalstudie entspricht.

Stichprobenplanung bei Replikationsstudien

- Liegt in der Literatur Publikationsbias und/oder P-Hacking vor, ist die FDR hoch und ein großer Anteil der publizierten Studien überschätzt den wahren Effekt θ.
- Verwendet man für die Stichprobenplanung einer Replikationsstudie den zu hohen Wert des Punktschätzers aus der Originalstudie, ist die tatsächliche Power der Replikationsstudie eventuell deutlich niedriger als geplant.
- Falls die Alternativhypothese gilt und in der Population ein Effekt $\theta \neq 0$ vorliegt, entspricht die Wahrscheinlichkeit, dass die Replikationsstudie signifikant wird, der Power. Ist die Power sehr niedrig, ist die Wahrscheinlichkeit, für eine erfolgreiche Replikation sehr niedrig, obwohl tatsächlich ein Effekt vorliegt!

- Fazit: Bei der Stichprobenplanung für Replikationsstudien sollte man nicht nur den Punktschätzer der Originalstudie berücksichtigen. Um eine hohe Aussagekraft sicherzustellen, sollte man für Replikationen immer konservativ von einem kleinen wahren Effekt ausgehen, d.h. große Stichproben erheben.

Exkurs: Many Labs 2



Klein et al., 2018

- Jeder Strich ist eine Replikation.
- Dreieck: Punktschätzung der Originalstudie (zwei Dreiecke, falls kulturelle Unterschiede in der Originalstudie berichtet)
- Für manche untersuchten Effekte ist der Schätzwert aus der Originalstudie größer als (fast) alle Replikationsstudien.
- Auch für untersuchte Effekte, für die die Replikationsstudien einen Effekt $\theta \neq 0$ nahelegen, ist der Schätzwert aus der Originalstudie manchmal deutlich höher als der Median (jeweils der dicke Strich) der Replikationsstudien.

Plan heute

Thema heute: Probleme, die das Vertrauen in metaanalytische Befunde in der Praxis stark einschränken können, sowie mögliche Lösungsansätze.

- ✓ Probleme von Metaanalysen:
  - ✓ (Heterogenität der Effekte)
  - ✓ Publikationsbias
  - ✓ Researcher Degrees of Freedom und P-Hacking
- ✓ Folgen: Replikationskrise in der Psychologie
- **Lösungsansatz: Open Science**
  - **Präregistrierung**
  - **Veröffentlichung der Rohdaten, Versuchsmaterialien und Analyseskripten**

Präregistrierung I

- Eine Möglichkeit, um zu verhindern, dass RDoF die Wahrscheinlichkeit für einen Fehler 1. Art und damit die FDR in der psychologischen Literatur erhöhen, ist es, den geplanten Ablauf einer Studie vor deren Durchführung auf einem öffentlichen Service (z.B. osf.io) mit Zeitstempel zu **präregistrieren** (Nosek, et al., 2018).
- Im Optimalfall sollten in einer Präregistrierung alle Punkte, bei denen RDoF auftreten können, genau festgelegt werden. Dies beinhaltet vor allem die Beschreibung …
  - der genauen Hypothesen und Forschungsfragen
  - des Materials, aller Versuchsbedingungen und aller erhobenen Variablen
  - der Datenerhebung und Zielpopulation mit Stichprobenplanung und Stoppregel
  - des genauen Ablaufs der Datenanalyse inklusive unvorhergesehener Probleme wie Ausreißer, fehlende Werte, Verletzung statistischer Annahmen, …
  - die Interpretation abhängig von den Ergebnissen der Analysen (Erfolgskriterien)
- Für maximale Transparenz: R-Code für die Analyse vor der Datenerhebung anhand simulierter Daten entwickeln und mit präregistrieren.

Präregistrierung II

Hauptziel von Präregistrierung ist die Steigerung von Transparenz in der Wissenschaft:

"The first principle is that you must not fool yourself, and you are the easiest person to fool." — Richard P. Feynman

- Präregistrierung soll nicht verhindern, vom ursprünglichen Studienplan abzuweichen, falls es dafür eine gute Begründung gibt (neue Erkenntnisse, Analysemethoden, etc.). Abweichungen von der Präregistrierung sollten aber transparent gemacht werden.

- Präregistrierung soll nicht verhindern, bereits verfügbare Daten neu zu analysieren. Auch hier kann eine Präregistrierung sehr sinnvoll sein und sollte zusätzlich enthalten, was bereits über die Daten bekannt ist und welchen Einfluss dieses Wissen z.B. auf die Auswahl der geplanten Analysen hatte.

- Präregistrierung soll keine explorative Forschung verhindern. Auch hier kann eine Präregistrierung sinnvoll sein, um die Qualität der geplanten Studie zu erhöhen.

Präregistrierung III

Präregistrierung ist kein Allheilmittel, sondern nur ein Baustein für eine niedrigere FDR:

- Präregistrierung soll rechtschaffene Wissenschaftlerinnen unterstützen, möglichst verlässliche Erkenntnisse zu produzieren. Betrügerinnen, die bereit sind, RoDF für absichtliches P-Hacking oder sogar schlimmere Vergehen wie Datenfälschung zu nutzen, werden immer noch Wege finden, zu betrügen („Preregistration Hacking“).
- Präregistrierung alleine hilft nicht gegen Publikationsbias:
  Angenommen, eine große Studie wird geplant, in der sehr viele Hypothesen präregistriert werden. Die Ergebnisse können immer noch einen Einfluss auf den Veröffentlichungsprozess haben. Nicht signifikante oder weniger überraschende Ergebnisse werden eventuell deutlich später oder gar nicht publiziert.

- Es ist umstritten, ob Präregistrierung zur Verbesserung psychologischer Theorien beitragen kann (Szollosi et al., 2019). Ein starker Fokus darauf, Präregistrierungen einzufordern, könnte auch von dem noch gravierenderen Problem der schlechten Qualität vieler psychologischer Theorien ablenken.

Registered Reports

Königsweg wissenschaftlicher Artikel (Chambers, 2019): **Registered Reports**

- Präregistrierung wird vor der Datenerhebung von externen Gutachterinnen bewertet. Es können Verbesserungen des Studienplans erarbeitet werden.
- Bei erfolgreicher Begutachtung verpflichtet sich die Zeitschrift, das Manuskript wie geplant zu veröffentlichen, unabhängig von den resultierenden Ergebnissen.
- Registered Reports schaffen deutlich weniger Anreize, signifikante Ergebnisse zu erzeugen. Gleichzeitig werden Anreize geschaffen, Studien mit hohem erwarteten Erkenntnisgewinn durchzuführen.

Open Science

- Neben der Präregistrierung von Studien gibt es weitere wichtige Initiativen zur Steigerung der Transparenz von (psychologischer) Forschung. Diese Initiativen werden häufig unter dem Begriff **Open Science** zusammen gefasst:

- **Open Data, Materials & Code:** Veröffentlichung aller Forschungsdaten, Versuchsmaterialien, Messinstrumente und Analyseskripten:
  - Forschungsdaten können von allen Wissenschaftlerinnen optimal genutzt werden.
  - Publizierte Analysen können von anderen Wissenschaftlerinnen überprüft werden.
  - Publizierte Studien können von anderen Wissenschaftlerinnen repliziert werden.
  - Standardisierung von Versuchsbedingungen und Messinstrumenten tragen zum Erkenntnisgewinn bei, da Forschungsergebnisse besser vergleichbar sind.
- **Open Access:** Freier Zugang zu wissenschaftlichen Artikeln
  - Öffentlich finanzierte Forschung sollte allen Menschen zu Gute kommen.
  - Veröffentlichung von Manuskripten auf Preprintservern (z.B. psyarxiv.com).

Fazit: Anwendung statistischer Methoden in der Psychologie

- Ein Großteil der Studien in der psychologischen Literatur verwendet statistische Hypothesentests mit viel zu geringer Stichprobengröße (damit niedrige Power).
- Zusätzlich scheinen Publikationsbias und fragwürdige Forschungspraktiken an der Tagesordnung zu sein.
- Alles spricht dafür, dass die FDR bei psychologischen Studien extrem hoch ist (eventuell sogar deutlich höher als 50%).

- Wir können den Ergebnissen einzelner psychologischer Studien also nicht vertrauen, ohne die Studien genauer angeschaut zu haben.
- Aufgrund der dargestellten Problemen können wir auch Metaanalysen mehrerer einzelner Studien nicht ohne Weiteres vertrauen, selbst wenn die eingeschlossenen Studien unauffällig erscheinen.

Konsequenzen für das eigene wissenschaftliche Arbeiten

- Überlegen Sie sich genau, welche Aussagen Sie überprüfen wollen und übersetzen Sie diese in spezifische Aussagen bzgl. der Parameter eines für Ihre Fragestellung geeigneten statistischen Modells.
  -> Thema aller kommenden Statistik Vorlesungen!

- Wählen Sie entsprechende inferenzstatistische Verfahren für genau diese Aussagen.
- Führen Sie eine Stichprobenplanung für diese Verfahren mit $\alpha = 0.005$, $1 - \beta = 0.8$ und einem kleinen Effekt durch. Falls Sie die benötigte Stichprobengröße nicht erheben können oder Sie keine konkreten inhaltlichen Hypothesen untersuchen:
  Keine Hypothesentests, sondern KIs (diese nicht als Hypothesentests interpretieren)
- Präregistrieren Sie geplante Analysen, am besten mit Veröffentlichung des R-Codes.
- Erheben Sie dann Ihre Daten und führen Sie die geplanten Analysen durch. Wenn Sie von der Präregistrierung abweichen, machen Sie dies transparent (kein Problem!).
- Veröffentlichen Sie Ihre Daten (soweit möglich), Materialien und Ihren R-Code.

Konsequenzen für das Beurteilen von anderen wissenschaftlichen Arbeiten

- Auch wenn die FDR für die Gesamtheit der psychologischen Studien sehr hoch ist, wird sie für einen nach sinnvollen Kriterien ausgewählten Teil der Studien deutlich niedriger sein.

- Obwohl dies eigentlich die Aufgabe der Fachzeitschriften wäre, müssen Sie diese Auswahl selbst treffen.

- (Sehr allgemeine) Daumenregeln dafür, dass Sie den Ergebnissen einer Studie/Metaanalyse (hoffentlich) vertrauen können:
  - Präregistrierung (bei Metaanalysen sowohl bezogen auf die einzelnen Studien als auch auf die Metaanalyse selbst) **und**
  - Große Stichprobe + hohe Power (bei Metaanalysen bezogen auf die Gesamtstichprobe) **und**
  - P-Werte der signifikanten Hypothesentests kleiner als 0.005 .

Exkurs: Quellenangaben

- Fanelli, D. (2011). Negative results are disappearing from most disciplines and countries. Scientometrics, 90, 891–904.
- Bakker, M., van Dijk, A., & Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science*, *7*(6), 543-554.
- Carter, E. C., Schönbrodt, F. D., Gervais, W. M., & Hilgard, J. (2019). Correcting for Bias in Psychology: A Comparison of Meta-Analytic Methods. Advances in Methods and Practices in Psychological Science, 2(2), 115–144.
- Wicherts, J. M., Veldkamp, C. L., Augusteijn, H. E., Bakker, M., Van Aert, R., & Van Assen, M. A. (2016). Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid p-hacking. *Frontiers in psychology*, *7*, 1832.
- John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological science*, *23*(5), 524-532.
- Wang, M. Q., Yan, A. F., & Katz, R. V. (2018). Researcher requests for inappropriate analysis and reporting: A US Survey of consulting biostatisticians. *Annals of internal medicine*, *169*(8), 554-558.
- Gelman, A., & Loken, E. (2013). The garden of forking paths: Why multiple comparisons can be a problem, even when there is no “fishing expedition” or “p-hacking” and the research hypothesis was posited ahead of time. *Department of Statistics, Columbia University*.
- Lakens, D. (2014). Performing high-powered studies efficiently with sequential analyses. *European Journal of Social Psychology*, *44*(7), 701-710.
- Bem, D. J. (2011). Feeling the future: experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of personality and social psychology*, 100(3), 407.
- Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. Psychological Science, 22, 1359–1366.
- Held, L. (2020), A new standard for the analysis and design of replication studies. J. R. Stat. Soc. A, 183: 431-448.
- Open Science Collaboration. (2015). Estimating the Reproducibility of Psychological Science. *Science*, 349, aac4716.
- Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams, R. B., Jr., Alper, S., . . . Nosek, B. A. (2018). Many Labs 2: Investigating variation in replicability across samples and settings. Advances in Methods and Practices in Psychological Science, 1, 443–490.
- Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T. H., Huber, J., Johannesson, M., ... & Altmejd, A. (2018). Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour*, *2*(9), 637-644.
- Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences*, *115*(11), 2600-2606.
- Szollosi, A., Kellen, D., Navarro, D., Shiffrin, R., van Rooij, I., Van Zandt, T., & Donkin, C. (2019). Is preregistration worthwhile?. *Trends in Cognitive Sciences*, 24(2), 94-95.
- Chambers, C. (2019). What’s next for registered reports?. *Nature*, 573, 187-189.