# Das Problem dysfunktionaler Reputationssysteme am Beispiel von Fehlverhalten und Diskriminierung in der Wissenschaft

Inaugural-Dissertation

zur Erlangung des Doktorgrades der Sozialwissenschaftlichen Fakultät

der Ludwig-Maximilians-Universität München

vorgelegt von

Andreas Schneck

2019

Erstgutachterin: Prof. Dr. Katrin Auspurg (Ludwig-Maximilians-Universität München)

Zweitgutachter: Prof. Dr. Thomas Hinz (Universität Konstanz)

Tag der mündlichen Prüfung: 10. Mai 2019

## Danksagung

Die vorliegende Dissertationsschrift wäre ohne die Inspiration und Unterstützung vieler Personen sicherlich nicht entstanden. Im Folgenden sei daher allen unmittelbar und mittelbar Beteiligten ganz herzlich gedankt.

Mein besonders herzlicher Dank gilt meiner Erstbetreuerin Prof. Dr. Katrin Auspurg. Sie gab mir angefangen vom Studium über die Promotionsphase stets wertvolle Rückmeldung sowie inspirierende Hinweise und hatte immer ein offenes Ohr für meine Nachfragen. Darüber hinaus ermöglichte sie auch die Vorstellung der Ergebnisse auf zahlreichen Konferenzen und förderte die eigenständige Weiterentwicklung der Arbeit, im vorliegenden Fall insbesondere im Bereich Data Science und Computational Social Science. Über die Dissertation hinaus hat mich aber auch die Mitarbeit in diversen Projekten von Prof. Dr. Katrin Auspurg an der Universität Konstanz sowie am Lehrstuhl an den Universitäten Frankfurt und München sehr bereichert und mir einen breiten Einblick in die Forschungslandschaft ermöglicht. Mein besonderer Dank gilt ebenso meinem Zweitbetreuer Prof. Dr. Thomas Hinz. Die gemeinsamen Artikel mit Prof. Dr. Katrin Auspurg und Prof. Dr. Thomas Hinz, von denen zwei in die vorliegende Dissertationsschrift Eingang gefunden haben, halfen ungemein bei der Einführung in das Schreiben und Konzipieren wissenschaftlicher Publikationen. Prof. Dr. Thomas Augustin sei an dieser Stelle ebenso gedankt für die Übernahme der Rolle des Drittprüfers meiner Disputation.

Mein Dank gebührt weiterhin den zahlreichen KollegInnen an den Universitäten Konstanz, Frankfurt und München, die gute Gespräche zum Gelingen der Arbeit beigesteuert sowie die Vorstellung von Teilprojekten der Arbeit in Kolloquien ermöglicht haben. Besonders hervorgehoben seien auch die TeilnehmerInnen des Seminars Analytische-Soziologie an der Venice International University sowie die TeilnehmerInnen der Konferenzen des MAER Networks, welche durch ihre Hinweise und Anmerkungen ganz erheblich zur inhaltlichen Schärfung beigetragen haben.

Nicht zuletzt danke ich allen meinen FreundInnen sowie meiner Familie für ihre Unterstützung.

München, Juni 2019 Andreas Schneck

Und er kommt zu dem Ergebnis:
Nur ein Traum war das Erlebnis.
Weil, so schließt er messerscharf,
nicht sein *kann*, was nicht sein *darf*.

Aus „Die unmögliche Tatsache“ von Christian Morgenstern

# Inhaltsverzeichnis

# Abbildungsverzeichnis

Das Abbildungsverzeichnis ist nach Kapiteln nummeriert, d.h. die erste Zahl bezeichnet das Kapitel, die zweite die fortlaufende Nummer innerhalb des Kapitels. Bei englischsprachigen Artikeln ist der englische Begriff Figure anstelle von Abbildung verwendet.

# Tabellenverzeichnis

Es gelten dieselben Bemerkungen zur Nummerierung und den Begriffen wie im Abbildungsverzeichnis.

# 1. Rahmenkapitel

## 1.1. Einleitung

Das erklärte Ziel von Wissenschaft, sei es in den Natur-, Geistes-, oder Sozialwissenschaften, ist es, Erkenntnis zu gewinnen. Dieses Streben nach Erkenntnisgewinn kann als Suche nach Wahrheit verstanden werden (Albert 1978; Descartes 2006). Jedem Suchprozess inhärent sind dabei (vermeintliche) Fehlschläge, welche eben nicht erwartungskonforme Ergebnisse erzielen. Die Wissenschaft selbst steht dabei immanent unter gesellschaftlichem Rechtfertigungsdruck, da sie sich nicht allein ökonomischen Prinzipien einer Verwertung der Ergebnisse auf dem Markt unterordnen muss, sondern vielmehr auch dem gesellschaftlichen Ziel der Erkenntnisgewinnung verpflichtet ist. Ihre Ergebnisse können demnach als Allgemeingut aufgefasst werden, welches grundsätzlich keinen partikularen Einzelinteressen nutzt, auch wenn das wirtschaftliche Wachstum und damit das Gemeinwohl mit dem Fortschritt in der Wissenschaft verbunden ist (Stephan 2010: 210). Diese mittelbare Loslösung vom Verwendungsdruck des Marktes führt zu Legitimationsproblemen, welche immer wieder in Kritik an der Wissenschaft und deren tatsächlichen Beitrag zum Erkenntnisfortschritt münden.

Dass diese Form der Kritik an und damit die (vermeintliche) Krise der Wissenschaft keineswegs ein neues Phänomen ist, beweist Merton (1973: 267, Erstveröffentlichung als Artikel 1942), der dies gar als Herausforderung ansah. Auch in neuerer Zeit werden verschiedene Bedrohungen der Legitimität von Wissenschaft diskutiert. Besonders hervorgehoben sei an dieser Stelle die Debatte zur Replikationskrise, die durch spektakuläre Fälle wissenschaftlicher Fälschungen (Levelt Commitee et al. 2012; Reich 2009) ausgelöst wurde. In einem groß angelegten Replikationsversuch der Open Science Collaboration konnten nur 39% der untersuchten Studien erfolgreich repliziert werden, d. h. den ursprünglichen Effekt der Vorgängerstudie erzielen (Open Science Collaboration 2015; für einen ähnlichen Befund in der Ökonomie s. Chang & Li 2015). Aus diesem Ergebnis resultiert, dass an der Erkenntnisleistung der Wissenschaft erhebliche Zweifel angebracht sind. Der Verwurf einer „corrupt research“ (Hubbard 2015), welche nur mehr Artefakte, denn faktuale Ergebnisse produziere, erscheint einerseits normativ gesehen ein zu vermeidendes allzu dystopisches Narrativ, andererseits muss die Erosion der Glaubwürdigkeit der Wissenschaft essentiell von der Wissenschaft selbst adressiert werden, um das eigene Fortbestehen zu sichern. Es stellt sich daher die dringende Frage, ob, in welchem Umfang und unter welchen Risikobedingungen bestimmte Praktiken in der Wissenschaft den Erkenntnisfortschritt behindern. Die vorliegende Dissertation untersucht anhand des Problems dysfunktionaler Reputationssysteme die Ineffizienz des wissenschaftlichen Erkenntnisfortschritts exemplarisch anhand zweier Beispiele: der Auswertungspraxen von Daten sowie der Auswahl von BewerberInnen auf besonders prestigeträchtige Positionen im Wissenschaftssystem (der Professur).

Die Untersuchung der Auswertungspraxen von Daten fokussiert dabei insbesondere das Phänomen des Publication Bias. Publication Bias liegt vor, wenn nur die Ergebnisse, die der Erwartung bzw. Hypothese der Forschenden entsprechen, veröffentlicht werden (Dickersin & Min 1993; für eine Übersicht über

weitere Publication Bias Definitionen s. Bassler et al. 2016). Ohne die Veröffentlichung bleiben Studienergebnisse, welche eben nicht den erwarteten Effekt zeigen, unsichtbar in der Schublade der Forschenden liegen (vgl. file drawer Rosenthal 1979) und können so keinen Beitrag zum Erkenntnisgewinn leisten. Petticrew (1998) beschreibt die Auswirkungen des Publication Bias eindrücklich anhand des griechischen Dichters und Denkers *Diagoras von Melos*, welcher die in einem Tempel dargestellten und die Macht des Meeresgottes Neptun verherrlichenden Bilder der aus dem Meer Geretteten als Gottesbeweis ablehnte und nur trocken auf die fehlenden Bilder der Ertrunkenen verwies, welche eben gerade nicht gezeigt würden. Nach dem gleichen Grundprinzip und besonders eindrücklich fallen die Folgen des Publication Bias in der Medizin aus. Liegt Publication Bias vor, so werden Patienten wirkungslosen, jedoch in Studien immer wieder als wirkungsvoll angepriesenen Medikamenten ausgesetzt (Godlee 2012), da die Ergebnisse, die auf die Unwirksamkeit der Medikamente verweisen, eben nicht sichtbar und zugänglich sind. Publikation Bias stellt dabei eine statistisch nicht begründbare Ungleichbehandlung von Forschungsergebnissen dar. Auch in qualitativen Studien, welche keine statistischen Methoden verwenden, lassen sich ähnliche Tendenzen zur Publikation erwartungskonformer Ergebnisse finden (Petticrew et al. 2008).

Im Gegensatz zu den Ungleichbehandlungen von (statistischen) Ergebnissen im Hinblick auf deren Publikation, zählt eine selektive Auswahl der Forschenden selbst – so sie meritokratischen Prinzipien folgend eine leistungsgerechte Bestenauswahl darstellt – zu den Pfeilern guter wissenschaftlicher Praxis. Trotz vermehrter Bemühungen zur Gleichstellung von Frauen und Männern in den letzten Jahren, finden sich immer noch Hinweise auf nicht leistungsgerechte Verbleibschancen im Wissenschaftssystem nach dem Geschlecht (Weisshaar 2017) und damit Anzeichen für Diskriminierung. Gerade Wissenschaftlerinnen werden trotz identischer Leistung teils als weniger kompetent und auch kollaborationswürdig beschrieben (Knobloch-Westerwick et al. 2013).

Die in der vorliegenden Dissertationsschrift vertretene These ist es, dass durch die Ungleichheit in der Behandlung von Forschungsergebnissen sowie auch in den nach Geschlecht unterschiedlichen Verbleibschancen von Forschenden in der Wissenschaft Ineffizienzen des Wissenschaftssystems offengelegt werden, welche den Erkenntnisfortschritt als Ziel aller Wissenschaft unterlaufen. Konkret hat dies zur Folge, dass wissenschaftliche Ergebnisse zu bloßen statistischen Artefakten absinken sowie die Forschungsleistung bei einer nicht-meritokratisch gerechtfertigten Ungleichbehandlung von Forschenden (z. B. aufgrund ihres Geschlechts) nicht den maximal erreichbaren Erkenntnisfortschritt gewährleistet. Diese Abweichung von der maximalen Erkenntnisleistung – entweder durch nichtpublizierte Ergebnisse oder dem Ausscheiden von kompetenten BewerberInnen – bei gleichem Mitteleinsatz – z. B. die Anzahl durchgeführter statistischer Tests bzw. die Anzahl an Professuren in der Wissenschaft – kann daher als ineffizient definiert werden.

Der Fokus der vorliegenden Arbeit liegt zum einen auf der innovativen Nutzung neuer Methoden der Datengenerierung, durch welche diese oft schwer zu beobachtenden Selektionsprozesse erstmals umfassend untersucht werden können. Zudem wird ein besonderes Augenmerk auf die Evaluation und Entwicklung von Testverfahren und Methoden gelegt, um den Effekt so genau als möglich zu erkennen als auch in dessen Ursachen zu erklären. Die vorliegende Dissertationsschrift positioniert sich daher zwischen den zwei Polen erklärender, die Ursachen identifizierender, sowie deskriptiver, um möglichst genaue Diagnose und Identifikation bemühter, Forschung (Leek & Peng 2015). Besonders hervorzuheben ist, dass eine valide Beschreibung, im vorliegenden Fall der verschiedenen Symptome von Ineffizienz im Wissenschaftssystem, einer Erklärung der Ursachen zwingend als Grundlagenforschung vorausgehen muss.

Das vorliegende Rahmenkapitel legt zunächst die handlungstheoretischen Grundlagen, welche zu einem ineffizienten Wissenschaftssystem führen können, dar. So werden zum einen die Normen des Makrosystems Wissenschaft, als auch die Anreize und Handlungsmöglichkeiten der Forschenden auf der Mikroebene erläutert. Darauf aufbauend werden die vier kumulativen Einzelbeiträge zusammenfassend in die Gesamtkonzeption der Dissertation eingeordnet, bevor final mögliche Implikationen zur Reduzierung bzw. Vermeidung möglicher Ineffizienzen, insbesondere im Hinblick auf die individuellen Anreizstrukturen, diskutiert werden.

## 1.2. Handlungstheoretische Einordnung

Ziel der vorliegenden theoretischen Einordnung ist es, im Sinne einer mechanistischen Erklärung eines empirisch beobachteten Zustands die Gründe für dessen Auftreten zu erklären (Elster 1989: 5). Im vorliegenden Fall ist die Beobachtung des Zustandes einer ineffizienten Wissenschaft auf der Makroebene anzusiedeln (M2 in Abbildung 1-1), sei es bezüglich der berichteten Forschungsergebnisse oder bei der Auswahl des Personals.

**Abbildung 1-1 Mikro-Makro-Erklärungsmodell**

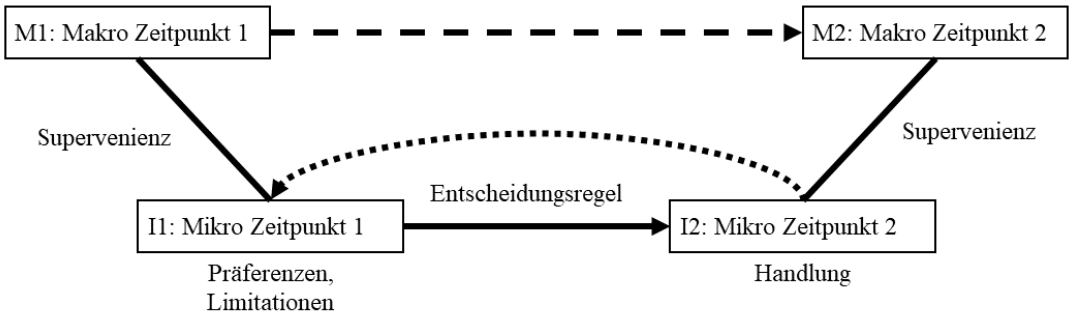



Abbildung angelehnt an Coleman (1990: 8) und Hedström & Bearman (2009: 10).

Die Erklärung dieses ineffizienten Zustandes kann jedoch, um ökologische Fehlschlüsse zu vermeiden, nicht auf der zeitlich vorausgehenden Makroebene (M1) angesiedelt werden, sondern konstituiert sich aus den Handlungen von einzelnen Akteuren[1] auf der Individual- oder Mikroebene (I).[2] Es wird daher im Folgenden ein im methodischen Individualismus (Coleman 1990 für einen in der Soziologie prominenten Vorreiter) bzw. strukturellen Individualismus (Hedström & Bearman 2009: 8) verankertes Mechanismus-Verständnis vertreten. Dies bedeutet jedoch nicht, dass gesellschaftliche Kontextbedingungen nicht auch für individuelles Handeln von Belang sein können. Im Sinne des strukturellen Individualismus kann hier sogar noch weiter gegangen werden, da auch Wechselwirkungen zwischen den Akteuren selbst theoretisch modelliert werden können (I2-I1, in Abbildung 1-1 gepunktete Linie). Beide Makroeigenschaften sind hierbei die Aggregationsform der jeweils zugehörigen Mikroeigenschaft und aus dieser ableitbar. So ist nicht nur I2 als Handlung aus dessen Aggregatform (M2) ableitbar, auch bezeichnet als Aggregationsregel, sondern ebenfalls die den Handlungsvoraussetzungen (I1) zugrunde liegenden Makroeigenschaften (M1), auch bezeichnet als Brückenhypothese. Hedström und Bearman (2009: 10) benennen dieses Phänomen als Supervenienz. Im Folgenden gilt es nun ausgehend von dem vorangehend erläuterten theoretischen Paradigma zuerst die im Wissenschaftssystem immanenten Normen, welche auf der Makroebene verortet sind, zu erläutern. Sodann werden die daraus sich ergebenden, im Individuum zu verortenden Handlungsvoraussetzungen sowie die individuelle Handlung und deren Makroergebnis diskutiert.

### *1.2.1. Makroebene*

Auf der Makroebene sollen als Kontextfaktoren insbesondere die normativen Rahmenbedingungen (M1) von (guter) Wissenschaft diskutiert werden. Diese beeinflussen im Folgenden die Handlungsbedingungen der Individuen (I1). Robert Merton nennt fünf Kernnormen, welche das Handeln von Forschenden gleichsam in Form eines Grundethos leiten: Universalismus, Kommunalismus, Uninteressiertheit, organisierter Skeptizismus (Merton 1973 Erstveröffentlichung als Artikel 1942) sowie Originalität (Merton 1957).

*Universalismus* meint dabei die Objektivität bzw. Werturteilsfreiheit der Forschenden selbst. So sollen dieser Norm zufolge Forschungsergebnisse unabhängig von den (sozialen) Eigenschaften der Forschenden, wie etwa (soziale) Herkunft oder Geschlecht, zugänglich gemacht und rezipiert werden. Merton illustriert dies an dem Beispiel einer „echt-deutsch[en]“ bzw. „hundred-percent American“ (Merton 1973: 270) Forschung, welche nach der Universalismusnorm nicht existieren kann. Diese völlige Loslösung von Personeneigenschaften lässt sich am Beispiel physikalischer Gesetze, wie der Schwerkraft, welche per se unabhängig von persönlichen Eigenschaften ihres Entdeckers[3] gelten, illustrieren. Anhand

[1] Nachfolgend wird der Akteursbegriff geschlechtsunabhängig verwendet.
[2] Dieses Mikro-Makro Modell wird in der Literatur aufgrund seiner Popularisierung durch James Coleman (Coleman 1990: 8) ‚Coleman's Badewanne' genannt, die Provenienz des Modells ist jedoch verschiedensten Autoren in der Soziologie zuzurechnen (s. Raub & Voss 2016 für eine detaillierte Diskussion).
[3] Isaak Newton (1643-1727)

dieses Vergleichs wird deutlich, dass es keine genuin amerikanische bzw. deutsche Schwerkraft geben kann. Die Universalismusnorm wirkt dabei auf die vorurteilsfreie Rezeption bestehender Forschung und deren Manifestation in Form von Zitationen sowie bei der Vergabe von beruflichen Positionen. Gleichstellungs-Policies bzw. allgemeine Gesetze zur Gleichstellung im Arbeitsmarkt sind hierbei institutionalisierte Formen der Universalismusnorm (vgl. § 7 Abs. 1 Allgemeines Gleichbehandlungsgesetz).

Die Norm des *Kommunalismus* hingegen bezeichnet die Pflicht der Zugänglichmachung von wissenschaftlichen Ergebnissen für die jeweilige Forschungsgemeinschaft, um darauf aufbauend weiteren Erkenntnisfortschritt zu ermöglichen. Die Bedeutung der Kommunalismusnorm (im Original „communism" Merton 1973: 271) lässt sich insbesondere an ihrem entgegengesetzten Konzept, dem Kapitalismus, einordnen (Macfarlane & Cheng 2008: 76). Anstatt wissenschaftliche Erkenntnisse für die Forschungsgemeinschaft verfügbar zu machen, wäre es im Sinne einer kapitalistischen Wissenschaftsnorm, analog zum Vorgehen bei Patenten, völlig selbstverständlich, Erkenntnisse, welche gegen das eigene Produkt sprechen sowie Informationen über das Produkt selbst, vor dem Zugriff anderer zu schützen, d. h. zu verheimlichen. Gerade bei theoriegeleiteter Forschung ist eine Verletzung der Kommunalismusnorm am ehesten an neu entwickelten Theorien zu erkennen, welche von den PrimärautorInnen als den ‚Erfindern' empirisch unablässig bestätigt werden, um ihre Theorie als ‚Produkt' am Leben zu halten. Gerade die vorangehend erwähnte Replikationskrise in der Psychologie ist hierfür ein gutes Beispiel, da oft bereits lang existierende Theorietraditionen, obschon falsch, gleichsam als „undead theories" (Ferguson & Heene 2012) in immerwährender, falsch positiver Bestätigung fortleben. Institutionelle Umsetzungen der Kommunalismusnorm lassen sich insbesondere im Bereich Open bzw. FAIR[4] Data (Wilkinson et al. 2016) finden, welche beginnen die Zugänglichkeit der Daten als Grundsatz guter Forschung zu etablieren (für die Soziologie z. B. Akademie für Soziologie 2019). Als institutionelle Umsetzung der Norm sind hier die bislang vereinzelten *Data Reproducibility Policies* einiger Zeitschriften zu nennen, die zugrunde liegenden Daten auf deren Richtigkeit zu überprüfen (Open Science Collaboration 2018).

Die dritte Norm der *Uninteressiertheit* bezieht sich auf die Einstellung der Forschenden ihrem Forschungsgegenstand gegenüber. Merton identifiziert die altruistische Suche nach wissenschaftlichem Fortschritt als antagonistisches Konzept zu den von eigenen normativen Ansichten geleiteten „pseudosciences" (Merton 1973: 277). Gerade der in den folgenden Kapiteln näher untersuchte Publication Bias stellt in dieser Hinsicht einen klaren Normverstoß dar, da in diesem Fall nach Bestätigung für (eigene) Theorien gesucht und eine Widerlegung zu vermeiden versucht wird. Eine institutionalisierte Form der Norm der Uninteressiertheit sind die in weiten Bereichen der Psychologie und Medizin bereits üblichen Pre-Registrierungen von Studien (Nosek & Lindsay 2018). Bei diesen werden bereits vor der Datenerhebung das genaue Studiendesign (insb. Fallzahlen, Power-Analyse) sowie die geplanten Auswertungen

[4] Akronym für: Findability, Accessibility, Interoperability, and Reusability (Wilkinson et al. 2016).

spezifiziert (Munafò et al. 2017). Da der Auswertungsablauf somit festgelegt ist, werden post-hoc Änderungen im Verlauf der Datenauswertung auf diese Weise zumindest sichtbar.

Beziehen sich alle bisher diskutierten Normen wissenschaftlichen Arbeitens auf den einzelnen Forschenden und seine Arbeit, kommt in der Norm des *organisierten Skeptizismus* eine stärkere Außenorientierung hinzu. Nun wird nicht mehr nur das eigene Handeln durch die Norm geleitet, sondern darüber hinaus auch das Handeln anderer hinterfragt. Organisierter Skeptizismus bedeutet in diesem Fall zwar auch die reflexive Selbstkritik von Forschenden gegenüber den eigenen Ergebnissen, jedoch insbesondere die kritische Würdigung anderer Forschender und ihrer Ergebnisse in der bestehenden Literatur. Merton führt als Antagonisten des organisierten Skeptizismus den unkritischen Respekt (Merton 1973: 278) an, welcher nicht an der Widerlegung bestehender theoretischer Konstrukte, sondern nur an deren konservierender Mumifizierung interessiert ist. Als institutionelle Normdurchsetzung können insbesondere die in der Physik verbreiteten *blind analyses* angeführt werden, welche die Datenanalyse zunächst nur an einem kleinen Teil der Daten ‚probehalber' durchführt und sobald die Modellierung an der Vor-Analyse erfolgreich abgeschlossen ist, sie auch auf den Hauptteil der Daten anwendet (MacCoun & Perlmutter 2015). In ähnlicher Weise sind an dieser Stelle die in jüngster Zeit vermehrt anzutreffenden Initiativen zur Replikation von Vorgängerstudien (Camerer et al. 2016; Open Science Collaboration 2015) anzuführen.[5] Diese sollen ebenso sicherstellen, dass der Befund über die Analysesituation hinaus übertragbar ist (z. B. durch neue Daten, geringe Veränderung der Operationalisierung).

Die letzte Norm der *Originalität*, eine von Robert Merton bemerkenswerterweise erst 1957 in den Kanon aufgenommene Norm, postuliert die Entdeckung, welche die wissenschaftliche Erkenntnis voranbringt, als Primärziel von Forschenden (Merton 1957: 639). Zentrales Momentum origineller Beiträge zum Forschungsfortschritt ist vor allem die Möglichkeit, die eigenen Forschungsergebnisse als erstes in diesem „race for priorities" (Merton 1970: 213) zu veröffentlichen. Das Postulat der Originalität von Forschungsergebnissen ist über die verschiedenen Disziplinen divers definiert, so unterscheiden sich selbst die Geistes- und Sozialwissenschaften mit einem eher auf die generelle Herangehensweise sowie auf die Methode gewählten Fokus deutlich (Guetzkow et al. 2004: 210). Ferner kann auch eine nicht bestätigte Hypothese als originell definiert werden. Dies geschieht beispielsweise in speziellen Zeitschriften für Nullergebnisse, wie sie in den letzten Jahren insbesondere in Teilen der Psychologie bzw. Biomedizin entstanden sind. Diese Zeitschriften haben es sich zur Aufgabe gemacht, nicht theoriekonforme sowie nicht-signifikante Ergebnisse zu veröffentlichen.[6]

Diese auf die Wissenschaftler selbst einwirkenden fünf Normen können anhand zweier Dimensionen aufgeteilt werden: die Herangehensweise der ForscherInnen selbst sowie die Einschätzung von Fremdwerken oder Personen betreffend. So sind die Normen des Universalismus, des Kommunalismus und

[5] vgl. auch speziell auf Replikationen spezialisierte Zeitschriften wie *International Journal for Re-Views in Empirical Economics (IREE).*

[6] z. B. *Journal of Articles in Support of the Null Hypothesis* (*JASNH), Journal of Negative Results in Biomedicine* (*JNRB).*

der Uninteressiertheit stark intrapersonal ausgerichtet. Die Norm des systematischen Zweifels wie auch der Originalität sind eher als Maßstab zur Beurteilung der Forschungsleistung Dritter anzusehen.

### *1.2.2. Mikroebene*

#### *Einordnung in die Rational Choice Theorie*

Auf der Mikroebene I1 (in Abbildung 1-1) wirken nun diese Normen aus Ebene M1 auf die Forschenden (nachfolgend Akteure) ein. Im Folgenden wird in der theoretischen Einordnung von zielgerichtet handelnden Akteuren ausgegangen, die danach streben ihren Nutzen zu maximieren (Coleman 1990: 17-19). Diese rationale Nutzenmaximierung schließt dabei nicht-rationale Handlungen nicht grundsätzlich aus (z. B. das Modell der begrenzten Rationalität Simon 1983), vereinfacht das handlungstheoretische Modell durch seine analytische Klarheit jedoch eminent und ermöglicht so erst eine nachvollziehbare Modellierung. Um eine schematische Einordnung zu ermöglichen, liegt der Fokus nachfolgend ausschließlich auf rationalem Handeln.

Für das Konzept des nutzenmaximierenden Akteurs erscheint es zunächst zentral, die Nutzendimensionen zu klären, welche der Akteur zu maximieren versucht. Wie bereits bei den Normen auf der Makroebene angedeutet, konstruiert in Form der Erstentdeckung insbesondere die Norm der Originalität eine Zieldimension für die Forschenden (Merton 1957; Diamond 1996: 8). Die Erstentdeckung wird dabei durch einen publizierten Beitrag im Erkenntnisfortschritt manifestiert. Erst die Publikation ermöglicht es Anerkennung der Forschungsgemeinschaft in Form von Zitationen der jeweiligen Publikation zu erlangen. Ein Kernproblem besteht insbesondere darin, dass in besonders stark beforschten Gebieten eine große Konkurrenz um Erstentdeckungen herrscht, da sie mit einem großen Reputationsgewinn einhergehen, während in kleineren Forschungsgebieten auch ein solches Ergebnis bedingt durch den hohen Spezialisierungsgrad allzu oft kaum wahrgenommen wird (Stephan 2010: 233). Für den Erfolg im Wissenschaftssystem essentiell ist zudem eine möglichst große Sichtbarkeit der Publikation („publish or perish“ vgl. Dijk et al. 2014; Skeels & Fairbanks 1968). Doch unabdingbare Vorbedingung für jeden originellen Artikel ist immer die Publikation desselben. Ist ein Forschungsbeitrag gar nicht publiziert, d. h. nicht der Forschungsgemeinschaft zugänglich gemacht, ist sowohl der Erkenntnisfortschritt als auch dessen Anerkennung durch die Forschungsgemeinschaft unmöglich.

Statistisch signifikante Ergebnisse werden häufig als ein Signal für Originalität und damit Qualität der Studie gewertet. Dies liegt zum einen an der Logik von Signifikanztests: So weist ein signifikantes Ergebnis mit einem *p*-Wert von beispielsweise 0,01 darauf hin, dass im Falle keines zugrundeliegenden wahren Effekts, die vorliegende Datenstruktur nur mit einer Wahrscheinlichkeit von 1% beobachtet werden kann. Bei nicht-signifikanten Ergebnissen ist hingegen keine solch klare Zuweisung möglich, da ein nicht-signifikanter *p*-Wert eben nicht auf das Fehlen eines zugrundeliegenden Effekts hindeutet. Nicht-signifikante Ergebnisse können zudem auf Mängel im Studiendesign insbesondere bei der statistischen Power (Cohen 1988) zurückzuführen sein. Signifikante Ergebnisse wirken daher oft als ein Signal vermeintlicher Forschungsqualität (vgl. die signaling theory von Spence 1973). Empirisch lässt sich

dies insbesondere an den höheren Zitationen von statistisch signifikanten Ergebnissen zeigen (Fanelli 2013).[7]

Um ihre Reputation zu maximieren können Forschende in Versuchung kommen, gegen eben jene Leitnormen insbesondere die Kommunalismusnorm als auch die Norm der Uninteressiertheit, z. B. durch Praktiken des Publication Bias, zu verstoßen. Im Folgenden soll dementsprechend die Theorie abweichenden Verhaltens (Becker 1968) näher betrachtet werden. Dieser zur Folge scheint ein normkonträres, abweichendes Verhalten für den Akteur genau dann lohnend, wenn der Erwartungswert möglicher Gewinne (*EU*) die erwarteten Verluste in Form von Sanktionen übersteigt. Für den vorliegenden Fall wissenschaftlichen Fehlverhaltens gilt, dass sich mögliche Sanktionen (*f*) nur im Falle einer Entdeckung (in diesem Fall in der Entdeckungswahrscheinlichkeit *p*) realisieren. Erst in diesem Fall verringert die Sanktion (*f*) die durch das abweichende Verhalten erzielten Gewinne (*Y*). Im Falle der Nichtentdeckung hingegen fallen nur die Gewinne ohne mögliche Sanktionen an. Der erwartete Nutzen des abweichenden Verhaltens lässt sich nach Becker (1968: 203) demnach formulieren als:

$$EU_j = p_j U_j\left(Y_j - f_j\right) + \left(1 - p_j\right) U_j(Y_j)$$

Als für die Wissenschaft besonders problematische Parameter erweisen sich einerseits die Sanktion (*f*) als auch die Entdeckungswahrscheinlichkeit (*p*). So resultiert zwar die entdeckte Fälschung von wissenschaftlichen Ergebnissen in einem Ausschluss aus dem Wissenschaftssystem (Levelt Commitee et al. 2012) bzw. sogar dem Verlust des akademischen Titels (Universität Konstanz 2013), die Entdeckungswahrscheinlichkeit für solch ein Fehlverhalten ist jedoch durch fehlende institutionelle Entdeckungsmechanismen eher gering. Zudem existieren für den Fall einer Entdeckung keine dezidierten Sanktionsmechanismen. Feigenbaum & Levy (1996) postulieren deswegen gar die Überflüssigkeit von Fälschungen, da das Ergebnis einer erfolgversprechenden Publikation auch auf anderen, kaum sanktionierten Wegen erreichbar sei, welche nichtsdestotrotz gegen die durch die Normen spezifizierte gute wissenschaftliche Praxis verstoßen. In der Simulationsstudie von Bakker et al. (2012) waren insbesondere kleinere Formen wissenschaftlichen Fehlverhaltens wie Formen des Publication Bias besonders effektiv, um mit einer hohen Wahrscheinlichkeit signifikante Ergebnisse zu ‚produzieren'. Obwohl Praktiken des Publikation Bias sowohl auf Seite der AutorInnen der Studien als auch auf Seite der jeweiligen HerausgeberInnen der Zeitschriften sowie der GutachterInnen möglich sind, gibt es in der bestehenden Literatur Evidenz dafür, dass insbesondere AutorInnen nur signifikante Ergebnisse verschriftlichen und zur Publikation einreichen (Franco et al. 2014). Hinzu kommt jedoch auch eine leichte Evidenz für die Selektion von signifikanten Ergebnissen bei der Begutachtung durch die Forschungsgemeinschaft (Epstein 1990,2004; Mahoney 1977). Für die Entscheidung der Herausgebenden lässt sich hingegen keine Bevorzugung von signifikanten Ergebnissen feststellen (Olson et al. 2002).

[7] Über die Reputation wissenschaftlicher Peers in Form von Zitationen hinaus sind zudem auch wissenschaftliche Preise zu nennen, welche besondere Aufmerksamkeit auf ausgezeichnete ForscherInnen lenken.

Bei Berufungsverfahren liegt eine Abkehr vom Prinzip des Universalismus vor, wenn Bewerberinnen am Zugang zu Professuren gehindert werden. Jedoch steht die Diskriminierung von Bewerberinnen aufgrund ihres Geschlechts in Form des Allgemeinen Gleichbehandlungsgesetzes (insb. § 7 Abs. 1) generell im gesamten Arbeitsrecht und damit auch beim Arbeitgeber Universität unter Sanktionsandrohung. Im Gegensatz zum Publication Bias ist daher eine explizite, sogar in Gesetzesform gegossene Sanktionsmöglichkeit vorhanden, welche durch ihre breitere Anwendung über den universitären Rahmen hinaus für größere Rechtssicherheit sorgt. Diese rechtlichen Rahmenbedingungen sind im Hochschulbereich zusätzlich durch eine erhöhte Entdeckungswahrscheinlichkeit gekennzeichnet, da eine Gleichstellungsbeauftragte[8] als direkt dem Rektorat zugeordnete Stelle den Fortgang und korrekten Ablauf der jeweiligen Berufungsverfahren vor allem im Hinblick auf mögliche Diskriminierungen überwacht (beispielsweise §4 Abs. 3 Landeshochschulgesetz Baden-Württemberg). Der konkreten Sanktion steht somit zudem eine doch substantielle Entdeckungswahrscheinlichkeit gegenüber, auch wenn diese mehr den Arbeitgeber, d. h. die Institution Universität selbst in Form von Schadenersatz für eine diskriminierende Ablehnung der Bewerbung trifft, denn konkret handelnde Personen, in diesem Fall Mitglieder der Berufungskommission.

***Spieltheoretische Einordnung***

In der bisherigen theoretischen Fassung der Handlungsmotive der Akteure wurde nur auf die einzelnen Akteure und deren Handlungsmotivation, nicht jedoch auf deren Interdependenzen eingegangen. Im Fall des Publikationsverhaltens von Forschenden, welches von Publikation Bias geprägt sein kann, sowie den Berufungen auf Professuren an Universitäten, welche durch Geschlechterdiskriminierung gekennzeichnet sein können, liegt eine insbesondere durch die Spieltheorie modellierbare Interaktivität der Handlungswahl zugrunde (vgl. gepunktete Linie in Abbildung 1-1). So sind alle Akteure als Teil der Wissenschaft von der Erosion der wissenschaftlichen Glaubwürdigkeit, wie eingangs skizziert, betroffen. Im Folgenden sollen zuerst die möglichen Handlungen in einer dichotomen Entscheidungssituation mit zwei Akteuren modelliert werden und darauf aufbauend ein Modell mit mehreren Akteuren, wie es die Wissenschaft realistischerweise darstellt, diskutiert werden.

Den im Zweipersonenfall modellierten Akteuren, A (Tabelle 1-1 & Tabelle 1-2 Spalten) und B (Tabelle 1-1 & Tabelle 1-2 Zeilen, unterstrichen), liegen dabei zwei Handlungsalternativen zugrunde, welche die vorangehend diskutierten Entscheidungen abbilden: Entweder die Anerkenntnis und Befolgung einer Norm (C – *Cooperation*) oder die Verweigerung normkonformen Verhaltens (D – *Defection*). Normkonform wäre im Fall des im Folgenden zunächst diskutierten Publication Bias die Veröffentlichung aller Forschungsergebnisse unabhängig ihres Ergebnisses und im Fall der anschließend besprochenen Berufungsverfahren die Einstellung unabhängig vom Geschlecht der BewerberInnen.

[8] In den jeweiligen Landeshochschulgesetzen unterschiedlich benannt.

**Tabelle 1-1 Publication Bias – Auszahlungsmatrix**

| | | Akteur A | |
|---|---|---|---|
| | | C (Normkonform) | D (Publication Bias) |
| Akteur B | C (Normkonform) | 4 / 4 (R) **PO** | 1 / 5 (T) |
| | D (Publication Bias) | 5 / 1 (S) | 2 / 2 (P) **NG** |

Publication Bias ist für sich genommen, wie bei der handlungstheoretischen Einordnung ausgeführt, mit geringen (Sanktions)kosten und einem positiven Nutzen gegenüber dem normkonformen Verhalten ausgestattet. Wie jedoch verhält sich diese Präferenz, wenn sie in einen größeren Handlungskontext und in Abhängigkeit der Entscheidung einer anderen Person eingebettet wird? In einer normkonformen Umgebung (R), in der beide Akteure, A und B, an Kooperation interessiert sind, profitieren beide Akteure, denn einerseits leidet die Glaubwürdigkeit der Wissenschaft nicht, andererseits ist der Wettbewerb zwischen den beiden Akteuren fair in der Hinsicht, dass kein Akteur auf Kosten des anderen profitiert. Da beide Akteure den gleichen Lohn für ihre kooperativen Mühen erhalten, wird diese Situation auch als *Reward* (R) bezeichnet (vgl. Bravetti & Padilla 2018: 1 für die Terminologie). Die im Beispiel gewählte Auszahlungsmatrix würde eine Belohnung von 4 Punkten pro Akteur erbringen.[9] Die Punkte können generell sichtbare Leistung wie z. B. Veröffentlichungen, Zitationen sowie Preise abbilden.

Eine solch gleichmäßige Aufteilung ist nicht im Fall der Diagonale T/S gegeben. Im ersten Fall (T) kooperiert Akteur A nicht, d. h. er begeht Publication Bias, wohingegen Akteur B weiterhin normkonform handelt. Diese Situation wird auch *Temptation* (T) genannt, da Akteur A der Versuchung erliegt aus dem kooperativen normkonformen Verhalten von Akteur B auf dessen Kosten Kapital zu schlagen. Am Beispiel würde dies bedeuten, dass die durch Publication Bias erzielten Ergebnisse des Akteurs A eine größere Sichtbarkeit erfahren, demzufolge stärker honoriert werden, wie die zwar normkonform erzielten, jedoch mit einer höheren Wahrscheinlichkeit unspektakulären Nullergebnisse des Akteurs B. Im zweiten Fall (S) kooperiert hingegen Akteur A, jedoch erliegt nun Akteur B der Versuchung und kann durch Mittel des Publikation Bias analog zur Situation T einen Vorteil in diesem Fall auf Kosten des Akteurs A erlangen. Die Fokalperson A erleidet durch ihr kooperatives Verhalten wie in der vorherigen Situation auch schon B einen Verlust, den *Suckers Payoff* (S) oder euphemistischen ‚Lohn des Gutgläubigen'. In der Situation P (*Punishment*) entscheiden sich beide Akteure, sowohl A als auch B, gegen ein normkonformes Verhalten und erhalten analog zur beidseitigen kooperativen Situation die gleiche, wenn auch durch die beidseitige Defektion geringere Anzahl an Punkten. Am Beispiel des Publication Bias wäre diese Situation als völlige Anomie aufzufassen, in der beide Akteure durch ihr nicht kooperatives Verhalten die Glaubwürdigkeit und den Einfluss der Wissenschaft bzw. ihres Fachgebiets mindern und sich so mittelbar selbst schädigen.

Da in der spieltheoretischen Grundkonstellation keine Deliberation über das eintretende Verhalten der anderen Akteure stattfinden kann, lassen sich die vier Auszahlungssituationen in eine ordinale Reihung

[9] Die gewählten Auszahlungsbeträge sind willkürlich gewählt. Die Modelle unterscheiden sich jedoch nicht in ihrem Ergebnis, solange die ordinale Rangordnung der verschiedenen Alternativen (s. u.) gewahrt bleibt.

bringen: T>R>P>S. Die Defektion ist bei gleichzeitiger Kooperation des Spielpartners (T) am lohnendsten, die beidseitig kooperative Variante (R) ist die zweite Präferenz. Beide sind erheblich vom Verhalten des Spielpartners abhängig. Wählt der Spielpartner die Defektion, wird aus der zweitbesten Spielsituation die schlechteste (S). Würde A nun wissen, dass B sich nicht normkonform verhält, so müsste A um extreme Einbußen zu vermeiden ebenfalls auf eine nicht normkonforme Handlungsalternative, in diesem Fall Publication Bias, umschwenken. Im Falle einer bereits normverletzenden Alternative wäre ein solcher, den Verlust minimierender Strategiewechsel hinfällig. Von beiden Akteuren begangener Publikation Bias als normverletzende Handlungsalternative (P) ist in diesem Fall daher das Equilibrium, (auch bekannt als Nash-Gleichgewicht NG, Nash 1950). Neben der Maximierung der jeweiligen individuellen Auszahlung führt die Pareto-Optimalität ein Globalkriterium ein, bei dem sich ein Spieler nicht weiter verbessern kann, es sei denn auf Kosten eines anderen Spielers (Diekmann 2013: 34). Im vorliegenden Fall ist das beidseitige normkonforme Verhalten (R), in welchem beide Akteure die Forschungsergebnisse ohne Ansehen der erzielten Ergebnisse publizieren und der Forschungsgemeinschaft zugänglichmachen, Pareto-optimal (PO).

Die im Falle des ausgeführten Beispiels vorliegende Situation erfüllt dabei die Grundannahmen eines Gefangenendilemmas (vgl. Hamburger 1973: 30). So hat jeder der beiden Spieler eine durch das Nash-Gleichgewicht determinierte klare dominante Strategie, im konkreten Fall würde dies Publication Bias implizieren. Die für beide Spielpartner abschreckendste Situation als worst-case Szenario ist dabei auch die jeweils optimale Strategie des Spielpartners. Das daraus resultierende Nash-Gleichgewicht (NG) ist dabei stabil, sodass kein Akteur einseitig abweichen würde. Zugleich ist es jedoch defizitär, da die dominante Strategie im Kollektiv immer zu niedrigeren Auszahlungen führt ($2*P < (T+S) < 2*R$) und daher global ineffizient ist (Hamburger 1973: 33).

Das bisherige Modell hat jedoch einen gravierenden Mangel, da es nur für zwei Spieler ausgelegt ist, ein im Falle des Wissenschaftssystems eher abwegiges Szenario. Vielmehr treten im Wissenschaftssystem zahlreiche Akteure im Wettbewerb um Anerkennung in Form von Publikationen und Zitationen gegeneinander an. Eine einfache Art dies zu modellieren ist das N-Personen Gefangenendilemma (Hamburger 1973; Schelling 1973), welches das vorhergehende Spiel unter gleichen Voraussetzungen, d. h. ohne strategisches Lernen sowie Deliberation mit weiteren Akteuren, durchspielt. Bei N Akteuren (N=5 in Abbildung 1-2) gibt es daher N-1 Spielsituationen (5-1=4). Für jeden Akteur lässt sich dabei der Erwartungswert der Punkte berechnen, gegeben der Entscheidungen der jeweils anderen Akteure. Der Erwartungswert des Auszahlungsbetrags über alle Runden für normkonformes Verhalten C(n) definiert sich daher über den Erwartungswert der Erträge mit kooperativen Akteuren (n) sowie unkoope-

rativen Akteuren (N-n). Mit kooperativen Akteuren (n) ist gegeben der eigenen Kooperation der maximale globale Auszahlungsbetrag R zu erreichen. Mit unkooperativen Akteuren (N-n) ist nur der geringste Auszahlungsbetrag S zu realisieren. Es gilt daher:[10]

$$C(n) = (n-1)R + (N-n)S$$

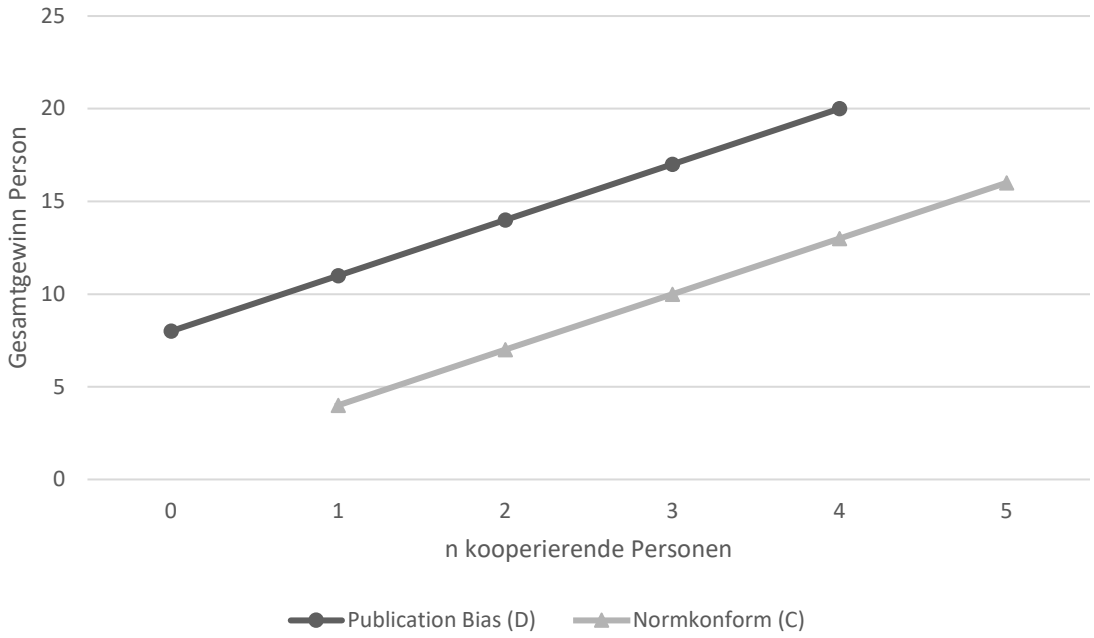
Der Ertrag aus defektierendem Verhalten (Publication Bias) berechnet sich analog. So ist mit n kooperierenden Spielpartnern der maximal mögliche individuelle Ertrag T zu erzielen. Bei N-n-1 Spielpartnern, welche ebenso wie der fokale Akteur defektieren, ist nur der geringere Auszahlungsbetrag P zu erreichen:

$$D(n) = nT + (N-n-1)P$$

Das Ergebnis lässt sich eindrücklich an Abbildung 1-2 nachvollziehen, in der die Auszahlungsfunktion der defektierenden Handlungsoption (Publication Bias) stets über der normkonformen liegt. Insbesondere ist hier hervorzuheben, dass wie in der Version mit zwei Personen der defektierende Akteur von der Kooperation weiterer Akteure profitiert. Das heißt, halten sich alle Forschenden an die Norm, lohnt es sich für den einzelnen Forschenden einseitig von der Norm abzuweichen, d. h. Publication Bias zu begehen. Die Version des Gefangenendilemmas mit mehreren Akteuren hat dabei zwei zusätzliche Grundannahmen in den beiden Auszahlungsfunktionen (Schelling 1973: 386), zum einen ist eine positive Steigung der Auszahlung zu erwarten je mehr Personen kooperieren, zum anderen dürfen sich beide Auszahlungsfunktionen nicht schneiden. Die erste Annahme weist darauf hin, dass sich die Kooperation anderer Akteure sich sowohl für kooperierende als auch nichtkooperierende Akteure lohnt: Für kooperierende Akteure, um den Auszahlungsbetrag (R) im Fall von beidseitiger Kooperation zu erlangen, im Fall von Defektion, um den maximalen Auszahlungsbetrag (T) anstatt des nichtkooperativen Nash-Gleichgewichts zu erhalten. Die zweite Annahme bezieht sich auf den Kern des Gefangenendilemmas der höheren Auszahlung im Fall von Defektion. Diese Grundannahme, dass Defektion in allen Spielsituationen zu überlegenen Auszahlungen führt, muss daher gewahrt bleiben, da sonst kein ineffizientes Nash-Gleichgewicht mehr bestünde.

[10] Zur besseren Verständlichkeit wurden die Formeln umgestellt, sie entsprechen jedoch der Auszahlungsfunktion von Hamburger (1973: 38).

**Abbildung 1-2 Publication Bias – Auszahlungen N-Personen Gefangenendilemma (*N* = 5)**




Das hier zur theoretischen Modellierung des Publication Bias herangezogene N-Personen Gefangenendilemma ist ein Spezialfall der Klasse der sozialen Dilemmata (vgl. eine entsprechende Einordnung des Publication Bias in Auspurg & Hinz 2017). Gemeinsamkeit aller sozialen Dilemmata ist das trotz rationalen Verhaltens auf der Mikro- bzw. Individualebene ineffiziente Ergebnis (Dawes 1980; Dawes & Messick 2000; Kollock 1998). Das alleinige Nash-Gleichgewicht auf der beidseitig defektierenden, d. h. Publication Bias begehenden Handlungsoption ist jedoch im Gegensatz zu anderen Spielsituationen wie etwa dem Chicken- und dem Assurance-Spiel nur beim N-Personen Gefangenendilemma zu finden (Kollock 1998: 187). Konkret bedeutet dies, dass im Fall des Publication Bias nur im Gefangenendilemma ein flächendeckendes Fehlverhalten die Folge ist. Die ebenfalls stark mit dem Gefangenendilemma verwandte Tragik der Allmende (Hardin 1968) hat eine gleiche Auszahlungsmatrix wie das Gefangenendilemma, d. h. jeder Anteilseigner der Allmende hat den Anreiz das eigene Vieh möglichst lange auf der Allgemeinweide grasen zu lassen, bis diese komplett leergefressen und damit die Ressourcen verbraucht sind. Diese Annahme begrenzter Ressourcen ist im Fall des Publication Bias jedoch nicht plausibel, da weder Zitationen noch weitere Anerkennungen wie im Fall der Allmende ‚übernutzt' werden können. Im vorliegenden Fall liegt daher klar ein Gefangenendilemma zugrunde.

Neben der Veränderung der Auszahlungsstruktur, welche es für den individuellen Akteur weniger attraktiv macht Publication Bias zu begehen, stehen zusätzlich zwei Lösungswege offen: die Konkurrenz mehrerer unabhängiger Subsysteme sowie die wiederholte Interaktion von Akteuren unter transparenten Vorinteraktionen.

Der erste Lösungsweg schließt an die stark differenzierte moderne Wissensproduktion an, welche insbesondere im Bereich der Natur- und in zunehmendem Umfang auch der Sozialwissenschaft nicht ausschließlich auf den akademischen Kontext begrenzt ist (Hessels & van Lente 2008). Vielmehr findet die Wissensproduktion in sich hybridisierenden Feldern auch in mehr oder weniger akademisch geprägten Forschungsinstitutionen oder wie im Bereich der Pharmaforschung z. T. komplett im freien Markt statt. Darüber hinaus besteht nicht nur ein Wettbewerb zwischen wissenschaftlichen und eher anwendungsbezogenen Feldern der Wissensproduktion, sondern auch zwischen verschiedenen Wissenschaftsdisziplinen. So werden beispielsweise sozialwissenschaftliche Themen zunehmend auch von NaturwissenschaftlerInnen, insbesondere PhysikerInnen bearbeitet (z. B. in der sozialen Netzwerkanalyse). Die durch das Gefangenendilemma entstehende Ineffizienz birgt, in Anlehnung an die Evolutionsbiologie, daher die Gefahr, dass andere effizientere und in diesem Fall kooperierende Populationen sich durchsetzen (Bravetti & Padilla 2018). Auf die Wissenschaft übertragen bedeutet dies, dass, sollten sich die Ineffizienzen wie das Problem des Publikation Bias nicht durch Kooperation lösen lassen, zumindest die Gefahr besteht, dass sich andere wissenschaftsimmanente oder wissensgenerierende Institutionen, welche weniger ineffizient sind, sich durchsetzen und so den Einfluss der einzelnen Wissenschaftsdisziplin bzw. der Wissenschaft insgesamt auf die (gesellschaftliche) Entwicklung zurückdrängen.

Die zweite Lösung des Kooperationsproblems und der daraus erwachsenen Ineffizienz kann durch eine Erweiterung des Gefangenendilemmas auf Situationen mit wiederholten Spielen und transparenten Verhaltensweisen in Vorinteraktionen erreicht werden. Zentral ist dabei, dass sich das Verhalten eines Akteurs immer am vorhergehenden Verhalten des jeweils anderen Akteurs orientiert. Im Fall der zwei Runden des Computerturniers von Axelrod hat sich die TIT FOR TAT Regel Rapoports, welche das Vorverhalten der jeweiligen Vorrunde spiegelt, jedoch kooperierend beginnt, als überlegen herausgestellt (Axelrod 1980a,b). Übertragen auf das Problem des Publication Bias bedeutet dies, dass, sofern Signale über ein normentsprechendes Verhalten verfügbar sind, konditional mit Kooperation geantwortet werden sollte.

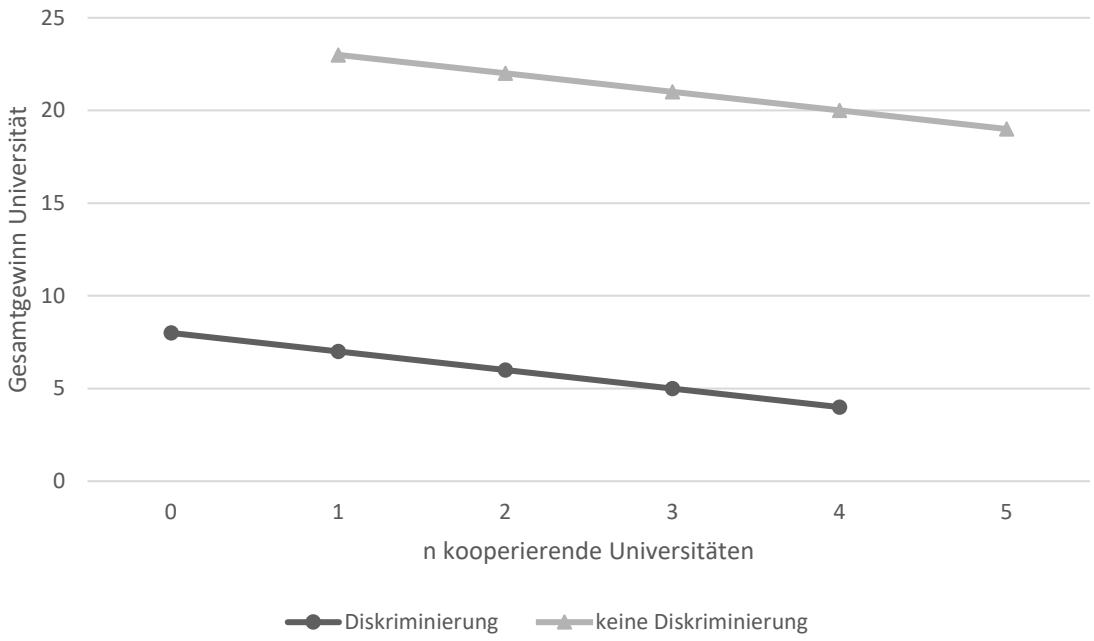
Im Gegensatz zur vorangehend erläuterten Ineffizienz im Wissenschaftssystem im Falle des Publikation Bias liegt die Ineffizienz im Falle der Benachteiligung von Bewerberinnen beim Zugang zu Professuren gänzlich anders. Die Diskriminierung von Bewerberinnen bei gleicher Leistungsfähigkeit aufgrund der Ablehnung ihrer askriptiven Eigenschaft Geschlecht, auch „tastes for discrimination“ (Becker 1971) genannt, bringt stets die Kehrseite mit sich, auf weniger qualifizierte Bewerber ohne diese askriptive Eigenschaft zurückgreifen zu müssen. Diskriminierende Arbeitgeber, in diesem Fall die Universität, müssen demzufolge ein geringeres Leistungslevel bei der gewünschten Personengruppe in Kauf nehmen, da kompetentere Personen aufgrund ihrer askriptiven Eigenschaft abgelehnt werden. Eine Präferenz männlicher Bewerber schädigt somit den diskriminierenden Akteur (Universität) sowie die diskriminierten Bewerberinnen und begünstigt die aufgrund der Diskriminierung bevorzugten männlichen Bewerber sowie die nicht diskriminierenden Akteure (Becker 1971: 22).

**Tabelle 1-2 Diskriminierung – Auszahlungsmatrix**

| | | Universität A | |
|---|---|---|---|
| | | C (Chancengleichheit) | D (Diskriminierung) |
| Universität B | C (Chancengleichheit) | 4 / 4 **PO, NG** | 5 / 1 |
| | D (Diskriminierung) | 1 / 5 | 2 / 2 |

Die Auszahlungsstruktur wie im Beispiel des Publikation Bias verkehrt sich hier ins genaue Gegenteil (vgl. Tabelle 1-2): Diskriminieren einzelne Akteure gelangt die jeweils nicht diskriminierende Universität zu kompetenteren Beschäftigten und damit einer höheren Auszahlung (5), wohingegen die diskriminierende Universität an Wettbewerbsfähigkeit verliert (1). Im Fall von jeweils diskriminierenden Universitäten wäre wie beim Beispiel des Publikation Bias der jeweils geringste kollektive Auszahlungsbetrag und damit die ineffizienteste Lösung gegeben. Im Gegensatz zu den vorangehend erörterten Dilemmasituationen ist die global optimalste Verhaltensweise, gänzlich unabhängig vom Verhalten der jeweils anderen Universität, das Unterlassen von Diskriminierung. Sowohl der höchste globale Auszahlungsbetrag (PO) als das Nash-Gleichgewicht (NG) liegen hier auf dem nicht-diskriminierenden Verhalten.

**Abbildung 1-3 Diskriminierung – Auszahlungen N-Personen Spiel (*N* = 5)**




Auch die Diskriminierung von Bewerberinnen im Zugang zu Professuren kann in die N-Personenstruktur übertragen werden (Abbildung 1-3). Die Auszahlungsfunktion des defektierenden Handlungsoption Diskriminierung liegen immer unter denen der kooperierenden Handlungsoption. Hier wird deutlich,

dass diskriminierende Akteure deutliche Kosten ihrer Diskriminierung zu tragen haben und sich das Unterlassen von Diskriminierung zu jedem Zeitpunkt auszahlt.[11]

## 1.3. Zusammenfassung und Einordnung der Beiträge

Die vorgelegte Dissertation setzt sich aus drei bereits veröffentlichten Zeitschriftenartikeln sowie einem unveröffentlichten Workingpaper zusammen (vgl. Tabelle 1-3). Bei den drei publizierten Beiträgen handelt es sich um zwei Artikel, die in der deutschsprachigen Soziologie (Auspurg et al. 2014 – Kapitel 2; 2017 – Kapitel 5) veröffentlicht sind, sowie einen Artikel, der in einer interdisziplinären, insbesondere in der Psychologie und Medizin verankerten Zeitschrift publiziert ist (Schneck 2017 – Kapitel 3).

**Tabelle 1-3 Übersicht über die Beiträge und die Eigenanteile**

| Kapitel | **Zeitschrift** | **Impact Factor** [a] | **KoautorInnen** | **Fremd-anteil** [b] | **Eigen-anteil** | **Gewich-tung** [c] | **Faktor** [d] |
|---|---|---|---|---|---|---|---|
| 2. | KZfSS [e] | 0,783 | Katrin Auspurg, Thomas Hinz | 20%, 20% | 60% | 1,5 | 0,9 |
| 3. | PeerJ | 2,469 | | | 100% | 1,5 | 1,5 |
| 4. | Working-paper | | | | 100% | 1 | 1 |
| 5. | ZfS [f] | 0,656 | Katrin Auspurg, Thomas Hinz | 30%, 20% | 50% | 1,5 | 0,75 |
| Total | | | | | | | 4,15 |

[a] 5-Jahres Impact Faktoren aus dem *Journal Citation Reports* der Zeitschrift im jeweiligen Veröffentlichungsjahr; [b] In der Reihung der KoautorInnen; [c] Für die Gewichtungsfaktoren vgl. Betreuungsvereinbarung; [d] Nach der Betreuungsvereinbarung ist in der Totale ein Faktor > 3 im Rahmen einer Dissertation zu erbringen; [e] Kölner Zeitschrift für Soziologie und Sozialpsychologie; [f] Zeitschrift für Soziologie.

Für die beiden Artikel, welche mit Koautoren verfasst wurden, sollen nachfolgend die jeweiligen Arbeitsanteile kurz dargelegt und begründet werden. Im Falle von Kapitel 2 hat der Verfasser der Dissertation einen 60 prozentigen Arbeitsanteil an der Studie vorzuweisen. Aufbauend auf die Vorläuferstudie von Auspurg & Hinz (2011a) wurde vom Verfasser die Datenbasis von 50 kodierten Artikel auf letztlich 108 Artikel erheblich erweitert. Der Verfasser wertete ferner die Daten aus und schrieb die erste Fassung des Artikels. Die weitere Ausarbeitung des Artikels wurde dabei von den beiden Koautoren unterstützt (jeweils 20%). Im Falle des Artikels zu den Berufungsverfahren (Kapitel 5) wurde vom Verfasser die umfangreiche Datenerhebung, welche sowohl die Digitalisierung von Archivdaten als auch das Zuspielen bibliometrischer Daten umfasste, unter Mithilfe von Hilfskräften vorgenommen. Der Verfasser wertete die gewonnenen Daten aus und verfasste den Daten & Methoden- und den Ergebnisteil des Artikels (50%). Katrin Auspurg verfasste die Einleitung, den Forschungsstand sowie den Theorieteil (30%). Thomas Hinz übernahm den Diskussionsteil sowie die finale Durchsicht des Manuskripts (20%).

[11] An dieser Stelle sei angemerkt, dass die Auszahlungsmatrix nur für den Fall von „tastes for discrimination" (Becker 1971) gilt. Im Falle von statistischer Diskriminierung (Arrow 1971; Phelps 1972), d. h. der rationalen Diskriminierung aufgrund von fehlender Information, welche durch bekannte Gruppenparameter ersetzt wird, wäre die diskriminierende Alternative sowohl das Pareto-Optimum als auch das Nash-Gleichgewicht.

### *1.3.1. Ausmaß und Risikofaktoren des Publication Bias in der deutschen Soziologie*

Der erste Beitrag der Dissertation (Kapitel 2) schließt in Erweiterung des Artikels von Auspurg & Hinz (2011a) an die Identifikation von Publikation Bias in der deutschen Soziologie sowie dessen Risikofaktoren an. Als Stichprobe wurden die beiden größten deutschen Soziologiezeitschriften, die Zeitschrift für Soziologie und die Kölner Zeitschrift für Soziologie und Sozialpsychologie, im Zeitraum von 2000-2010 untersucht. Die methodische Herausforderung der Untersuchung bestand in den sehr heterogenen Fragestellungen innerhalb der Soziologie. Eine Untersuchung von Publication Bias wie in effekthomogenen Meta-Analysen vorgeschlagen (vgl. den Sammelband von Rothstein et al. 2005 für adäquate Tests) war somit nicht möglich. Wie schon in der Vorläuferuntersuchung von Auspurg & Hinz (2011a) wurde die Verteilung der erfassten statistischen Testwerte mittels des Caliper-Tests (Gerber & Malhotra 2008a) auf Auffälligkeiten untersucht. Der Caliper-Test untersucht hierbei nur einen kleinen Abschnitt der Testwerte rund um das interessierende Signifikanzniveau, um welches Publication Bias erwartet wird. Üblicherweise wird der Caliper relativ zum Schwellenwert des Signifikanzniveaus definiert (z. B. 5% um das 5%-Signifikanzniveau).[12] Eine Häufung von knapp signifikanten Werten im Verhältnis zu knapp nicht-signifikanten Werten kann dabei als Anzeichen von Publication Bias gesehen werden, da obgleich die genaue Verteilungsform der Testwerte unbekannt ist, doch eine Stetigkeit der Testwerteverteilung angenommen werden kann, welche in einem kleinen Intervall zu einer Gleichverteilung der knapp signifikanten und knapp nicht-signifikanten Werte um einen willkürlichen Schwellenwert wie das *a priori* gesetzte Signifikanzniveau führen sollte (ähnlich eines Regression Discontinuity Designs vgl. Lee & Lemieux 2010).

Für Testwerte um das am weitesten verbreitete 5%-Signifikanzniveau (Cohen 1994) ließ sich eine überzufällige Häufung knapp signifikanter Ergebnisse feststellen, welche auf Publication Bias hindeuten. Für das weniger verbreitete 10%-Signifikanzniveau hingegen konnte nur eine deutlich schwächere Tendenz ausgemacht werden. Neben der Diagnose von Publication Bias wurden auch die Risikofaktoren, welche diesen bedingen, untersucht. Aufgrund der sehr kleinen Fallzahlen im Analysesample der Testwerte in einem kleinen Ausschnitt um das Signifikanzniveau, konnten jedoch aus diesen Ergebnissen keine belastbaren Schlussfolgerungen gezogen werden. In der Tendenz zeigte sich jedoch, dass eine hohe Anzahl an berichteten Koeffizienten und damit eine größere Wahrscheinlichkeit irgendeinen signifikanten Effekt berichten zu können, das Risiko eines Publication Bias dämpft.

Ebenso wie für die Vorläuferstudien (Gerber & Malhotra 2008a,b; Gerber et al. 2010; Ridley et al. 2007) als auch im Falle der nachfolgend publizierten Studien (Berning & Weiß 2015; Kühberger et al. 2014) wurden zwar eine große Anzahl an Artikeln gelesen und deren Effektstärken kodiert. Jedoch reduzierte sich die Fallzahl in diesen Studien von 1.000 berichteten Testwerten aus 50 bis über 100 Studien stark auf 50 bis 200 Testwerte in den Calipern. Zwar steigt per Definition bei der Verbreiterung der Caliper

[12] Beispiel: Signifikanzniveau 5% impliziert einen z-Wert von 1,96. Im Falle des 5%-Calipers würden alle Werte im Intervall 1,96±0,096 betrachtet, d. h. 0,05*1,96.

die Fallzahl, jedoch sinkt parallel die Fähigkeit des Tests, Publication Bias von anderen Faktoren der zugrunde liegenden Testwerteverteilung zu unterscheiden. Die geringe Anzahl der Werte in den engsten, jedoch präzisesten Calipern führt hierbei zu einer geringen statistischen Power, d. h. einer geringen Wahrscheinlichkeit einen existierenden Publication Bias auch wirklich zu entdecken. Unter dem Problem einer geringen statistischen Power leiden besonders multivariate Verfahren, welche es erlauben die Risikofaktoren zu identifizieren. Um das durch die Artikel und Testwerteanzahl bedingte Problem einer nur sehr geringen statistischen Power abzumildern, wurden zwei methodische Weiterentwicklungen des Caliper-Tests vorgenommen: zum einen die Berechnung der Caliper aufgrund der den Artikeln zugrunde liegenden *t*-Werten anstatt *z*-Werten, zum anderen die Imputation von Nachkommastellen, um mögliche Artefakte zu vermeiden. Die Verwendung der *t*-Werte, welche eine fallzahlabhängige Modellierung der Caliper für jeden einzelnen Testwert ermöglichen, zeigte dabei leicht konservativere, jedoch konsistente Ergebnisse. Die Imputation von Nachkommastellen bei den kodierten Teststatistiken, welche oftmals ab der zweiten Nachkommastelle gerundet werden, zeigte ebenfalls konsistente jedoch leicht stärkere Hinweise auf Publikation Bias.

Zwar erlaubt die vorliegende Studie ebenso wie ihre Vorgänger- und Nachfolgestudien die Diagnose von Publication Bias, es können jedoch aufgrund der geringen Fallzahl nur sehr vage Schlussfolgerungen über die Entwicklung des Publication Bias über die Zeit, sowie weitere Risikofaktoren gezogen werden. Zudem ist die Wahl des gewählten Calipers kaum empirisch zu rechtfertigen, so verwenden Gerber & Malhotra (2008a: 19) den 5%, 10%, 15% als auch 20% Caliper, wohingegen im vorliegenden Dissertationsbeitrag das 3%, 5%, 10% als auch 15% Intervall gewählt wurde. Ferner stellt die Vielzahl an berichteten Calipern ein Interpretationsproblem dar, da das gleiche Sample beginnend mit dem kleinsten Caliper (3% in der vorliegenden Studie) immer weiter erweitert wurde. Die berechneten Signifikanzwerte sind daher nicht mehr statistisch unabhängig (Holm 1979). Dieses Problem konnte im vorliegenden Dissertationsbeitrag durch korrigierte *p*-Werte adressiert werden, welche jedoch die Power der Tests ihrerseits stark reduzieren.

### *1.3.2. Examining publication bias – A simulation-based evaluation of statistical tests on publication bias*

Im zweiten Beitrag der Dissertation (Kapitel 3) werden die Probleme, welche sich im ersten Beitrag stellten, insbesondere die Frage nach der idealen Größe des verwendeten Calipers, wieder aufgenommen. Der Beitrag vergleicht dabei die Qualität verschiedener Testverfahren auf Publication Bias im Hinblick auf deren statistische Power und die Falsch-Positiv-Rate. Ein guter Test auf Publication Bias sollte dabei mit einer angebbaren, vorher durch das Signifikanzniveau spezifizierten Wahrscheinlichkeit Publication Bias feststellen, obwohl dieser eben nicht besteht (konsistente Falsch-Positiv-Rate). Zugleich sollte die statistische Power des Tests, d. h. die Wahrscheinlichkeit einen existierenden Publication Bias auch als einen solchen zu entdecken, so hoch wie möglich sein. Hierfür wurden unterschiedli-

che Tests auf Publication Bias, welche insbesondere für Meta-Analysen zu einer bestimmten Fragestellung entworfen wurden, namentlich Egger's Test (Egger et al. 1997; auch bekannt als Funnel Asymmetry Test – FAT, Stanley & Doucouliagos 2014), p-uniform (van Aert et al. 2016; van Assen et al. 2015) sowie der test for excess significance (TES; Ioannidis & Trikalinos 2007), mithilfe von Monte Carlo Simulationen evaluiert. Im Anschluss an den ersten Beitrag wurde zudem der Caliper-Test (CT) als zusätzliches Testverfahren mit unterschiedlichen Intervallbreiten (3, 5, 10, 15%) integriert. Insbesondere die Inklusion des TES als auch des CT stellen dabei ein Novum innerhalb der bisherigen Literatur dar.

Ein Vergleich von sich in der statistischen Konzeptionierung unterscheidenden Publication Bias Tests erscheint von besonderem Interesse, da sich bemerkenswerterweise trotz vergleichbarer Anwendung in Meta-Analysen eine anhand der Fächergrenzen orientierte Spezialisierung auf verschiedene Tests etabliert hat. Dies zeigt die nachfolgende bibliographische Analyse deutlich. Hierfür wurde die Literaturdatenbank *Web of Science* mittels der Stichwortsuche nach Publikationen durchsucht, welche sowohl den jeweiligen Test auf Publication Bias als auch die Phrase „publication bias" enthielten.[13] Zudem wurden die jeweils gennannten Forschungsgebiete, in welchen die jeweiligen Publikationen von der Literaturdatenbank *Web of Science* verortet wurden, kodiert. An den in Tabelle 1-4 dargestellten Ergebnissen lässt sich ein stark nach Forschungsrichtungen geteiltes Muster erkennen.

**Tabelle 1-4 Bibliographische Analyse der in Kapitel 3 untersuchten Tests auf Publication Bias**

| | Stichwortsuche [a] | | Schneeballsuche [b] | | |
|---|---|---|---|---|---|
| | Forschungsgebiet | Publikationen | Schlüsselpublikation (SP) | Gebiet Zitation | Zitationsanzahl SP |
| Egger's Test [c] | Medizin | 1.053 | Egger et al. (1997) | Medizin | 19.266 |
| FAT [d] | Wirtschaftswiss. | 4 | Stanley & Doucouliagos (2014) | Sozialwiss. | 113 |
| p-uniform | Psychologie | 9 | van Assen et al. (2015) | Sozialwiss./ Psychologie | 41 |
| Test for excess significance | Psychologie | 4 | Ioannidis & Trikalinos (2007) | Psychologie | 212 |
| Caliper Test | Soziologie/ Mathematik | 8 | Gerber & Malhotra (2008a) | Wirtschaftswiss. | 63 |

[a] Suche im *Web of Science* nach Testname sowie „publication bias"; [b] Suche der Schlüsselpapiere sowie deren Zitationen; [c] Äquivalent zum FAT; [d] Äquivalent zu Egger's Test

So wird Egger's Test vornehmlich in der Medizin eingesetzt, wie die hier erzielten 1.053 Suchergebnisse für den Suchbegriff „Egger's Test" zeigen. Dieser Trend folgt damit der Empfehlung der in der Medizin einflussreichsten Richtlinie zur Erstellung von Meta-Analysen, dem Cochrane Handbuch (Higgins & Green 2008). Der in der Grundrichtung äquivalente FAT ist eher in den Wirtschaftswissenschaften verbreitet. Ebenso wie der FAT (4 Ergebnisse, vgl. Tabelle 1-4) erzielen auch p-uniform (9) und der TES (4) nur eine deutlich geringe Anzahl an Suchergebnissen, wobei die letzten beiden Tests insbesondere

[13] Suche am 02.01.2019

in der Psychologie zu verorten sind. Der CT (8) hingegen wird insbesondere in Publikationen der Soziologie bzw. der Mathematik genannt.

Neben der Stichwortsuche wurden in der Schneeballsuche auch die Anzahl der Zitationen sowie das Forschungsgebiet der Publikationen untersucht, welche die Schlüsselpublikationen, in denen das Testverfahren erstmals vorgestellt wurde, zitieren. Die Befunde aus der Stichwortsuche werden hierbei bestätigt, so weist Egger's Test mit Abstand auch die höchsten Zitationszahlen auf (19.266), welche ebenso wie bei der Schlagwortsuche vornehmlich der Medizin entstammen. Auch bei den übrigen Tests bestätigt sich der Befund der bibliographischen Suche, so wird p-uniform (41) sowie der TES (212) vorwiegend in der Psychologie rezipiert. Der FAT (113) ebenso wie der CT (63) werden hingegen eher in den Sozialwissenschaften, welche neben der Psychologie auch die Soziologie und die Wirtschaftswissenschaften einschließt, zitiert. Die gezeigte disziplinspezifische Anwendung von Publication Bias Tests ist hierbei weniger einem inhaltlichen Kriterium, als vielmehr Pfadabhängigkeiten wie einer in einem bestimmten Fächerbereich verorteten Schlüsselpublikation geschuldet.

Der Dissertationsbeitrag (Kapitel 3) löst die Tests aus ihrem jeweiligen Fächerkontext und vergleicht ihre Leistungsfähigkeit einen Publication Bias zielgenau entdecken zu können. Hierfür wurden in der Monte Carlo Simulation die Anzahl der Testwerte, auf deren Basis alle eingesetzten Publikation Bias Tests beruhen, die Fallzahlen der Primärerhebungen als auch die zugrunde liegende wahre Effektstärke, inklusive einer effektheterogenen Bedingung modelliert. Als zentraler Faktor wurde die Neigung Publication Bias zu begehen, sofern ein nicht-signifikantes Ergebnis erzielt wurde, in den Primärstudien variiert. Hierbei wurden entweder 0% im Falle keines Publication Bias, 50% oder gar 100% der nicht-signifikanten Ergebnisse einem Publication Bias Treatment unterzogen. Dieses Treatment war ebenfalls zweigeteilt, einerseits wurde Publication Bias in der besonders in den experimentellen Forschungsbereichen verbreiteten kompletten Neuerhebung (vgl. file-drawer, Rosenthal 1979) oder aber in Form eines iterativen Hinzunehmens weiterer Variablen operationalisiert (*p-hackin*g). *p-hacking* wurde dabei durch die Aufnahme von mehreren Collider-Variablen, welche ein hohes Risiko bergen komplette statistische Artefakte hervorzubringen (Greenland et al. 1999; für eine aktuelle Problematisierung in der Soziologie Breen 2018), modelliert. Das Ergebnis der Monte Carlo Simulation lässt sich in Tabelle 1-5 zusammenfassen.

Zur Einordnung der Ergebnisse können drei trennende Dimensionen gebildet werden: der vermutete Typ des Publication Bias als entweder *file-drawer* oder *p-hacking*, die Effektstruktur der untersuchten Studien sowie die Richtung des Publication Bias. Neben dem bereits vorangehend erläuterten Typ des Publication Bias bezeichnet die Effektstruktur die Homogenität bzw. Heterogenität der untersuchten Studien. Unter Homogenität werden dabei Untersuchungen zu einer spezifischen Fragestellung gefasst, wohingegen Heterogenität verschiedene Fragestellungen und damit zugrundeliegende Effekte abbildet (z. B. Untersuchung von Publication Bias in einer gesamten wissenschaftlichen Disziplin). Als Richtung

eines Publication Bias kann die Präferenz für statistisch signifikante Ergebnisse unabhängig vom Vorzeichen des jeweiligen Effekts (zweiseitiger Publication Bias) oder für eine bestimmte statistisch signifikante Effektrichtung (einseitiger Publication Bias) definiert werden. Durch die statistischen Voraussetzungen sowohl von Egger's Test (im Folgenden aufgrund der Deckungsgleichheit durch FAT bezeichnet) als auch p-uniform kann hierbei nur einseitiger Publication Bias (Empfehlungen für zweiseitigen Publication Bias in Tabelle 1-5 in Klammern) untersucht werden.

**Tabelle 1-5 Empfohlene Tests auf Publication Bias für verschiedene Anwendungsszenarien**

| | | Publication Bias Typ | |
|---|---|---|---|
| | | *file-drawer* | *p-hacking* |
| Effektstruktur | Homogen | FAT (TES) | TES |
| | Heterogen | FAT (CT) | CT |

Sofern abweichend, geeignete Tests für zweiseitigen Publication Bias in Klammern.

Unter allen evaluierten Publication Bias Tests erwies sich der FAT im Hinblick auf seine Falsch-Positiv-Rate am konsistentesten. Die empirisch beobachtete Fehlerwahrscheinlichkeit sollte dabei, unabhängig von den weiteren variierten Dimensionen, der *a priori* gewählten Fehlerwahrscheinlichkeit von 5% möglichst nahekommen. Zwar ist eine kleinere Fehlerwahrscheinlichkeit als 5% grundsätzlich nicht problematisch, jedoch geht mit dieser eine kleinere statistische Power einher. Von den untersuchten Tests fielen lediglich der 10%- und 15%-CT durch überhöhte Falsch-Positiv-Raten auf und sollten daher nicht angewandt werden.

Auch bei der statistischen Power war der FAT im Fall von *file-drawer* Publication Bias sowohl bei homogenen als auch heterogenen Effekten klar überlegen. Da der FAT jedoch nur bei einseitigem Publication Bias anwendbar ist, bietet sich der TES als Alternative an, sollte die Vermutung auf zweiseitigen Publication Bias bestehen. Der CT hat unter heterogenen Effekten, insbesondere wenn eine größere Anzahl an Studien auf Publication Bias untersucht werden, ebenso wie bei zweiseitigem Publication Bias Vorteile. Im Fall von *p-hacking* ist hingegen der TES mit einer deutlich besseren statistischen Power ausgestattet und sowohl bei ein- als auch zweiseitigem Publication Bias anwendbar. Im Falle von *p-hacking* und Effektheterogenität bietet sich im ein- und zweiseitigen Fall der CT als Test mit der größten statistischen Power an. Hierbei ist zu berücksichtigen, dass der CT immer eine recht große Anzahl an Testwerten benötigt. Der 5%-CT bietet dabei einen guten trade-off zwischen konsistenter Falsch-Positiv-Rate und hinreichender statistischer Power. Im Anschluss an die offene Frage nach der richtigen Breite des Calipers aus dem ersten Beitrag der Dissertation (Kapitel 2) lässt sich als Ergebnis feststellen, dass insbesondere der 5%-CT zum Test auf Publication Bias gerade in einem effektheterogenen Feld wie einer gesamten Disziplin geeignet ist.

### *1.3.3. Are really most of our research findings false? An empirical estimation of trends in statistical power, publication bias and the false discovery rate in psychological journals (1975-2017)*

Der dritte Dissertationsbeitrag (Kapitel 4) schließt an die beiden vorangegangenen Beiträge an und erweitert sie in zwei Aspekten. Zum einen wurde durch einen automatisierten Export von Testwerten die Datenbasis um ein Vielfaches erweitert, was insbesondere der Anwendung des CT entgegenkommt, welcher wie im vorangegangenen Beitrag (Kapitel 3) erläutert, eine Vielzahl an eingeschlossenen Testwerten benötigt, um mit einer angemessenen statistischen Power ausgestattet zu sein. Zum anderen wurde der CT über seine Fähigkeit der bloßen Diagnose von Publication Bias hinaus im Hinblick auf die Abschätzung von dessen Folgen für die Wissenschaft erweitert. In einem ersten Schritt wird auf die Datengrundlage eingegangen. In einem zweiten Schritt werden dann die statistischen Grundlagen sowie Ergebnisse des Beitrags skizziert und eingeordnet.

Während im ersten Dissertationsbeitrag 1.618 Testwerte aus 108 Artikeln erfasst werden konnten – eine Anzahl, die gerade bei der Ursachenanalyse von Publication Bias schnell an die Grenzen einer akzeptablen statistischen Power stoßen lässt – konnten im vorliegenden Beitrag durch den automatisierten Export 736.596 Testwerte aus 42.234 Artikeln extrahiert werden.[14] Die automatisierte Routine exportierte einerseits die bibliographischen Daten als auch die jeweiligen Volltexte.[15] Die ausgezeichnete Möglichkeit zu solch einem automatisierten Export boten die in den Jahren von 1975-2017 erschienenen quantitativ empirischen Zeitschriftenartikel, welche in von der APA (American Psychological Association) herausgegebenen Zeitschriften veröffentlicht sind, da das Publication Manual der APA, beginnend ab der 2. Auflage (American Psychological Association 1974) stark standardisierte Berichtspflichten von statistischen Ergebnissen vorschreibt. Die so erhobenen Daten entsprechen der klassischen Big Data Definition der 3Vs, *Volume*, *Velocity* und *Variety* (Laney 2001). Dabei bezieht sich *Volume* auf die schiere Masse der Daten, welche im vorliegenden Fall zunächst in Form von 53.861 vollständigen Forschungsartikel vorlagen, welche im Folgenden mit Methoden des *text mining* anhand der durch die im Publication Manual der APA formalisierten Standardform des Berichtens statistischer Testwerte in einen Datensatz aller Testwerte transformiert werden konnten. Auch die Geschwindigkeit der Datenerfassung, *Velocity*, ist besonders im Vergleich mit der manuellen Kodierung der Testwerte im ersten Dissertationsbeitrag (Auspurg et al. 2014) augenscheinlich. Die Eigenschaft der *Variety* des Datenmaterials lässt sich am besten anhand der bibliographischen Daten, welche zumeist als semistrukturierte XML-Daten vorliegen, sowie den aus Datenorganisationssicht völlig unstrukturierten Textdaten nachvollziehen.

In Anlehnung an die beiden ersten Dissertationsbeiträge lässt sich die insbesondere für Big Data zentrale Frage nach dem „warum“ (Salganik 2018: 14), bzw. dem Mehrwert dieser Daten stellen. Der Vorteil

[14] Aus den Artikeln kodierbare Teststatistiken, welche zur Berechnung des CT benötigt werden.
[15] Die Exportroutine wurde in Python programmiert.

der automatisierten Vorgehensweise liegt nicht gemeinhin in der Natur der erhobenen Daten als Big Data, vielmehr konstituiert sich der Mehrwert aus den statistischen Möglichkeiten, welche versprechen das Erkenntnisinteresse, in diesem Fall die Veranschaulichung und Entwicklung von Publication Bias und dessen Folgen, valider zu ergründen. Im vorliegenden Fall besteht der Mehrwert insbesondere in der Herausforderung des verwendeten Testverfahrens, des CT. So zeigte sich in den ersten beiden Dissertationsbeiträgen (Kapitel 2 & 3), dass der CT erst bei einer großen Anzahl an Testwerten eine akzeptable statistische Power aufweist. Im Falle einer Untersuchung der Entwicklung des Publication Bias über die Zeit hinweg multipliziert sich dieses Problem, da im Fall der vorliegenden 43 Untersuchungsjahre (1975-2017), den Ergebnissen des zweiten Dissertationsbeitrags in Kapitel 3 folgend,[16] ca. 43.000 Testwerte erforderlich wären, um den Zeittrend ohne weitere Restriktionen schätzen zu können.[17] Die Verwendung automatisierter Verfahren zur Datengewinnung bieten daher einen ganz konkreten analytischen Mehrwert, im vorliegenden Fall eine ausreichende statistische Power zur Untersuchung von Zeittrends. Dieser Mehrwert, *Value*, wird in der Literatur auch als Übergang von Big Data zu Smart Data bezeichnet (Iafrate 2015; Schöch 2013).

Das fundamentale Problem herkömmlicher Daten, welches sowohl bei Big als auch Smart Data unvermindert fortbesteht, sind mögliche Verzerrungen in den Daten (vgl. das Konzept des Total Survey Errors in der Umfrageforschung Groves 2009: Kapitel 2), deren Abwesenheit essentiell für die Validität der Daten ist (*Validity*). In der Anwendung auf Big bzw. Smart Data lassen sich die möglichen Verzerrungen in Zeilenfehler, welche die Selektion der Untersuchungseinheiten betreffen, und Spaltenfehler, welche die Messung betreffen, aufgliedern (Biemer 2017: 268-270).[18] Da im vorliegenden Beitrag eine Vollerhebung aller empirischer Artikel, welche in den betreffenden Zeitschriften veröffentlicht wurden, durchgeführt wurde, sind Zeilenfehler wenig einschlägig. Doch Spaltenfehler, d. h. die systematische Fehlkodierung oder gar Nichtkodierung von Testwerten, können gleichwohl problematisch sein. Aus diesem Grund wurden 20 zufällig ausgewählte Artikel nochmals manuell kodiert. Durch den Vergleich der manuellen sowie automatisierten Kodierung konnte herausgefunden werden, dass einerseits der Datensatz kaum fälschlicherweise exportierte Testwerte enthält (2,66%), andererseits der Exportalgorithmus zu einem höchstmöglichen Anteil alle relevanten Testwerte auch erkennt (88,2%, unter Einbeziehung von nicht exportierbaren Tabellen 69,41%). Insgesamt kann konstatiert werden, dass der in der Programmiersprache Python erstellte Exportalgorithmus eine adäquate Datenqualität liefert.

Die zweite Neuerung des Beitrags ist die Möglichkeit, im Gegensatz zum ersten Dissertationsbeitrag in Kapitel 2, die Prävalenz von Publikation Bias, d. h. den Anteil von durch Publikation Bias signifikant gewordenen Ergebnissen, zu schätzen. Zudem konnte aufbauend auf den exportierten Testwerten auch die statistische Power der jeweiligen Tests berechnet werden. Durch die Kombination beider Maßzahlen

[16] Gegeben einer adäquaten statistischen Power bei 1.000 Testwerten.
[17] Eine mögliche Restriktion wäre die Beschränkung auf einen linearen Zeittrend mittels zweier Zeiträume (z. B. 1974-1993 bzw. 1994-2017).
[18] Biemer (2017: 271) nennt zudem Zellenfehler.

war es damit möglich, den prominenten jedoch ausschließlich theoretisch fundierten Artikel „Why Most Published Research Findings Are False“ von Ioannidis (2005) mit empirischem Leben zu füllen. Anhand der *False Discovery Rate* wurde die Frage geklärt, wie viele signifikante Ergebnisse bedingt durch eine zu niedrige statistische Power und Publikation Bias lediglich als statistische Artefakte zu bezeichnen sind.

Die statistische Power in den untersuchten Artikeln lag mit nur 41,7% im Durchschnitt weit unter den von Cohen (1988) geforderten 80%. Hoffnung gibt jedoch der leicht positive Trend insbesondere der letzten Jahre. Weniger optimistisch hingegen stimmt die nahezu komplett stagnierende Entwicklung des Publication Bias: 22,7% der zunächst nicht-signifikanten Ergebnisse werden bedingt durch Publication Bias zu signifikanten Ergebnissen. Beide defizitären Größen zusammengenommen, einerseits die zu niedrige statistische Power und andererseits der substantielle Publication Bias, führen in der Folge dazu, dass im Mittel 32,6% (*False Discovery Rate*) der signifikanten Ergebnisse, welche in den Jahren 1975-2017 in von der APA herausgegebenen psychologischen Zeitschriften veröffentlicht wurden, nicht auf einen substantiellen Effekt verweisen, sondern vielmehr statistische Artefakte darstellen.

### *1.3.4. Berufungsverfahren als Turniere: Berufungschancen von Wissenschaftlerinnen und Wissenschaftlern*

Wie bereits in Kapitel 1.1. dargelegt, verschiebt der vierte Beitrag der Dissertation (Kapitel 5) den Fokus möglicher Ineffizienzen im Wissenschaftssystem von Forschungsergebnissen auf die Auswahl des Forschungspersonals. Konkret werden in diesem vierten Disserationsbeitrag erstens die Entwicklung der Frauenanteile über den gesamten Verlauf eines Berufungsverfahrens und zweitens die nach Geschlecht unterschiedlichen Leistungsvoraussetzungen untersucht, die das schlussendliche Erreichen einer Professur bedingen.

Besonders in der Medienberichterstattung wird zum Monitoring der Chancengerechtigkeit oftmals der Frauenanteil an Professuren herangezogen. Dieser hat sich in den letzten zehn Jahren zwar von 9% auf über 23% erhöht (Gemeinsame Wissenschaftskonferenz (GWK) 2018: 8), jedoch ist die Geschlechterparität mit Ausnahme des Bundeslandes Berlin, welches im Jahr 2017 45% Frauen auf Professuren berief (Grigat 2018), noch nicht erreicht. Somit ist der Verlust an Forscherinnen auf dem Weg zur Professur eklatant, da der Frauenanteil sowohl bei den Studienabschlüssen (51%) als auch Promotionen (45%) deutlich stärker an eine Gleichverteilung angenähert ist (Gemeinsame Wissenschaftskonferenz (GWK) 2018: 8). Die Unterrepräsentanz von Frauen insbesondere in Führungsebenen ist dabei selbstverständlich nicht auf die Wissenschaft beschränkt, sondern zeigt sich auch in anderen Bereichen wie der Privatwirtschaft (McCook 2013).

Zum Monitoring der Chancengerechtigkeit ist der Anteil der von Frauen besetzten Professuren jedoch aus zwei Gründen nur bedingt geeignet: So sollte einerseits immer nur auf den aktuellen Stand der Berufungen des jeweiligen Jahres, also auf Neuberufungen, bezuggenommen werden, da sonst kompensatorische Ungleichbehandlungen auftreten können, welche zwar eine Gleichverteilung im Bestand der

Professuren herstellen, jedoch gerade zu Chancenungleichheit im jeweiligen Verfahren zulasten der überrepräsentierten Gruppe führen können. Zweitens, sollte sich bei allen Verfahren auf den jeweiligen Stand der Bewerbungen bezogen werden, da ansonsten Prozesse der Selbstselektion als Diskriminierung eingestuft werden könnten. Neben der Gesamtschau der Frauenanteile an den Professuren scheint daher insbesondere die Entwicklung des Frauenanteils über den Verfahrensverlauf untersuchenswert, um so gerade jene Verfahrensstufen ausfindig machen zu können, an denen Bewerberinnen verloren gehen (vgl. den Begriff der *leaky-pipeline* Alper 1993).

Zur Untersuchung der ersten Fragestellung, der Entwicklung des Frauenanteils über den Verfahrensverlauf, wurden 235 Berufungsverfahren einer mittelgroßen deutschen Universität von 2001-2013 über den Verfahrensverlauf, angefangen vom Frauenanteil bei den Bewerbungen, der Erstauswahl, den Vorstellungsvorträgen, der Liste und schlussendlich des ersten Listenplatzes nachverfolgt. Ebenso wie in Vorläuferuntersuchungen in Österreich (Wroblewski & Leitner 2013) sowie den Niederlanden (Brouns 2000) findet sich über den Verfahrensverlauf keine Veränderung des Frauenanteils. Damit zeigen sich keine Hinweise auf strukturelle Ungleichbehandlungen innerhalb desselben. Dies ist insofern interessant, da sich die Wettbewerbsbedingungen in der Wissenschaft insbesondere zwischen den Niederlanden und Deutschland bzw. Österreich stark unterscheiden. So weist das niederländische System einen höheren Anteil an unbefristeten bzw. *tenured* Stellen aus, wohingegen in Deutschland und Österreich fast ausschließlich Professuren als unbefristete Dauerstellen angelegt sind, welche zudem einen deutlich geringeren Anteil an der Anzahl der Gesamtbeschäftigten haben (Kreckel 2008: 17-18). Darüber hinaus zeigt der vorliegende Dissertationsbeitrag jedoch, dass insbesondere im Übergang zwischen den Qualifikationsphasen der Promotion und der Habilitation, aufgeteilt nach Fächern, ein mit 15 bzw. knapp 2 Prozentpunkten geringerer Frauenanteil als auf Stufe der Bewerbung vorliegt. Die *leaky-pipeline* scheint daher insbesondere am Übergang von der Qualifikationsphase in das Wissenschaftssystem zu bestehen (diesen Schluss legen auch Hechtman et al. 2018 für die Bewerbung um die Verlängerung von Research Grants nahe).

Die zweite Fragestellung hingegen untersucht unterschiedliche Leistungen sowie Leistungsvoraussetzungen im Zugang auf Professuren. Hierfür wird der Fokus von der Betrachtung von Aggregatdaten über den Verfahrensverlauf hin auf die Individualdaten von BewerberInnen, welche auf die Berufungsliste gesetzt wurden, gelegt. Auf diese Weise ist es möglich die in der Literatur berichtete Ungleichbehandlung von Frauen zu untersuchen, welche sich in einer Benachteiligung in Form von einem Mehr an benötigten Publikationen (De Paola & Scoppa 2015; De Paola et al. 2018) wie auch einer Bevorzugung in Form von einer geringeren Anzahl an Publikationen (Jungbauer-Gans & Gross 2013; Lutter & Schröder 2016), welche für die Erlangung einer Professur von Nöten sind, äußern kann. Zentrales Argument des Beitrags ist dabei die Turnier-Charakteristik von Berufungsverfahren, wie sie in der Arbeitsmarktforschung als *labor queue* (Fernandez & Mors 2008) diskutiert wird. BewerberInnen konkurrieren dabei nicht mit allen anderen BewerberInnen, sondern nur mit solchen, welche sich auf die gleiche Vakanz, d. h. Professur bewerben. Aufgrund der unterschiedlichen Publikationskultur der Fächer (insb.

Mehrfachautorschaften, *high-impact journals*, Wichtigkeit von Monographien, etc.) kann sich die Publikationsleistung der jeweiligen BewerberInnen, gemessene an der Anzahl an *peer-reviewed* Publikationen, zwischen den Verfahren deutlich unterscheiden. Dies wird in den verwendeten Daten deutlich, so publizieren Frauen ohne Berücksichtigung der *labor queue* Struktur 11 *peer-reviewed* Publikationen weniger, mit ihrer Berücksichtigung jedoch nur noch knapp 4. Zusammenfassend ist zu konstatieren, dass sich insgesamt keine Ungleichbehandlung und damit Diskriminierung von auf der Berufungsliste stehenden BewerberInnen im Zugang auf eine Professur zeigt.

## 1.4. Synthese

Die vorliegende Dissertation untersuchte Ineffizienzen im Wissenschaftssystem am Beispiel von Fehlverhalten und Diskriminierung als Ergebnis dysfunktionaler Reputationssysteme. Zusammenfassend lässt sich feststellen, dass die ersten drei Dissertationsbeiträge (Kapitel 2-4) zum Publication Bias Evidenz für Ineffizienzen finden konnten, während es dem letzten Beitrag (Kapitel 5) zufolge keine ungleichen Chancen für Frauen gibt, auf Professuren berufen zu werden. Im Folgenden sollen die Erkenntnisleistungen der vier Dissertationsbeiträge anhand einer theoretischen Rückbindung eingeordnet sowie mögliche Interventionen zur Verringerung der beobachteten Ineffizienzen diskutiert werden. Abschließend soll zudem auf Desiderata für die weitere Forschung hingewiesen werden.

### *1.4.1. Theoretische Rückbindung*

Die theoretische Haupterkenntnis der vorliegenden Dissertationsteile zum Publication Bias liegt in der Erklärung des Auftretens von Publikation Bias durch die zugrunde liegende spieltheoretische Situation eines Gefangenendilemmas. Obwohl alle WissenschaftlerInnen von einer unverzerrten und von Praktiken des Publikation Bias freien Wissenschaft profitieren, lohnt sich für einzelne Forschenden ein abweichendes Verhalten. Die daraus abgeleitete Implikation lautete, dass ein Publication Bias in der bestehenden Literatur feststellbar sein müsste. Dies konnte in den beiden Beiträgen zum Publication Bias in der deutschen Soziologie in Kapitel 2 sowie in der internationalen Psychologie in Kapitel 4 bestätigt werden. Der durch den automatisierten Export von Testwerten in Kapitel 4 untersuchbare Zeittrend des Publication Bias widerspricht der oft vertretenen These, dass der im Zeitverlauf zunehmende Publikationsdruck zu einem steigenden Publication Bias führt. Vielmehr zeigte sich in diesem dritten Beitrag eine bemerkenswerte Konstanz der Prävalenz von Publication Bias.

Im Gegensatz zum Publication Bias liegt bei der Diskriminierung von Bewerberinnen auf Professuren keine Struktur eines Gefangenendilemmas vor. Ganz im Gegenteil, in diesem Setting führt eine Ungleichbehandlung bei entsprechend nicht-diskriminierender Konkurrenz zu Wettbewerbsnachteilen. Die Ergebnisse in Kapitel 5 bestätigen diese theoretische Prämisse, da keine Evidenz für Diskriminierung von Bewerberinnen im Zugang auf Professuren gefunden wurde. Ebenso zeigten sich unter Kontrolle der Publikationsleistung weder eine Benachteiligung noch eine Bevorzugung von Bewerberinnen. Dieses in Kapitel 5 erzielte Nullergebnis könnte kurioserweise somit selbst ein Hinweis zum Publication Bias in der Literatur zu Geschlechterungleichheiten sein.

### *1.4.2. Mögliche Interventionen*

Auf Basis der Ergebnisse der Dissertation lassen sich sowohl für die Ineffizienzen bei Publication Bias, als auch für die nichtexistierenden Ineffizienzen im Bereich der Auswahl von Bewerberinnen auf Professuren Interventionen diskutieren.

Zur Reduktion von Publication Bias lassen sich die bereits in Kapitel 1.2.1 exemplarisch diskutierten Interventionen anführen. Insbesondere im Anschluss an die die Kooperation begünstigenden Verhaltensweisen im wiederholten Gefangenendilemma (Axelrod & Hamilton 1981) ist hier auf Interventionen zu verweisen, welche das tatsächliche Commitment von Forschenden zu einer transparenten Wissenschaft ohne Publication Bias auch für andere Akteure nachvollziehbar halten. Akteure sind dadurch in der Lage im Anschluss an Becker (1968) nicht-kooperierendes Verhalten, welches zeitlich zurückliegt, zu erkennen und schlussendlich zu sanktionieren. Vorangegangenes kooperatives Verhalten wird so in das Reputationssystem positiv eingespeist.

Um die Transparenz des gesamten Forschungsprozesses zu erhöhen, bieten sich Pre-Registrierungen von Studien an. Die geplanten Auswertungen werden so vor der eigentlichen Datenerhebung genau festgelegt, Abweichungen sind im Nachhinein freilich möglich, werden jedoch als solche schnell ersichtlich. Schwerer umsetzbar sind Pre-Registrierungen bei Analysen bereits bestehender Daten. Sollten die erstellten und analysierten Daten im Sinne eines *public* oder *scientific use* zur Verfügung gestellt sein, wäre es überlegenswert, von Forschenden zur Erlangung des Datensatzes einen Pre-Analyseplan einzufordern. Im Zuge dessen ist sicherlich auch eine stärkere Trennung von erklärenden inferenzstatistischen Verfahren sowie explorativen Verfahren erforderlich. Bietet sich z. B. eine lineare Regression (OLS) hervorragend zum Testen eines vorher spezifizierten Modells an, ist ein händisches exploratives Vorgehen, sei es mittels stufenweiser Regression oder ähnlicher Verfahren, in diesem Fall ineffizient. Neuere Verfahren wie *lasso* (Tibshirani 1996) oder *elastic net* (Zou & Hastie 2005), um nur einige Verfahren aus dem Bereich des Maschine-Learning zu nennen, scheinen für explorative Analysen deutlich vielversprechender.

Neben Pre-Registrierungen ist in einem weiteren Schritt die Verfügbarkeit von Forschungsdaten zentral. Das bereits in Kapitel 1.2.1 angeführte Konzept von FAIR Data (Wilkinson et al. 2016) bietet hierbei eine Grundlage, so müssen Forschungsdaten nach diesem Konzept einerseits auffindbar (*findability*) sowie über einen längeren Zeitraum zugänglich sein (*accessibility*). FAIR Data schließt dabei nicht nur den Datensatz, sondern auch die für die Nachvollziehbarkeit nötigen Auswertungsskripte mit ein. Dies bedeutet, dass Daten vorzugsweise bei einer Zentralinstitution[19] bzw. einem Datenhoster[20] archiviert werden sollten. Die oft geübte Praxis Daten auf der eigenen Homepage zu verlinken, steht diesem Ansinnen diametral entgegen. Denn die Daten sind auf diese Weise über längere Zeiträume oft nicht gesichert zu erreichen, da sich z. B. kleinere Verweisstrukturen auf der Homepage ändern oder diese

[19] z. B. GESIS Datorium, Open Science Framework
[20] z. B. Figshare, GitHub

schlichtweg nicht weitergeführt wird. Neben der grundlegenden Zugänglichkeit der Daten ist zudem auch deren praktische Verwendbarkeit wichtig. Das bedeutet, dass die verwendeten Forschungsdaten in einem zumindest potentiell zugänglichen Format (*interoperability*) gespeichert werden sollten. Wurden Open-Source Auswertungsprogrammen wie R bzw. Python verwendet, ist diese Voraussetzung problemlos gewährleistet. Falls lizensierte und daher kostenintensivere Programme (z. B. Stata, SAS, SPSS) eingesetzt wurden, ist die Interoperabilität zumeist durch Zusatzpakete der Open Source Programme gewährleistet, dies sollte jedoch im Einzelfall vor der Veröffentlichung von den AutorInnen nachgeprüft werden. Die Wiederverwendbarkeit der Daten (*reusability*) fokussiert gänzlich auf die Strukturiertheit der Daten, welche einer außenstehenden Person deren Nutzung erst ermöglicht. Es sollte daher auf eine klare Benennung der jeweiligen Variablen und Ausprägungen geachtet sowie im Falle komplexerer Daten auch eine angemessene Dokumentation zur Verfügung gestellt werden. Um diese Standards von FAIR Data flächendeckend zu etablieren, können diese insbesondere von Zeitschriften oder auch Drittmittelgebern eingefordert werden, so schreibt die Europäische Kommission im Rahmen aller Projekte des einflussreichen Horizon 2020 Forschungsförderprogramms FAIR Data explizit vor (Europäische Kommission 2016).

Während sich die bisher genannten Interventionen der Pre-Registrierung und Datenarchivierung auf im Forschungsprozess verankerte Abläufe beziehen, betrifft die dritte, auf die Reduktion von Publikation Bias bezogene Intervention den Bericht von Teststatistiken in den Publikationen selbst. Nur durch konsistente Vorgaben, wie Effekte in den Publikationen berichtet werden sollen, sind großangelegte Analysen zum Publication Bias wie in Kapitel 4 möglich. Ein besonders positiv hervorzuhebendes Beispiel sind die im Publication Manual der APA (für die aktuelle Ausgabe vgl. American Psychological Association 2010) festgeschriebenen Berichtspflichten. Neben einer automatisierten Kontrolle auf Publication Bias ermöglichen es formal standardisierte Berichtspflichten zudem, kleinere Fehler und Ungenauigkeiten in den Publikationen zu identifizieren (für Rundungsfehler bzw. Betrug vgl. Nuijten et al. 2016). Auch für die Leser der Publikationen besteht anhand der berichteten Teststatistiken die Möglichkeit, die in der Publikation gezogenen Schlüsse anhand der Testwerte zumindest einer ersten Plausibilitätsprüfung zu unterziehen. Zwar sind Pre-Registrierungen sowie Datenpolicies bei weitem besser geeignet, Publikationen kritisch zu hinterfragen, ein einheitliches Berichtsformat in den Publikationen hat jedoch den Vorteil besonders zeiteffizient für den interessierten Leser zu sein, da eine erste Überprüfung schon parallel zur Lektüre stattfinden kann.

Da im Falle der Diskriminierung im Zugang auf Professuren eine annähernd gleiche Zugangschance für Bewerberinnen und Bewerber festgestellt wurde, sind keine spezifischen Interventionen einschlägig. Besonders hervorzuheben ist, dass große Förderprogramme wie das seit 2008 vom Bundesministerium für Bildung und Forschung ins Leben gerufene Professorinnenprogramm neben positiven Schritten in Richtung des Förderziels, wie die Erhöhung des Frauenanteils bei Professuren (GESIS Leibniz-Institut für Sozialwissenschaften 2017: 3) nicht in einer Schlechterstellung von Bewerbern resultiert ist.

### *1.4.3. Forschungsdesiderata*

Die Desiderata an die zukünftige Forschung lassen sich in zwei Stränge aufteilen: Zum einen sollte das anhand der strukturierten Erhebung vorliegende deskriptive Potenzial der Daten weiter genutzt, zum anderen sollte neben der Deskription auch die Ursachenanalyse weiter vertieft werden.

Monitoring-Instrumente erlauben es, aufbauend auf den in der vorliegenden Dissertation geschaffenen strukturellen Datenerhebungstools, Indikatoren guter wie schlechter wissenschaftlicher Praxis wie in Kapitel 4 gezeigt weiter zu untersuchen. Im Falle des Publication Bias ist durch die automatisierte Datenerhebung eine Fortschreibung der Datenbasis mit moderatem Aufwand möglich. Dies erlaubt es einerseits auch die zukünftige Entwicklung von Publication Bias, der statistischen Power, sowie in deren Folge, der *False Discovery Rate* über die Zeit zu verfolgen, andererseits können auf diese Weise einzelne Zeitschriften die in ihren Artikeln bestehenden Trends über die Zeit nachvollziehen sowie diese mit anderen Zeitschriften verglichen werden. Ähnlich eines *Impact Factors* bei Zeitschriften könnte anhand dieser Indikatoren ein Qualitätsmaß für die in Zeitschriften veröffentlichten Studien etabliert werden.

Im Fall des Monitorings von Ungleichheiten in Berufungsverfahren wäre aus Gleichstellungsgesichtspunkten eine bundesweite Erhebung der über die Verfahrensstufen fortgeschriebenen Frauenanteile wie in Österreich bereits 2013 erhoben (Wroblewski & Leitner 2013) äußerst vielversprechend, da dies die Datenbasis des nur auf einer Universität beruhenden Beitrags in Kapitel 5 stark erweitern würde. Das Fehlen eines solchen Monitoring-Instruments erscheint gerade vor dem Hintergrund des seit 2008 bestehenden Professorinnenprogramms frappant, da so die Auswirkungen des Programms nur unzureichend evaluiert werden können.

Ebenfalls sollten die in Kapitel 2 begonnenen Analysen zu den Risikofaktoren des Publication Bias fortgesetzt werden. Für die vorliegenden Daten aller von der APA herausgegebener Zeitschriften wäre die Möglichkeit gegeben, die in Kapitel 1.2.1 sowie im vorangehenden Abschnitt erläuterten Interventionen experimentell zu testen. Einzelne Zeitschriften könnten so Interventionen wie eine verbindliche Hinterlegung von Daten implementieren und dann im Anschluss deren Auswirkungen auf die eingesetzten Indikatoren guter wie schlechter wissenschaftlicher Praxis (Publication Bias, statistischer Power und *False Discovery Rate*) evaluieren. Darüber hinaus wäre es ebenfalls möglich, die Auswirkungen einer solchen Intervention auf die Zitationszahlen der einzelnen Artikel, welche eine höhere Glaubwürdigkeit aufweisen müssten und daher öfter zitiert werden sollten, zu untersuchen.

Ziel der weiteren Forschung sollte es sein, aufbauend auf dem in der vorliegenden Dissertation entwickelten deskriptiven Instrumentarium die Ursachen der Ineffizienz im wissenschaftlichen Erkenntnisfortschritt zu untersuchen. Statt in einer wiederholten Problemdiagnose gefangen zu bleiben, müssen vielmehr diese Ursachen der Ineffizienz adressiert werden, um einen nachhaltigen wissenschaftlichen Erkenntnisfortschritt zu sichern.

# 2. Ausmaß und Risikofaktoren des Publication Bias in der deutschen Soziologie

Auspurg, K., T. Hinz und A. Schneck. Erschienen in: Kölner Zeitschrift für Soziologie und Sozialpsychologie 2014 (66):549-573. https://doi.org/10.1007/s11577-014-0284-3

## 2.1. Einleitung

Die Rezeption von Forschung orientiert sich in der wissenschaftlichen *community* und der Öffentlichkeit stark an der statistischen Signifikanz von Ergebnissen. Ob ein Effekt statistisch „signifikant" ist, hängt letztlich von lediglich per Konvention gesetzten statistischen Schwellenwerten ab. In der Wissenschaft hat sich das 5%-Signifikanzniveau etabliert (Cohen 1994; Fisher 1973).[21] Das Problem des Publication Bias (PB) schließt an diese willkürlichen Signifikanzschwellen an. Nach der Definition von Dickersin ist der PB „a tendency toward preparation, submission and publication of research findings based on the nature and direction of the research results" (Dickersin 2005: 13). Ergebnisse, welche die von den ForscherInnen gewählte Signifikanzschwelle unterschreiten, demnach „statistisch signifikant" sind, werden unabhängig von ihrer Qualität und Aussagekraft mit größerer Wahrscheinlichkeit niedergeschrieben, eingereicht, veröffentlicht und schlussendlich auch zitiert als nicht-signifikante Ergebnisse (Egger & Smith 1998; Mahoney 1977). In der Folge stellt die veröffentlichte Forschung nur eine Teilmenge der tatsächlich durchgeführten Forschung dar, und was noch problematischer ist: aufgrund der wegfallenden, nicht-signifikanten Ergebnisse handelt es sich um einen *verzerrten* Ausschnitt der gesamten durchgeführten Forschung. Ein PB kann durch das systematische Ausblenden von nicht-signifikanten Effekten dazu führen, dass die Wirkung von Maßnahmen deutlich überschätzt wird und im Extremfall völlig wirkungslose (sozialpolitische) Empfehlungen ausgesprochen werden.[22] Generell steht das Phänomen des PB einer guten wissenschaftlichen Praxis entgegen, wonach alle Forschungsergebnisse unabhängig von ihrem Resultat der Öffentlichkeit zugänglich gemacht werden sollten.

Im vorliegenden Beitrag wird das Vorliegen eines PB in der deutschen quantitativ-empirisch arbeitenden Soziologie anhand der Jahrgänge 2000-2010 der Kölner Zeitschrift für Soziologie und Sozialpsychologie (KZfSS) sowie der Zeitschrift für Soziologie (ZfS) untersucht.

Im Zentrum der Analyse stehen Anreize und Kosten des PB für Autoren sowie Sanktions- oder Überwachungsmöglichkeiten durch andere Akteure, wie etwa die Herausgeber der Zeitschriften. Inwieweit hängt das Risiko eines PB mit fehlenden Überwachungsmöglichkeiten (etwa aufgrund eines exklusiven Datenzugangs) zusammen? Wie sieht es mit den Möglichkeiten gegenseitiger Kontrolle in Auto-

[21] In den Sozialwissenschaften werden neben dem 5%-Niveau das konservativere 1%-Niveau und das bei kleinen Fallzahlen verbreitete 10%-Signifikanzniveau verwendet (Labovitz 1968; Skipper et al. 1967).

[22] Besonders eindrücklich lassen sich die Folgen in der Medizin veranschaulichen, in der die Unterdrückung nicht-signifikanter Studienergebnisse dazu führen kann, dass völlig wirkungslose Medikamente eingesetzt werden. Ein aktuelles Beispiel ist das Grippe-Mittel Tamiflu. Jefferson et al. (2012: 5) konnten in der Forschung zur Wirksamkeit einen klaren PB feststellen. Folgen sind u. a. sehr hohe Anschaffungskosten eines in der Wirkung überschätzten Präparats (über 70 Mio. Euro allein auf Bundesebene; Deutscher Bundestag 2013).

renteams aus? Verstärkte Ursachenforschung scheint vor dem Hintergrund der zwar präsenten Problematisierung des PB (z. B. Nuzzo 2014) und der stetigen Entwicklung neuer Diagnosemethoden (z. B. Leggett et al. 2013; Simonsohn et al. 2014b), zugleich aber einem fehlenden Wissen zu den Risikofaktoren relevant (Auspurg & Hinz 2011a). Für die Diagnose eines PB wird im vorliegenden Beitrag das Verfahren des Caliper-Tests (Gerber & Malhotra 2008a,b) eingesetzt.

## 2.2. Signifikanztests und PB

Grundprinzip der Inferenzstatistik ist es, von Merkmalsverteilungen eines zufälligen Ausschnitts (Stichprobe) aus der Grundgesamtheit (GG) auf Merkmalsverteilungen der GG zu schließen (Greene 2012: 434). In der Stichprobe beobachtbare Effekte, wie etwa ein Mittelwertunterschied eines Merkmals nach Geschlecht, können hierbei zufällig (allein durch die Stichprobenziehung) entstanden sein oder auf „wahre" Effekte (im vorliegenden Beispiel einen Geschlechtsunterschied) in der GG hinweisen. Das Signifikanzniveau, der Fehler 1. Art, gibt hierbei die Wahrscheinlichkeit an, dass ein Effekt irrtümlich aufgrund der Stichprobenbeobachtungen angenommen wird, obwohl dieser in der GG nicht vorliegt (Greene 2012: 1062). Es wird also fälschlicherweise die Nullhypothese, dass kein Effekt vorliegt, verworfen. Grundsätzlich gilt, dass sich Merkmalsverteilungen von Zufallsstichproben mit steigender Fallzahl den Merkmalsverteilungen der GG annähern (die Standardfehler der Schätzer werden kleiner – es sind also geringere zufällige Abweichungen vom Schätzwert zu erwarten). Mit steigender Fallzahl wird es demzufolge immer unwahrscheinlicher, in der Stichprobe starke Effekte festzustellen, obwohl in der GG keine Effekte bestehen. Allerdings werden selbst geringe Effekte bei entsprechend großer Stichprobe leichter als „überzufällig" und somit als „signifikant" betrachtet.. In jedem Fall sagt die statistische Signifikanz nichts über die Effektstärke und damit die praktische Relevanz von Ergebnissen aus.

Die Orientierung an strikten Signifikanzniveaus wurde in den Sozialwissenschaften schon früh kritisiert (Labovitz 1968; Skipper et al. 1967), ist jedoch immer noch weit verbreitet, wie sich etwa an der Verwendung von Sternchen zur Veranschaulichung des Signifikanzniveaus zeigt. Auch wenn sich *p*-Werte von knapp unter und knapp über 5% nur marginal in der Irrtumswahrscheinlichkeit, aufgrund der realisierten Stichprobe eine Nullhypothese zurückzuweisen, obwohl sie für die GG zutrifft, unterscheiden, führt eine strikte Orientierung an Signifikanzniveaus (signifikant oder nicht) zur diametral unterschiedlichen Interpretation der Ergebnisse.

Die Willkürlichkeit der Signifikanzniveaus erscheint vor allem vor dem Hintergrund des PB problematisch: Stellt die veröffentlichte Forschung nur eine anhand der Signifikanzniveaus ausgewählte Teilmenge der tatsächlich durchgeführten Forschung dar, ist die Annahme der Gesamtschau aller Forschungsergebnisse verletzt (Sutton & Pigott 2006: 227). Ein 5%-Signifikanzniveau bedeutet *per definitionem*, dass im Mittel jeder 20. Signifikanztest ein signifikantes Ergebnis zeigt, auch wenn in der GG kein Effekt besteht. Werden die signifikanten Ergebnisse nun bevorzugt beachtet und veröffentlicht,

werden Nulleffekte eher verkannt (Rosenthal 1979: 638) und statistische Artefakte für wahre Effekte gehalten.

Ein PB wird dabei vermutlich vornehmlich durch Selektion oder gar Manipulation von den Autoren selbst hervorgerufen (Dickersin 2005: 13). Hierbei lassen sich grundsätzlich zwei Vorgehensweisen unterscheiden. Erstens kann versucht werden, durch die wiederholte Erhebung von Daten (vgl. objective publication bias; Begg 1994: 400) statistisch signifikante Ergebnisse zu finden. Erhobene (Teil-) Projekte mit nicht-signifikanten Ergebnissen werden unterschlagen (vgl. mit dem Konzept des file drawer effects von Rosenthal 1979). Dieses Vorgehen erscheint insbesondere in Disziplinen mit kleinen Fallzahlen verbreitet (z. B. in der Psychologie). Die zweite Strategie ist die Re-Analyse des Datensatzes durch den Ausschluss von „Ausreißern", einem methodisch unbegründeten Wechsel der Auswertungsmethode oder dem theoretisch unbegründeten Austausch von Kontrollvariablen, bis die erwünschte Signifikanz erreicht wird (vgl. subjective publication bias; Begg 1994: 400). Im Gegensatz zum objektiven PB werden hier nicht die Daten selbst, sondern vielmehr die Ergebnisse der Auswertungen unterschlagen.[23]

Die Untersuchung von Forschungsergebnissen auf einen PB hat in den letzten Jahren in Meta-Analysen weite Verbreitung gefunden (s. z. B. Ferguson & Brannick 2012).[24] Als Voraussetzung für den Test auf einen PB müssen die in einer Meta-Analyse zusammengefassten Studien in einem möglichst ähnlichen Setting durchgeführt worden sein und sich auf ein- und denselben Effekt beziehen (um nicht den Verdacht eines „statistischen Fruchtsalats" zu rechtfertigen, s. Brüderl 2004). Unterscheidet sich die Fragestellung in den zu untersuchenden Studien stark oder soll gar eine gesamte Forschungsdisziplin wie im vorliegenden Fall die Soziologie untersucht werden, ist von sehr heterogenen Effekten auszugehen, für die eine statistische Zusammenfassung in Form von durchschnittlichen Effekten in Meta-Analysen keinen Sinn macht (für die wenigen bestehenden Meta-Analysen in der Soziologie siehe etwa Weiß & Wagner 2008). In der Soziologie wird noch wenig kumulative Forschung betrieben, sodass nur in seltenen Ausnahmefällen die kritische Masse von mindestens zehn Studien erreicht wird, die für PB-Testverfahren in Meta-Analysen erforderlich ist (Sterne et al. 2000: 1127).

## 2.3. Theoretischer Rahmen

Erste Erkenntnisse zur Auftrittswahrscheinlichkeit von einem PB lassen sich bereits Mertons Wissenssoziologie entnehmen, nach der zwei Grundprinzipien für den wissenschaftlichen Erfolg von Forschenden zentral sind: die Erstentdeckung (*priority*) sowie deren Innovationsgehalt (*originality*) (Merton 1957). Problematisch im Hinblick auf einen PB ist insbesondere das zweite Grundprinzip. Originalität

[23] Als verwandte Strategien lassen sich zudem das nachträgliche Zuschneiden der Hypothese auf die Ergebnisse (sogenanntes HARKING: Hypotheses after the results are known; Kerr 1998) sowie die nachträgliche Anpassung von Signifikanzniveaus anführen.

[24] Meta-Analysen fassen mehrere Untersuchungen zu einem Effekt zusammen und bieten dank ihrer insgesamt höheren Fallzahlen der aggregierten Einzelstudien einen genaueren Aufschluss über den „wahren" Effekt.

entspricht der auf Unterschiede abstellenden Alternativhypothese, während empirisch aber oft die weniger spektakuläre Nullhypothese nicht zurückgewiesen werden kann. Zudem werden Replikationen von Forschungsergebnissen oder gleichzeitige Entdeckungen dann als „unnecessary duplication[s]" (Merton 1961: 479) angesehen.

Systematischere Annahmen zum Auftreten eines PB lassen sich durch die Analyse der Anreiz- und Kostenstrukturen der beteiligten Akteure gewinnen. Nach der Rational-Choice Theorie wählt jeder Akteur stets jene Handlungsalternative, die unter Randbedingungen seinen Nutzen maximiert. Die Wissenschaft lässt sich dabei als eine Art Turnier von Forschenden betrachten, deren Ziel es ist, die Zitationen ihrer Werke und andere Auszeichnungen zu maximieren (Feigenbaum & Levy 1993: 216). Dies geschieht unter sehr selektiven Wettbewerbsbedingungen, unter denen Belohnungen und – speziell im deutschen Wissenschaftssystem – auch Positionen im Sinne eines „winner-take-all contest" (Stephan 2010: 222) vergeben werden. Aufsehenerregende Publikationen, Aufsätze in (hoch renommierten) Zeitschriften und häufige Zitationen sind dabei wichtige Erfolgskriterien, um sich gegenüber der Konkurrenz durchzusetzen (Feigenbaum & Levy 1996: 263; Stephan 2010: 223).

Um Zitationen zu erreichen, ist es für Forschende in einem ersten Schritt zunächst einmal wichtig, den eigenen Artikel überhaupt veröffentlichen zu können. Dieser Schritt stellt zumindest in den Zeitschriften mit geringen Annahmequoten eine große Hürde dar, so wurden etwa im American Sociological Review in den Jahren 2005-2010 nur zwischen 6% und 10% der eingereichten Manuskripte publiziert.[25] Ein PB lässt sich in diesem Wettbewerb als Strategie von Autoren verstehen, durch die Signifikanz der Ergebnisse die Bedeutung der eigenen Studie herauszustellen und so die Wahrscheinlichkeit einer Publikationszusage, und im Anschluss daran die (bei den hochgerankten Zeitschriften wiederum höhere) Zitationswahrscheinlichkeit zu maximieren.

Der Maximierung von Zitationen und Reputation stehen jedoch auch Kosten gegenüber. Kennen die Autoren die Norm, wonach Nullergebnisse nicht unterschlagen werden dürfen, dann bereitet der PB zumindest moralische Kosten (Slote 1985: 165). Ein Großteil der Forschenden scheint sich durchaus bewusst zu sein, dass ein Hintrimmen signifikanter Ergebnisse ethisch verwerflich ist (Necker 2012). Allerdings ist die bloße Norm ethisch korrekten Verhaltens sicher kein wirksamer Hemmfaktor, insbesondere wenn eigene Karrierechancen von den Publikationen abhängen – dafür sprechen jedenfalls die vielen empirischen Resultate, die eine hohe Prävalenz von PB anzeigen (Überblicke z. B. in Auspurg & Hinz 2011a; Dickersin 2005; Ferguson & Brannick 2012).

Den vermutlich wichtigeren Kostenaspekt stellt die drohende Sanktion im Falle einer Entdeckung des Fehlverhaltens dar. Nach Becker lassen sich die Kosten von Normverstößen als Produkt von Sanktionsschwere und Entdeckungswahrscheinlichkeit verstehen (Becker 1968: 177). In Bezug auf die Sanktionsschwere gilt allerdings, dass ein PB im Gegensatz zum offenen Betrug (wie dem Manipulieren von

[25] http://www.asanet.org/journals/previous_editors_reports.cfm (Zugriff am 21.03.2014).

Daten) weit weniger stark sanktioniert wird.[26] Feigenbaum & Levy (1996) zeigen insgesamt gar die Überflüssigkeit von offenem Betrug auf, da andere Strategien wie eben das Trimmen auf Signifikanz mit viel geringerem Aufwand bei zugleich geringerer Sanktionsschwere in der Lage sind, die gewünschten signifikanten Ergebnisse zu erzielen (es müssen nur einzelne Beobachtungen gelöscht, aber nicht ganze Datensätze erfunden werden). Diesen Überlegungen zufolge ist der PB vermutlich auch ein weitaus verbreiteteres Phänomen als der offene Betrug (Fanelli 2009: 6; Necker 2012).

Dies ist auch deshalb anzunehmen da Herausgeber und Gutachtende als *gatekeeper* ebenfalls kaum Anreize haben dürften, einen PB zu verhindern: Herausgeber, weil sie an möglichst hohen Zitationen ihrer Zeitschrift interessiert sind, schließlich wird die Reputation der Zeitschrift primär über den zitationsbasierten *Journal Impact Factor* gemessen; Gutachtende, weil sie den (zeitlichen) Aufwand der Begutachtung durch die Orientierung an scheinbaren Qualitäts-Proxys wie der Signifikanz gering halten können (eine solche Heranziehung von Proxy-Informationen lässt Arrow 1973; Phelps 1972 vermuten).

Alles in allem dürfte es für einen rationalen Forschenden nach der aufgezeigten Kosten-Nutzenstruktur daher lohnend erscheinen, primär signifikante Ergebnisse niederzuschreiben und zur Veröffentlichung einzureichen. Zwar besteht das soziale Optimum in einer vom PB unverzerrten Wissenschaft, da in dieser durch die Befolgung wissenschaftsethischer Normen ein höherer Wissensfortschritt erreicht wird. Um dieses Optimum zu erreichen ist jedoch das Vertrauen erforderlich, dass alle anderen Akteure ebenfalls die bewusste Bevorzugung von statistisch signifikanten Ergebnissen unterlassen und nicht die mit einem PB verbundenen individuellen Wettbewerbsvorteile nutzen (s. zur Annahme eines solchen sozialen Dilemmas: Auspurg & Hinz 2011a; Kerr 1998). Für viele Akteure dürfte unter den skizzierten Wettbewerbsstrukturen die Alternative, einen PB zu begehen, daher vermutlich lohnender erscheinen: *In der untersuchten Literatur sind Anzeichen auf einen PB zu finden ($H_1$).*

Im Vergleich zu den führenden US-amerikanischen Zeitschriften sind die Annahmequoten selbst in den zwei renommiertesten deutschsprachigen Soziologie-Zeitschriften, der ZfS und der KZfSS, deutlich höher, so konnten in den hier untersuchten elf Jahrgängen rund ein Drittel aller eingereichten Manuskripte veröffentlicht werden.[27] Die Konkurrenz um die limitierten Veröffentlichungsgelegenheiten ist also in der deutschen Soziologie erheblich geringer, womit zu erwarten ist: *Ein PB tritt in der deutschsprachigen Soziologie in geringerem Ausmaß auf als in den höher gerankten US-amerikanischen Zeitschriften ($H_2$).*

[26] Ein aufgedeckter Betrugsfall kann den Ausschluss aus der wissenschaftlichen Gemeinschaft nach sich ziehen, wie die vielen *Retractions* im Falle von Diederik Stapel oder der Verlust des Doktortitels im Falle von Hendrik Schön belegen (Stroebe et al. 2012). Ein PB ist dagegen nahezu sanktionslos, da es weitaus schwieriger bis unmöglich ist ein *vorsätzliches* Fehlverhalten nachzuweisen.

[27] Daten aus den Editorials der ZfS 2002-2011, sowie aus dem Autorenmerkblatt zum Entscheidungsverfahren der KZfSS (Daten zu den Jahren 2000-2006; http://www.uni-koeln.de/kzfss/konventionen/ksents.htm; Zugriff am 21.03.2014).

Einflüsse sind zudem auf der Kostenseite zu erwarten. Zunächst sollte das Produzieren von signifikanten und die Hypothesen bestätigenden Effekten umso leichter fallen, je „höher" die Freiheitsgrade des wissenschaftlichen Arbeitens sind. Beschränkungen der „Freiheitsgrade" können dabei projektintern auftreten: Mit zunehmender Anzahl der in den *einzelnen* Studien getesteten Hypothesen dürfte es immer schwieriger werden, einzelne Effekte durch Techniken des *Signifikanztunings* (Weglassen von Beobachtungen etc.) in der gewünschten Weise zu modifizieren. Denn mit der Manipulation einzelner Effekte werden zumindest die mit demselben Modell geschätzten weiteren Effekte ebenfalls tangiert, sodass es mit steigender Variablenanzahl immer anspruchsvoller sein sollte, viele Effekte signifikant zu rechnen: *Je mehr Hypothesen getestet werden, desto geringer ist das PB-Risiko ($H_3$).*[28]

Zwar sind Herausgeber wie vorangehend erörtert an hohen Zitationsraten interessiert, die Reputation ihrer Zeitschrift hängt aber zugleich auch von der gewissenhaften Einhaltung wissenschaftlicher Standards ab. Herausgebervorgaben können die Transparenz der Analysen durch mehr oder weniger ausführliche Dokumentationspflichten, durch Vorgaben wie die verbindliche Angabe der exakten Signifikanzniveaus (statt lediglich Sternchen) oder die Verpflichtung, Daten und Analysefiles für Replikationen zur Verfügung zu stellen, erhöhen. Werden keine Daten sowie Analysefiles zur Verfügung gestellt, ist eine Replikation zur Aufdeckung von Praktiken des *Signifikanztunings* nicht möglich (Feigenbaum & Levy 1993: 119). Generell ist daher anzunehmen, dass mit einer starken Dokumentationspflicht und zugänglichen Daten das Risiko für einen PB abnimmt: *Je umfangreicher die Berichtspflichten zu Datenmaterial und Analysen, desto geringer ist das PB-Risiko ($H_{4a}$). Wenn Daten allgemein zugänglich sind, dann kommt es seltener zu einem PB ($H_{4b}$).*

Soziale Kontrolle kann allerdings nicht nur durch externe Gutachtende, Herausgeber oder weitere Forschende erfolgen, sondern bereits durch am Projekt beteiligte Koautoren. Etwa führt nach der *Social Control Theory* von Hirschi (1969) eine stärkere Bindung zu normkonformen Akteuren zu einer wahrscheinlicheren Normbefolgung. Koautoren erhöhen zudem als „Mitwisser" die Gefahr eines sozialen Tadels (Auspurg & Hinz 2011a). Um das einen PB unterstützende Fehlverhalten gemeinsam zu tragen, wäre eine Einigung auf die Verletzung der Norm erforderlich und überdies das Vertrauen, dass es keinen *whistle blower* gibt, der die Reputation in der *Community* beschädigen könnte. Die Größe des Autorenteams beeinflusst neben der Entdeckungswahrscheinlichkeit zugleich aber auch andere Nutzenparameter, wie die Sanktionsschwere. So ermöglichen erst etwaige Mittäter eine Verantwortungsdiffusion.[29]

[28] Eine alternative Interpretation wäre, dass bereits wenige signifikante Ergebnisse die Publikationschancen hinreichend erhöhen, so dass es weniger wichtig ist, ob weitere Effekte ebenfalls signifikant sind. Ein solcher „Grenznutzen" der Anzahl signifikanter Ergebnisse für Publikationschancen ist bislang allerdings rein spekulativ, es gibt hierfür u. W. kein zwingendes theoretisches Argument (auch wenn dieses gelegentlich angeführt wird, siehe etwa Auspurg & Hinz 2011a).

[29] Zudem kann mit der Zahl der Autoren der Druck steigen, eine erfolgreiche Arbeit zu produzieren, da zumindest einer der Forschenden dringend auf eine erfolgreiche Publikation (beispielsweise für einen Ruf auf eine Professur) angewiesen ist. Allerdings kann man hier auch umgekehrt argumentieren, dass mit der Größe des Teams die Wahrscheinlichkeit steigt, dass zumindest ein hochreputierter Autor beteiligt ist, der aufgrund seines anerkannten Status nicht mehr auf signifikante Ergebnisse für Publikationszusagen angewiesen ist.

Insgesamt ist der „Nettoeffekt“ der Teamgröße daher schwer vorherzusagen. Vermutlich birgt aber gerade die Möglichkeit von *whistle blowern* das abschreckendste Risiko, dass das andernfalls schwer zu entlarvende Fehlverhalten überhaupt beanstandet wird. Denn zumindest für bewusste Datenfälschungen ist bekannt, dass diese insbesondere von Insidern in Form von Koautoren aufgedeckt werden (Stroebe et al. 2012: 673f.): *Je mehr Autoren an einer Publikation beteiligt sind, desto geringer ist das PB-Risiko ($H_5$).*

## 2.4. Forschungsstand und Methoden

### *2.4.1. Prävalenz signifikanter Ergebnisse*

Erste Erkenntnisse über die Prävalenz signifikanter Ergebnisse berichtet Sterling in seiner Untersuchung von vier US-amerikanischen Psychologie-Zeitschriften aus dem Jahr 1959. Er stellte fest, dass über 95% der Artikel überwiegend zum 5%-Niveau signifikante Effekte berichten (Sterling 1959: 32). Diesen Befund konnten Sterling et al. in ihrer Replikation aus dem Jahr 1995 mit über 92% solcher Artikel bestätigen. In der Medizin hingegen sind die Anteile überwiegend signifikanter Ergebnisse mit 43-82% deutlich geringer (Sterling et al. 1995: 109). Eine starke Häufung von Artikeln mit überwiegend signifikanten Ergebnissen wurde überdies für US-amerikanische Soziologie-Zeitschriften (80%) beobachtet (Wilson et al. 1973: 144). Auch in der deutschen Soziologie findet sich, bei einer analogen Methodik zu Sterling et al. (1959; 1995), in der KZfSS, der ZfS sowie der Sozialen Welt eine deutliche Überrepräsentanz signifikanter Ergebnisse. Im untersuchten Zeitraum (1965-1976) zeigen sich in 74% der Artikel zum überwiegenden Teil signifikante Ergebnisse, insgesamt sind gut 60% der getesteten Koeffizienten zum 5%-Niveau signifikant (Sahner 1979: 271). In einer Replikation für die Jahrgänge 2000-2010 lässt sich mit Anteilen von 66% bzw. 55% nur ein leicht rückläufiger Trend feststellen (Editorial der ZfS 2012) . Fanelli (2012) stellt hingegen im Zeitraum 1990-2007 und für internationale Zeitschriften insbesondere in den Sozialwissenschaften einen Anstieg signifikanter Ergebnisse fest.

Die Ergebnisse zur Prävalenz signifikanter Ergebnisse haben trotz ihrer Hinweise auf einen PB den großen Nachteil, dass sie ebenso zu Alternativerklärungen passen. So lässt sich einwenden, dass durch theoretisch gut fundierte Untersuchungen der Anteil an positiven Ergebnissen steigt (Diekmann 2011: 631). In theoretisch gut aufgestellten Disziplinen sollten demnach signifikante Ergebnisse eher die Regel als die Ausnahme sein. Ein ähnlich gelagerter Einwand gilt dem Vergleich von in Zeitschriften publizierten versus nicht-publizierten Ergebnissen (die sich z. B. in Studienverzeichnissen oder Konferenzbeiträgen finden lassen). Signifikante Studien haben insgesamt eine deutlich höhere Chance veröffentlicht zu werden (s. z. B. Dickersin et al. 1992; Easterbrook et al. 1991; Stern & Simes 1997), dieser Unterschied kann aber neben einem PB auch in der besseren methodischen und theoretischen Qualität der Studien begründet sein.

### *2.4.2. Auffälligkeiten in Testwerteverteilungen*

Alternative Testmethoden setzen daher noch näher an dem Phänomen des PB an, indem sie prüfen, ob sich Auffälligkeiten in den Verteilungen statistischer Testwerte (Teststatistiken)[30] finden, die auf ein Hintrimmen signifikanter Werte schließen lassen.

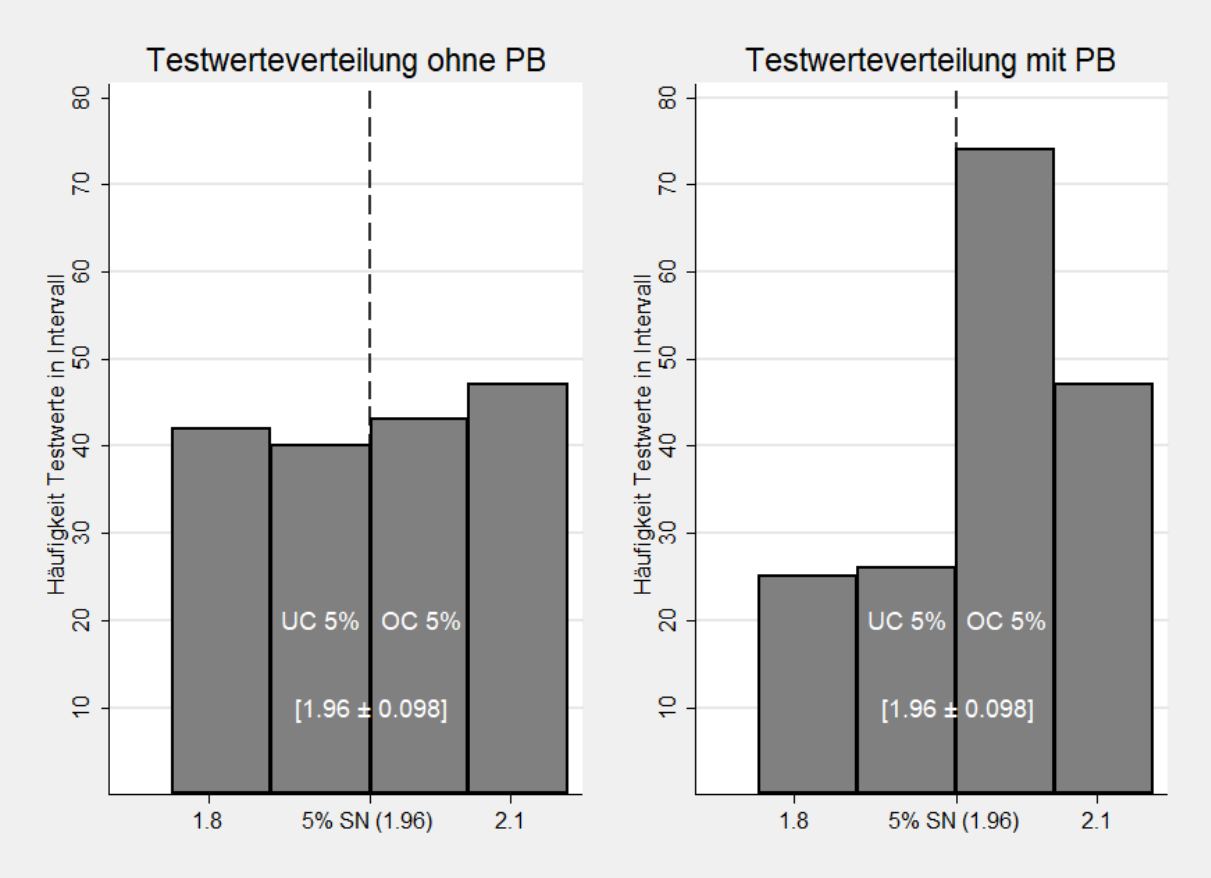
Die genaue Verteilung von Testwerten wie *p*-, *t*- oder *z*-Werten ist zwar in heterogenen Forschungsfeldern unbekannt, jedoch lässt sich die Testwerteverteilung zumindest als stetig annehmen (s. für eine mathematische Herleitung: Gerber & Malhotra 2008a: 11f.). Die Testwerteverteilung sollte daher– liegt kein PB vor – keine auffälligen Sprünge an den rein willkürlich gesetzten Signifikanzschwellen zeigen. Werden nun aber nicht-signifikante Ergebnisse unterdrückt und signifikante Ergebnisse bevorzugt ausgewählt, entsteht an der Signifikanzschwelle ein Sprung in der Verteilung, der die Annahme der Stetigkeit verletzt. Um dies testen zu können, wird von einigen Autoren eine Verteilungsannahme über die unbekannte Testwerteverteilung getroffen. Masicampo & Lalande (2012) sowie Leggett et al. (2013) finden dabei, dass *p*-Werte knapp unterhalb des 5%-Signifikanzniveaus deutlich häufiger auftreten, als man erwarten würde (Leggett et al. 2013: 2305; Masicampo & Lalande 2012: 2274). Beide Studien zeigen damit Anzeichen für eine Überrepräsentanz knapp signifikanter Ergebnisse und damit für einen PB.[31] Allerdings haben diese Studien den Nachteil, die lediglich zwischen 0 und 1 definierte *p*-Werte Verteilung, welche insbesondere kleine *p*-Werte sehr komprimiert und damit unscharf darstellt, zu verwenden. Zudem wird die zweifelhafte Annahme exponentialverteilter *p*-Werte im gesamten Testwertebereich [0; 1] getroffen.[32]

Das von Gerber & Malhotra (2006) vorgeschlagene Verfahren des Caliper-Tests (CT)[33] vermeidet hingegen weitreichende Verteilungsannahmen und stützt sich lediglich auf die vorangehend erläuterte, konservativere Annahme einer stetigen Testwerteverteilung. Bei stetiger Testwerteverteilung liegt um die Signifikanzschwelle bei Betrachtung kleiner Intervalle („Caliper“) näherungsweise eine Binomialverteilung der Testwerte mit einer 50%-Wahrscheinlichkeit für Werte über oder unter der Signifikanzschwelle vor. Man spricht von einem x%-Caliper, wenn ein Intervall von jeweils x% der Signifikanzschwelle um diese Schwelle herum betrachtet wird. Wenn kein PB vorliegt, sollten in diesem Intervall gleich viele Werte knapp signifikant (Over-Caliper; OC) und knapp nicht signifikant (Under-Caliper; UC) sein; bei Vorliegen eines PB ist dagegen ein Fehlen von Werten im UC und eine Häufung von Werten im OC zu beobachten. Dies ist in Abbildung 2-1 schematisch dargestellt: Bei Vorliegen eines

[30] Die Begriffe ‘Teststatistiken’ und ‘Testwerte’ werden im Folgenden synonym verwendet, gemeint sind beispielsweise *p*-, *t*-, oder *z*-Werte.
[31] Auch Brodeur et al. (2013) finden in ihrer grafischen Analyse deutlich mehr knapp signifikante als nichtsignifikante *z*-Werte.
[32] Diese ist nur empirisch abgeleitet, jedoch nicht theoretisch begründet.
[33] Ursprünglich geht der CT auf eine Idee von Edward Tufte zurück, der die Methodik in dem unveröffentlichten (und inzwischen leider auch verschollenen Manuskript) „Evidence Selection in Statistical Studies of Political Economy: The Distribution of Published Statistics“ anwendet (Gerber & Malhotra 2008b: 315). Wie auch Gerber und Malhotra anmerken, ist es angesichts des Forschungsthemas fast eine Ironie, dass dieses aus dem Jahr 1985 stammende Manuskript nie veröffentlicht wurde (oder werden konnte?).

PB (rechte Grafik) ist die Verteilung unterhalb der Signifikanzschwelle (gestrichelte Linie bei einem *t*-Wert von 1.96. Dieser Wert markiert das 5%-Signifikanzniveau markiert) unterbesetzt, während eine auffällige Häufung von Werten im OC zu beobachten ist. Da die Signifikanzschwellen willkürlich gewählt sind, gibt es kaum eine andere Ursache für solche Muster als einen PB. Mit Binomialtests und vergleichbaren Testverfahren lässt sich dann feststellen, ob der Sprung an der Signifikanzschwelle signifikant von der ohne einen PB zu erwartenden Gleichverteilung abweicht.

**Abbildung 2-1 Schematische Darstellung des CT**




*Erläuterung:* Dargestellt ist das Testprinzip für den 5%-Caliper um das 5%-Signifikanzniveau, welches einem t-Wert von 1.96 entspricht (gestrichelte Linie). Der 5%-Caliper deckt alle t-Werte ab, die in den Bereich von 1,862 bis 2,058 fallen. Die Verteilung dieser Werte wird durch die beiden Histogrammbalken links (UC) und rechts (OC) von der gestrichelten Linie abgebildet. Bei Vorliegen von PB ist der OC deutlich stärker besetzt als der UC, während bei Fehlen von PB in etwa eine stetige Gleichverteilung zu beobachten ist.

Der Caliper sollte dabei so klein wie möglich gewählt werden, um die Annahme der gleichverteilten Binomialverteilung (Anteile von 50% über und unter der Signifikanzschwelle) rechtfertigen zu können (Gerber & Malhotra 2006: 12; für nähere Erläuterungen s. Onlineanhang).[34] Jedoch ist zu beachten, dass ein kleiner Caliper alle Testwerte, die außerhalb des Calipers liegen, ausschließt und damit die statistische Power des Binomialtests verringert. In der Studie von Auspurg und Hinz (2011a: 652f.) verbleiben von ursprünglich 587 Testwerten im 3%-Caliper nur 22 Testwerte im OC und 11 im UC. Der Caliper

[34] Siehe http://www.uni-koeln.de/kzfss/materialien/KS-66-4-Auspurg.zip, aktuell (27.01.2018) leider nicht mehr abrufbar aufgrund von Umstellungen der Homepage der KZfSS.

sollte daher auch in Abhängigkeit von den insgesamt erfassten Testwerten hinreichend groß gewählt werden, um eine ausreichende statistische Power zur Entdeckung des PB zu erreichen.

In ihren drei, den CT einsetzenden Studien stellen Gerber und Malhotra in der US-amerikanischen Soziologie (Gerber & Malhotra 2008a: 17) und ebenso in Zeitschriften der Politikwissenschaft (Gerber & Malhotra 2008b: 318) mehr knapp zum 5%-Niveau signifikante als knapp nicht signifikante Testwerte und damit deutliche Hinweise auf einen PB fest. In den von Gerber und Malhotra untersuchten Artikeln in soziologischen Zeitschriften sind signifikante Werte im 5%-Caliper um das 5%-Signifikanzniveau mit über 78% deutlich häufiger vertreten als nicht-signifikante Werte. Dieses Ergebnis ist unter der Nullhypothese einer Gleichverteilung mit einer verschwindend geringen Chance von 1:100.000 zu erwarten (Gerber & Malhotra 2008a: 21). Ähnliche Muster wurden von den Autoren auch für zwei in der Politikwissenschaft verbreitete Fragestellungen (zum *economic voting* und *negative advertising*) gefunden (Gerber et al. 2010).

Auspurg & Hinz (2011a) stellen in ihrer Anwendung des CT auf eine Zufallsauswahl von 50 Artikeln, die in der ZfS, der KZfSS und der Sozialen Welt im Zeitraum von 2000-2010 erschienen sind, ebenfalls Hinweise auf einen PB fest, jedoch erscheint dort die Überrepräsentanz knapp signifikanter Testwerte im 3%- (mit 67%) und im 5%-Caliper (mit 60%) für das 5%-Signifikanzniveau deutlich weniger stark ausgeprägt als in den US-amerikanischen Zeitschriften (Auspurg & Hinz 2011a: 653).[35] Die Ergebnisse von Auspurg und Hinz konnten auch von Weiß & Berning (2013) repliziert werden.[36]

Neben einer reinen Deskription untersuchen Auspurg & Hinz (2011a) sowie Brodeur et al. (2013) zudem den Einfluss von Text- und Autorenmerkmalen auf die Testwerteverteilung. Auspurg & Hinz (2011a) finden in bivariaten Analysen, dass explizit formulierte Hypothesen sowie Alleinautorschaften das Risiko für einen PB tendenziell erhöhen (Auspurg & Hinz 2011a: 654). Allerdings handelt es sich lediglich um eine Pilotstudie, die aufgrund der geringen Fallzahl keine aussagekräftigeren multivariaten Analysemethoden verwendet und auch nur begrenzte Aussagen zur Prävalenz des PB in der deutschen Soziologie zulässt.[37]

In einem grafischen Vergleich ihrer Kerndichteschätzer zeigen Brodeur et al. (2013), dass bei Artikeln, die zur Verdeutlichung des Signifikanzniveaus Sternchen verwenden, ebenso wie bei Wissenschaftlern am Anfang ihrer Karriere (gemessen am akademischen Alter sowie noch keiner Herausgeberschaft bei Zeitschriften) knapp signifikante Ergebnisse deutlich häufiger anzutreffen sind als knapp nicht-signifikante Ergebnisse (Brodeur et al. 2013: 9). Die Verfügbarkeit der Daten scheint hingegen keinen

[35] Im Gegensatz zu den anderen berichteten Studien verwenden Auspurg & Hinz (2011a) die genaueren *t*- statt *z*-Werte, was allerdings nur bei wenigen Artikeln mit sehr kleinen Fallzahlen zu unterschiedlichen Ergebnissen führt.
[36] Die Autoren kommen jedoch in ihrer Schlussfolgerung trotz signifikant überrepräsentierter Testwerte im OC des 3%-Caliper zu dem Ergebnis, dass kein PB vorliegt.
[37] Auch Weiß & Berning (2013) können den Effekt der Mehrfachautorenschaft replizieren. Die Autoren analysieren überdies den Effekt des Status von Autoren, allerdings sind die Berechnungen und Ergebnisse aufgrund der knappen, überwiegend grafischen Darstellung in Form einer Posterpräsentation schwer zu deuten.

Effekt zu haben. Allerdings wird hier lediglich explorative Forschung mit ad-hoc Annahmen und grafischen Inspektionen betrieben.

## 2.5. Daten und Methoden

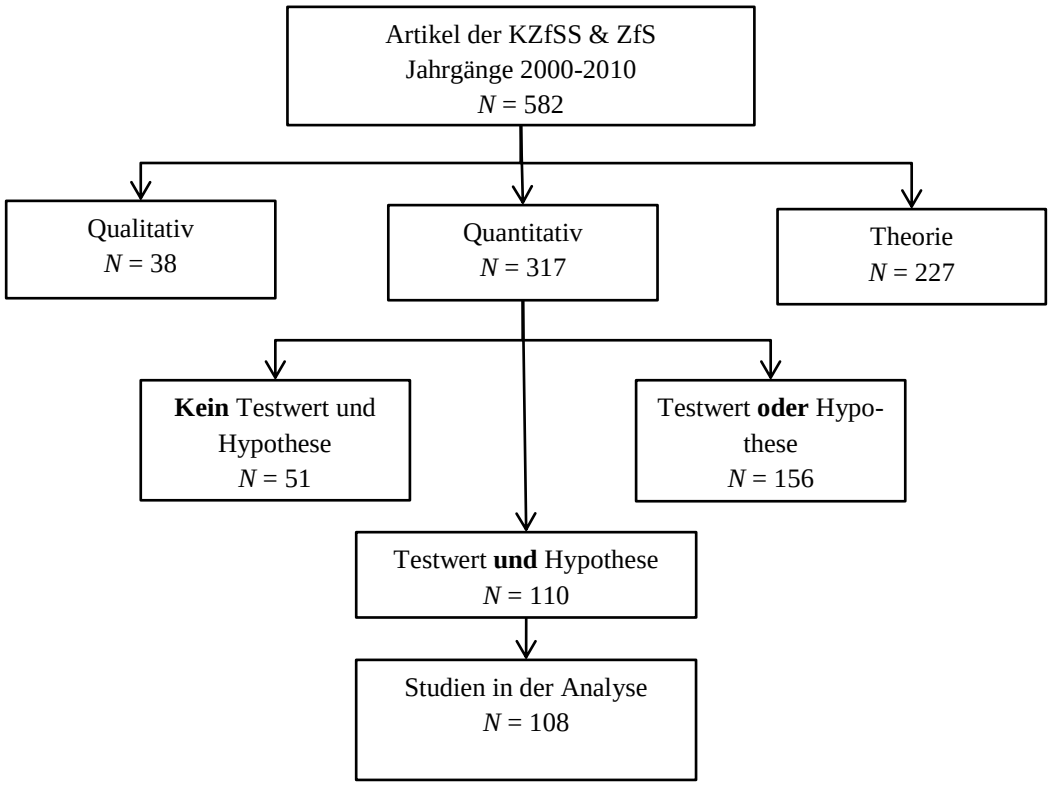
Grundgesamtheit der vorliegenden Untersuchung bilden alle im Zeitraum von 2000 bis 2010 in der KZfSS sowie der ZfS erschienenen Artikel. Beide Zeitschriften stellen mit deutlichem Abstand die meist rezipierten Soziologie-Zeitschriften des deutschsprachigen Raumes dar, was eine gewisse Prävalenz für einen PB erwarten lässt.[38] Die vollständige Erfassung eines Elfjahreszeitraums war von dem Gedanken motiviert, besser als in der Vorläuferstudie (die nur einen zufälligen Ausschnitt beobachtete) eine solide Abschätzung der Prävalenz des PB zu erreichen. Eine Begrenzung auf elf Jahre war zudem aus forschungspragmatischen Gründen erforderlich (die Kodierung der einzelnen Koeffizienten ist sehr zeitintensiv, weil dazu alle Artikel gelesen werden müssen). Überdies wurde der zusammenhängende Veröffentlichungszeitraum der zwei führenden Zeitschriften deshalb gewählt, da in der vorliegenden Analyse keine Trendhypothesen zu zeitlichen Veränderungen interessieren und diese Eingrenzung eine möglichst hohe Standardisierung von möglichen Drittvariablen verspricht.

Der in Abbildung 2-2 dargestellte Kodierprozess[39] kann in drei Phasen aufgeteilt werden: In einem ersten Schritt werden alle Artikel der ZfS ($N$ = 322) sowie der KZfSS ($N$ = 260) anhand ihrer Ausrichtung aufgeteilt. Es wird zwischen theoretischen, qualitativen sowie quantitativen Artikeln ($N$ = 317) unterschieden. Studien, die sowohl quantitative als auch qualitative Methoden einsetzen ($N$ = 9), werden dabei als quantitative Forschung kodiert. Ein großes Problem stellt die Berechnung der Testwerte dar, da entweder der Testwert direkt vorliegen oder aus Koeffizient und Standardfehler berechnet werden muss. Studien, die nur Regressionskoeffizienten mit Sternchen zur Verdeutlichung der Signifikanz berichten oder ausschließlich deskriptiv arbeiten ($N$ = 129), keine Hypothesen aufstellen ($N$ = 27), sowie Studien die weder Standardfehler noch Hypothesen berichten ($N$ = 51), können aus diesem Grund nicht in den Datensatz aufgenommen werden. Die Anzahl der verwendbaren Texte reduziert sich damit drastisch von 317 auf 110. Gründe für diese Reduktion der Fallzahlen sind vor allem explorative Forschung bzw. rein deskriptive Auswertungen. Von den quantitativen Studien lassen sich mit diesen Auswahlkriterien insgesamt nur knapp 35% für die nachfolgenden Auswertungen berücksichtigen ($N$ = 108 Studien).[40] Diese deutliche, aber unvermeidliche Reduktion der einzubeziehenden Studien geht vermutlich mit einer Unterschätzung der Prävalenz des PB einher, da dieser der theoretischen Betrachtung zufolge mit schlecht dokumentierten Analysen einhergehen dürfte. Die hier primär beabsichtigten Analysen zu den Risikofaktoren sollten gleichwohl aussagekräftig sein.

[38] Vgl. ein Impact Factor der ZfS von 0,604 bzw. 0,481 bei der KZfSS (Journal Citation Report 2012).
[39] Die Darstellung des Kodierprozesses orientiert sich an den Empfehlungen der PRISMA Guidelines (Moher et al. 2009: 3).
[40] Zwei Studien fallen weg, da kein $t$-Wert berechenbar war ($p$ = 0 für alle getesteten Koeffizienten) oder umgekehrte Hypothesen (Nullhypothese sollte nicht verworfen werden) aufgestellt wurden.

Werden in Studien zum Test der Hypothesen mehrere Modelle berichtet, so werden (abweichend zu den Studien von Gerber & Malhotra 2008a,b) stets die durch Hypothesen spezifizierten Testwerte aus dem „vollen Modell“, also dem Modell mit allen Kontrollvariablen, erfasst. Nur wenn Hypothesentestungen sich explizit nur auf Modelle ohne vollständige Aufnahme aller Kontrollvariablen beziehen, werden die Testwerte aus diesen Modellen erhoben. Diese Methodik der Beschränkung auf ein Modell hat den Vorteil, dass sich Doppelerfassungen von Testwerten zum gleichen Effekt weitestgehend vermeiden lassen. Zudem erscheint für die Frage, ob sich eine Hypothese bestätigt oder nicht, das volle Modell aussagekräftiger. Insgesamt wurden 1618 Testwerte aus 108 Artikeln erfasst. Mit 68 zusätzlichen Artikeln geht diese Studie damit deutlich über die Vorläuferstudie von Auspurg & Hinz (2011a) hinaus.

**Abbildung 2-2 Flussdiagramm zum Kodierprozess**




Aus den Artikeln werden auch relevante Kovariaten erfasst. Die *Mehrfachautorenschaft* wird dichotom kodiert, da lediglich eine sehr geringe Anzahl an Artikeln ($N$ = 16) mehr als zwei Autoren aufweist. Insgesamt wurden knapp 60% der Artikel von mehr als einem Autor verfasst (s. Tabelle 2-1 für deskriptive Statistiken). Die *Anzahl der in den Artikeln getesteten Hypothesen bzw. Koeffizienten* unterscheidet sich stark, so werden in einer Studie 82 Koeffizienten getestet, wohingegen in anderen Studien nur ein einziger auf Hypothesen bezogener Koeffizient berichtet wird. In Anbetracht dieser stark rechtsschiefen Verteilung wird diese Variable logarithmiert.

**Tabelle 2-1 Deskriptive Statistiken der verwendeten Artikel**

| | Mittelwert | *SD* | Median | Minimum | Maximum |
|---|---|---|---|---|---|
| **Anzahl Koeffizienten (log)** | 2,423 | 0,958 | 2,485 | 0 | 4,407 |
| **Data Policy ZfS** | 0,417 | 0,495 | 0 | 0 | 1 |
| **Berichtspflicht ZfS** | 0,185 | 0,390 | 0 | 0 | 1 |
| **Datenzugang** | 0,556 | 0,499 | 1 | 0 | 1 |
| **Mehrfachautorenschaft** | 0,593 | 0,494 | 1 | 0 | 1 |
| **implizite Hypothesen** | 0,333 | 0,474 | 0 | 0 | 1 |

Als Proxy für strengere *Berichtspflichten für Autoren* werden die in der ZfS ab dem Jahr 2009 bestehenden Standards in den Autorenrichtlinien verwendet, die Autoren vorschreiben, immer auch Standardfehler und deskriptive Statistiken zu berichten sowie die „Formulierungen von Items und (Um-) Codierungen von Merkmalsausprägungen" einschließlich ihrer genauen Datenquellen wiederzugegeben.[41] Insgesamt fallen knapp 19% der Artikel unter diesen Standard. Der *Datenzugang* wird durch zwei Variablen operationalisiert. Zunächst als generelle Verfügbarkeit der Daten: Als öffentlich zugänglich gilt ein Datensatz, sofern er über GESIS verfügbar ist (z. B. ALLBUS). Darüber hinaus wurde auch in den Artikeln selbst und im Internet nach Angaben zum Datenzugang recherchiert. Alle Datensätze, die als verfügbar (für Sekundäranalysen) ausgewiesen sind, werden – auch wenn dies noch die Zustimmung des Rechteinhabers oder andere Formalitäten erfordert – als „zugänglich" kodiert. Insgesamt ist dieses Kriterium für 55% der Artikel erfüllt. Als alternative Operationalisierung wird die ab 2002 geltende *Data Policy* der ZfS gewählt, mit deren Einführung alle Autoren in einer Verpflichtungserklärung zustimmen, die für die Replikation ihrer Ergebnisse nötigen Daten und Methodenanleitungen auf Anfrage zur Verfügung zu stellen (Diekmann et al. 2002). Knapp 42% der verwendeten Artikel sind in der ZfS im Jahr 2002 oder später erschienen und unterliegen deshalb der *Data Policy*.[42] Auf eine Unterscheidung zwischen der ZfS und der KZfSS wird in den Analysen aufgrund der großen Überschneidung zwischen Zeitschrift und Berichtspflicht bzw. *Data Policy* verzichtet. Ferner wird als Kontrollvariable aufgrund des Befundes von Auspurg & Hinz (2011a), dass bei explizit formulierten Hypothesen ein PB stärker auftritt als bei impliziten, nach der Art der Hypothesenformulierung unterschieden: Implizite Hypothesen sind in den Artikeln erwähnte Annahmen, die im Gegensatz zu expliziten Hypothesen aber nicht ausdrücklich von den Autoren selbst als Hypothesen ausgeflaggt werden (Kodierbeispiel s. Onlineanhang). Im vorliegenden Datensatz sind in gut 33% der Artikel die Hypothesen lediglich implizit formuliert.

[41] Siehe die Autorenhinweise: http://www.zfs-online.org/index.php/zfs/information/authors (Zugriff am 21.03.2014).

[42] Die verbleibenden 58% der 108 in den Analysen berücksichtigten Artikel sind in der ZfS 2000 bzw. 2001 sowie in der KZfSS erschienen.

***Testverfahren und Schätzmodelle***

Für die nachfolgenden Auswertungen wird der CT herangezogen (siehe Kapitel 2.4.2). Insbesondere im Hinblick auf die in der Literatur sehr stark variierenden Fallzahlen und damit Freiheitsgrade werden allerdings abweichend zu den Vorläuferstudien *t*-Werte, und nicht *z*-Werte zur Berechnung der Caliper verwendet (die mit *z*-Werten verbundene Normalverteilungsannahme ist nur näherungsweise, also nur bei hinreichend großen Fallzahlen, erfüllt). Entsprechend wurde eine freiheitsgradabhängige Signifikanzschwelle zur Bestimmung des OC und UC gewählt (Details zur Berechnung der Calipers s. Kapitel 2.8).[43] Zudem werden für die Testwerte zufällige Nachkommastellen imputiert, um robuste Ergebnisse im Hinblick auf potentielle Rundungsungenauigkeiten zu erhalten, welche die Annahme der stetigen Testwerteverteilung verletzen könnten.

Vor dem Hintergrund des Problems multipler Tests beschränken sich die Analysen auf nur vier Caliper (Intervalle von 3, 5, 10 und 15% um das Signifikanzniveau). Je mehr Tests auf einen PB durchgeführt werden, desto größer ist die Wahrscheinlichkeit, dass ein Test rein per Zufall statistische Signifikanz erreicht. Eine Methode zur Korrektur dieses Problems stellt die an die Bonferroni-Korrektur angelehnte Vorgehensweise von Holm (1979) dar. Die *p*-Werte werden hierbei nach ihrer Größe aufsteigend sortiert und sequentiell mit der Anzahl der Tests minus des jeweiligen Schritts multipliziert, um damit auf die Anzahl der Tests angepasste (vergrößerte) Signifikanzniveaus zu erhalten (symbolisiert mit „k.*p*“*)*. Das von Holm vorgeschlagene Verfahren ist aufgrund der höheren statistischen Power der klassischen Bonferroni-Korrektur vorzuziehen (Holm 1979: 67).

Für bivariate Analysen von Zusammenhängen mit den vier Calipern werden für kontinuierliche unabhängige Variablen der Korrelationskoeffizient *r* nach Pearson und für dichotome abhängige Variablen das Zusammenhangsmaß *phi* verwendet. Als abhängige Variable dient jeweils die binäre CT-Variable mit den beiden Ausprägungen UC (0) und OC (1).

Aufgrund der hohen Anzahl an erfassten Artikeln lassen sich (anders als in der Vorläuferstudie) auch multivariate logistische Regressionsmodelle schätzen. Zur besseren Interpretierbarkeit der Ergebnisse werden durchschnittliche Marginaleffekte (*Average Marginal Effects*, kurz AMEs) berechnet. Diese erlauben auch eine bessere Vergleichbarkeit von Modellen (Auspurg & Hinz 2011b). Die AMEs zeigen (gemessen in Prozentpunkten) die durchschnittliche Veränderung der Wahrscheinlichkeit an, dass Testwerte in den OC statt UC fallen, bei marginaler Änderung der jeweiligen unabhängigen Variablen. Dabei wird für jeden Caliper (3, 5, 10 und 15%) ein separates Modell geschätzt. Auf Kontrollvariablen außer der Art der Hypothese (explizit oder implizit) wird im Hinblick auf die Instabilität der Maximum-Likelihood Schätzung bei geringen Freiheitsgraden (kleine Fallzahlen und hohe Anzahl zu schätzender Parameter) verzichtet (Hart & Clark 1999). Aus diesem Grund werden über die logistischen Regressi-

[43] Die Freiheitsgrade der Modelle wurden in den Studien oft nicht berichtet. Es wird daher die Fallzahl als Proxy verwendet.

onsmodelle hinaus auch die bei kleinen Fallzahlen stabileren linearen Regressionsmodelle (*Linear Probability Model*, LPM) verwendet. Die Werte beider Modelle sollten jedoch sehr ähnlich ausfallen (Greene 2012: 687). Um der Mehrebenenstruktur (L1: Testwerte; L2: Artikel) Rechnung zu tragen, werden die Modelle mit geclusterten Standardfehlern geschätzt (Rogers 1994).

## 2.6. Ergebnisse

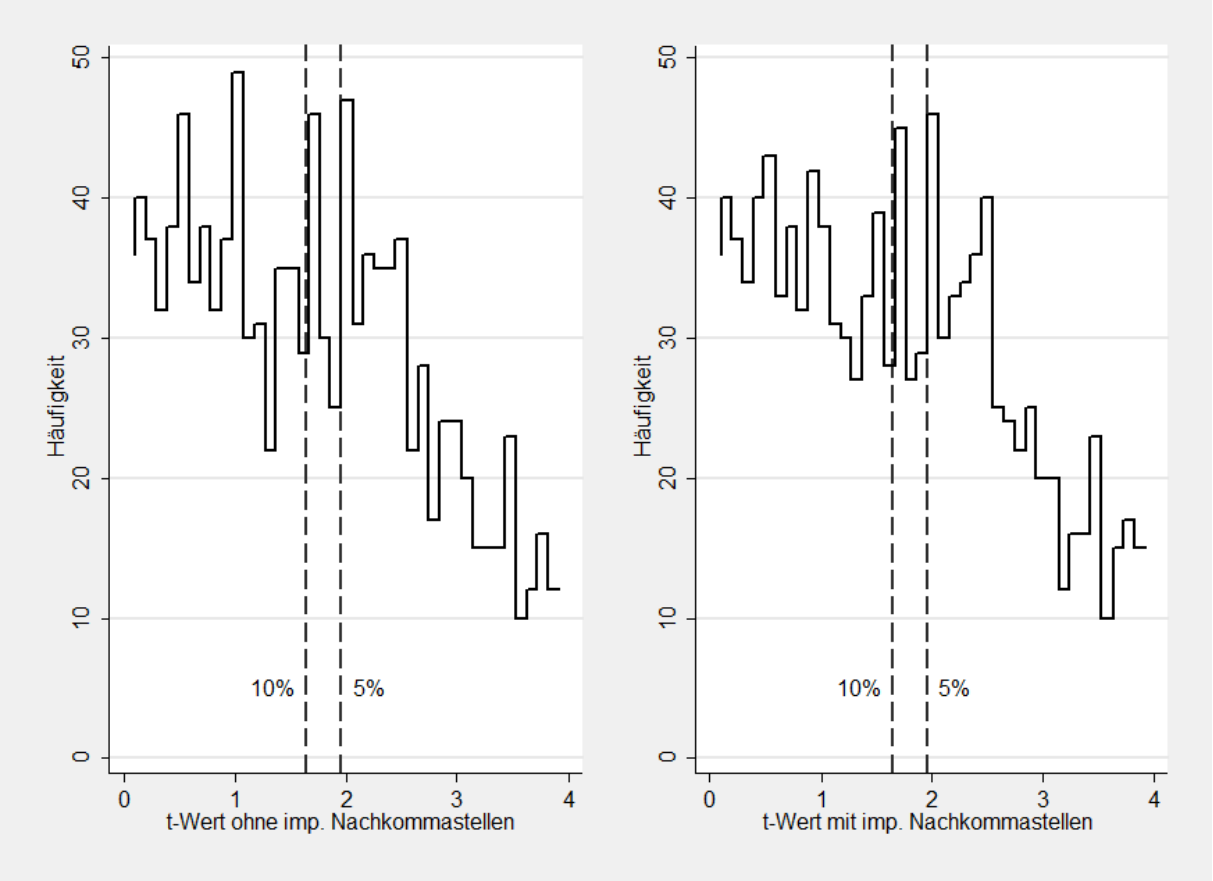
Die starke Verbreitung des 5%-Signifikanzniveaus wird auch in den untersuchten Artikeln deutlich. Insgesamt stützen sich gut 49% aller Koeffizienten auf das 5%-Signifikanzniveau als höchstes berichtetes Signifikanzniveau. Weitere 43% der Koeffizienten werden anhand des 10%-Niveaus beurteilt.[44] Im Folgenden wird zuerst die Verteilung der Testwerte (Abbildung 2-3) näher betrachtet, um darauf aufbauend den CT auf die in der Soziologie am weitesten verbreiteten Signifikanzniveaus (5% bzw. 10%) anzuwenden.

In den vorliegenden Daten sind 56,2% der 1618 getesteten Koeffizienten auf dem 5%-Signifikanzniveau signifikant. Dieser Wert liegt um knapp vier Prozentpunkte leicht unter den Ergebnissen von Sahner (1979). Dabei berichten 58,2% der 108 Artikel überwiegend auf dem 5%-Niveau signifikante Ergebnisse, was einen deutlicheren Rückgang im Vergleich zu Sahner (1979) um 16 Prozentpunkte und eine noch stärkere Abweichung von den Ergebnissen in der US-amerikanischen Psychologie (Sterling 1959) bedeutet.

Bei der grafischen Inspektion der Daten (Abbildung 2-3, linke Seite) fällt für die Testwerte ab dem Erreichen der kritischen Signifikanzschwellen für das 5% und 10%-Niveau (*t*-Werte größer 1,96 bzw. 1,64 – siehe die gestrichelten Linien) ein abrupter Anstieg der Häufigkeit auf. Ergebnisse, die diese Signifikanzschwellen knapp überschreiten, kommen fast doppelt so häufig vor als solche, die knapp unter dem jeweiligen Signifikanzniveau liegen. Diese für einen PB sprechenden Ergebnisse bleiben auch nach der Bereinigung etwaiger Rundungsungenauigkeiten stabil. Die bei den Testwerten um 0,5 bzw. 1 auftretenden Schwankungen sind hingegen schwer zu erklären. Die Imputation von zufälligen Nachkommastellen zur Korrektur von möglichen Rundungsungenauigkeiten (s. Abbildung 2-3, rechts) löst dieses Problem nur in Teilen auf.

[44] Nur 1% der Koeffizienten werden anhand des 1%-Signifikanzniveaus diskutiert. Das verwendete Signifikanzniveau wird selten explizit ausgewiesen, die Kodierung stützt sich daher primär auf die Legenden unter Ergebnistabellen. In gut 6% der Fälle konnte keinerlei Signifikanzniveau ermittelt werden.

**Abbildung 2-3 Histogramm der *t*-Werteverteilung zum 5%-Caliper**




Im univariaten Test bestätigen sich die grafischen Befunde auch bei Berücksichtigung der freiheitsgradabhängigen Signifikanzschwellen (*t*-Werte). Dies gilt zunächst für die Ergebnisse des CT um das 5%-Signifikanzniveau (Tabelle 2-2). Im 3%- und 5%-Caliper sind statistisch signifikante Studien mit 64% bzw. 62% deutlich überrepräsentiert. Im 10%- bzw. 15%-Caliper nähern sich die Werte mit 56% und 53% der erwarteten Gleichverteilung (d. h. 50%) an. Die ungleiche Verteilung der *t*-Werte ist für den 3%- und 5%-Caliper zum 5%-Niveau signifikant.[45] Der Unterschied ist zudem auch im 10%-Caliper auf dem 10%-Niveau signifikant. Nach der Holm-Korrektur der *p*-Werte für mehrfache Tests ist für den engsten Caliper nur noch ein schwach signifikanter Unterschied festzustellen (10%-Niveau). Der Unterschied im 5%-Caliper verfehlt dabei das 10%-Signifikanzniveau nur knapp. Zumindest in den engen Calipern ist das Auftreten einer derart asymmetrischen Verteilung ohne Bevorzugung signifikanter Ergebnisse sehr unwahrscheinlich (lediglich eine Chance von rund 1:30 bzw. selbst nach der konservativen Holm-Korrektur 1:10 im 3- bzw. 5%-Caliper). Das Vorliegen eines PB ist somit sehr wahrscheinlich (*H1*).

[45] In Analogie zu Gerber und Malhotra (2006, S. 316f.) werden die *p*-Werte des einseitigen Binomialtests berichtet.

**Tabelle 2-2 Caliper-Tests zum 5%-Signifikanzniveau**

| | N | OC | UC | % OC | *p*-Wert [a] | k. *p*-Wert [a] |
|---|---|---|---|---|---|---|
| 3%-Caliper | 50 | 32 | 18 | 0,640 | 0,032 | 0,097 |
| 5%-Caliper | 71 | 44 | 27 | 0,620 | 0,028 | 0,114 |
| 10%-Caliper | 133 | 75 | 58 | 0,564 | 0,083 | 0,165 |
| 15%-Caliper | 217 | 115 | 102 | 0,530 | 0,208 | 0,208 |

[a] einseitiger *p*-Wert

Betrachten wir nun die Ergebnisse des CT um das 10%-Signifikanzniveau (Tabelle 2-3). Im Fall des weniger verbreiteten 10%-Signifikanzniveaus ist die Besetzung des OC und UC insgesamt deutlich schwächer. Zwar sind auch hier knapp signifikante Studien mit 58% im 5%-Caliper, bzw. 55% und 52% im 10%- und 15%-Caliper, häufiger vertreten als knapp nicht-signifikante. Im sensibelsten 3%-Caliper ist jedoch eine exakte Gleichverteilung festzustellen. Nach der Holm-Korrektur erreicht keiner der Caliper statistische Signifikanz (k.$p > 0,4$).

Die Prävalenz von einem PB ist damit deutlich geringer als in der US-amerikanischen Sozialwissenschaft. So ist der Anteil gerade signifikanter Koeffizienten in allen Calipern (5%, 10%, 15%) signifikant geringer ($p < 0,045$) als in der Studie von Gerber & Malhotra (2008a).[46] Diese deutlichen Unterschiede bleiben auch nach der Holm-Korrektur auf dem 10%-Niveau signifikant (vgl. Tabelle 2-9). Somit scheint die US-amerikanische Forschung zumindest im Bereich der Soziologie erwartungsgemäß (*H2*) deutlich stärker von einem PB betroffen zu sein als die deutschsprachige.

**Tabelle 2-3 Caliper-Tests zum 10%-Signifikanzniveau**

| | *N* | OC | UC | % OC | *p*-Wert [a] | k. *p*-Wert [a] |
|---|---|---|---|---|---|---|
| 3%-Caliper | 34 | 17 | 17 | 0,5 | 0,568 | 0,568 |
| 5%-Caliper | 65 | 38 | 27 | 0,585 | 0,107 | 0,429 |
| 10%-Caliper | 118 | 65 | 53 | 0,551 | 0,156 | 0,467 |
| 15%-Caliper | 173 | 89 | 84 | 0,514 | 0,381 | 0,761 |

[a] einseitiger *p*-Wert

Die Risikofaktoren für die Ungleichverteilung als Indiz für einen PB werden nachfolgend für das 5%-Signifikanzniveau näher untersucht. Das 10%-Signifikanzniveau wird aufgrund von niedrigeren Fallzahlen und damit einhergehender geringer Power von den Analysen ausgeschlossen. Der stärkste Effekt in den bivariaten Analysen findet sich mit $r = -0,184$ für die logarithmierte Koeffizientenanzahl (vgl. Tabelle 2-4). Je mehr Koeffizienten Autoren in ihr Modell aufnehmen, desto geringer ist das PB-Risiko. Dieser Effekt ist jedoch, auch bedingt durch die geringe Fallzahl, nicht signifikant ($p = 0,202$). Die Mehrfachautorenschaft, sowie die Berichtspflicht, die *Data Policy* der ZfS und der Datenzugang zeigen

[46] Ergebnisse von Zwei-Stichproben *z*-Tests. Zur besseren Vergleichbarkeit mit den Ergebnissen von Gerber & Malhotra (2008a) wurden die Ergebnisse mit *z*-Werten berechnet.

in der bivariaten Analyse ebenfalls keine substanziellen Effekte auf das PB-Risiko in den verschiedenen Calipern ($r < 0{,}042$).

**Tabelle 2-4 Bivariate Korrelationen von knapp signifikanten Ergebnissen (OC statt UC) mit Randbedingungen (5%-Signifikanzniveau, unterschiedliche Caliper-Breiten)**

| | 3%-Caliper | | 5%-Caliper | | 10%-Caliper | | 15%-Caliper | |
|---|---|---|---|---|---|---|---|---|
| | ***r*** | ***p*-Wert** | ***r*** | ***p*-Wert** | ***r*** | ***p*-Wert** | ***r*** | ***p*-Wert** |
| Anzahl Koeffizienten (log) | -0,184 | 0,202 | -0,135 | 0,263 | -0,078 | 0,371 | -0,063 | 0,358 |
| Data Policy ZfS | 0,001 | 0,832 | 0,008 | 0,448 | 0,022 | 0,086 | 0,000 | 0,848 |
| Berichtspflicht ZfS | 0,000 | 0,977 | 0,000 | 0,972 | 0,001 | 0,770 | 0,001 | 0,665 |
| Datenzugang | 0,004 | 0,670 | 0,011 | 0,368 | 0,000 | 0,833 | 0,002 | 0,549 |
| Mehrfachautorenschaft | 0,042 | 0,145 | 0,035 | 0,113 | 0,008 | 0,309 | 0,005 | 0,316 |
| implizite Hypothesen | 0,002 | 0,768 | 0,001 | 0,842 | 0,002 | 0,567 | 0,006 | 0,257 |

*Anmerkungen*: Bei dichotomen Variablen wird als Korrelationskoeffizient *phi* berichtet. Alle *p*-Werte stammen von zweiseitigen Tests.

Gleichwohl sollen auch multivariate Modelle geschätzt werden, da ausgeprägte Korrelationen zwischen den Variablen bestehen und sich Effekte womöglich erst unter Kontrolle der anderen Variablen zeigen.[47] Es werden im Folgenden AMEs aus logistischen Regressionen mit allen Kovariaten berichtet (s. Tabelle 2-5).[48] Lediglich die beiden auf die ZfS bezogenen Variablen (Berichtspflicht und *Data Policy*) werden in getrennten Modellen geschätzt, um die andernfalls auftretenden Multikollinearitätsprobleme zu umgehen.

Im kleinsten berechneten Ausschnitt um das 5%-Signifikanzniveau, dem 3%-Caliper, lässt sich mit steigender Koeffizientenanzahl ($H_3$) sowie bei Mehrfachautorenschaft ($H_5$) eine Reduktion der Wahrscheinlichkeit eines knapp signifikanten Ergebnisses um knapp 15 Prozentpunkte beobachten. Beide Ergebnisse sind jedoch, auch bedingt durch die geringe Fallzahl im kleinsten Caliper, nicht signifikant ($p = 0{,}167$ bzw. 0,223). Für die weiteren getesteten Variablen, die *Data Policy* der ZfS ($H_{4a}$), die Datenzugänglichkeit (gemessen durch die potenzielle öffentliche Verfügbarkeit des Datensatzes; $H_b$) sowie die Kontrollvariable *Hypothesenformulierung* lässt sich kein klarer Effekt feststellen ($p > 0{,}5$). Dies gilt auch aufgrund der über die einzelnen Caliper wechselnden Vorzeichen der Koeffizienten. Ebenso findet sich für die Berichtspflicht ($H_{4b}$) nur ein Nulleffekt (vgl. Tabelle 2-10).

[47] Multikollinearitätsprobleme sind jedoch nicht zu befürchten ($r < 0{,}45$).
[48] Die vorhergesagten Werte von beiden Modellen (Logit und LPM) korrelieren hoch ($r > 0{,}99$), es ist von unverzerrten Schätzern der Maximum-Likelihood basierten logistischen Regression auszugehen.

**Tabelle 2-5 Logistische Regression von knapp signifikanten Ergebnissen (OC statt UC) auf Randbedingungen (zum 5%-Signifikanzniveau, unterschiedliche Caliper-Breiten)**

| | Variablen | AME | *SE* | *p*-Wert |
|---|---|---|---|---|
| 3%-Caliper | **Anzahl Koeffizienten (log)** | -0,156 | 0,113 | 0,167 |
| | **Data Policy ZfS** | 0,034 | 0,149 | 0,819 |
| | **Datenzugang** | 0,071 | 0,144 | 0,621 |
| | **Mehrfachautorenschaft** | -0,150 | 0,123 | 0,223 |
| | **implizite Hypothesen** | -0,103 | 0,179 | 0,566 |
| | ***N*, Cluster, pseudo $R^2$** | 50 | 34 | 0,061 |
| 5%-Caliper | **Anzahl Koeffizienten (log)** | -0,112 | 0,095 | 0,239 |
| | **Data Policy ZfS** | -0,073 | 0,123 | 0,557 |
| | **Datenzugang** | 0,056 | 0,146 | 0,702 |
| | **Mehrfachautorenschaft** | -0,127 | 0,129 | 0,323 |
| | **implizite Hypothesen** | 0,033 | 0,141 | 0,812 |
| | ***N*, Cluster, pseudo $R^2$** | 71 | 43 | 0,045 |
| 10%-Caliper | **Anzahl Koeffizienten (log)** | -0,073 | 0,055 | 0,188 |
| | **Data Policy ZfS** | -0,138 | 0,082 | 0,093 |
| | **Datenzugang** | -0,016 | 0,084 | 0,845 |
| | **Mehrfachautorenschaft** | -0,065 | 0,076 | 0,391 |
| | **implizite Hypothesen** | -0,032 | 0,091 | 0,722 |
| | ***N*, Cluster, pseudo $R^2$** | 133 | 58 | 0,027 |
| 15%-Caliper | **Anzahl Koeffizienten (log)** | -0,041 | 0,053 | 0,433 |
| | **Data Policy ZfS** | 0,003 | 0,075 | 0,970 |
| | **Datenzugang** | -0,039 | 0,083 | 0,639 |
| | **Mehrfachautorenschaft** | -0,073 | 0,074 | 0,327 |
| | **implizite Hypothesen** | -0,063 | 0,082 | 0,442 |
| | ***N*, Cluster, pseudo $R^2$** | 217 | 73 | 0,010 |

*Anmerkungen: N* = Anzahl Testwerte; Cluster = Anzahl Artikel. Alle *p*-Werte stammen von zweiseitigen Tests.

Im 5%-Caliper (s. ebenfalls Tabelle 2-5) bleiben die Mehrfachautorenschaft sowie die Anzahl der Koeffizienten in ihrem jeweils negativen Vorzeichen stabil, beide Effekte nehmen jedoch mit zunehmendem Caliper in ihrer Stärke ab. Einzig der Koeffizient der *Data Policy* erreicht im 10%-Caliper einen schwach signifikanten, negativen Effekt. Dieses Ergebnis ist jedoch vor dem Hintergrund des wechselnden Vorzeichens in den anderen Calipern mit Vorsicht zu interpretieren. Das pseudo-$R^2$ nach McFadden als Modellgütemaß ist in allen Calipern als eher gering einzustufen ($pR^2 < 0{,}061$). Der Rückgang der Effekte mit zunehmender Breite der Caliper ist hierbei konsistent mit der Annahme, dass insbesondere in engen Calipern ein PB nachzuweisen ist.

## 2.7. Diskussion

Der vorliegende Beitrag hatte das Ziel, das Phänomen des PB in deutschsprachigen Soziologie-Zeitschriften zu untersuchen und mittels einer detaillierten Erfassung von Kontextmerkmalen mögliche Ursachen zu identifizieren.

Ähnlich wie in US-amerikanischen Zeitschriften finden sich auch in der deutschen Soziologie auf der Datengrundlage von elf Jahrgängen der KZfSS und ZfS (2000-2010) Anzeichen für einen PB. So lässt sich um das am weitesten verbreitete 5%-Signifikanzniveau eine deutlich asymmetrische Verteilung der herangezogenen Testwerte feststellen. Die Ergebnisse des CT sprechen insgesamt dafür, dass ein Hintrimmen signifikanter Ergebnisse stattfindet, denn es gibt eigentlich keine Alternativerklärungen für die beobachteten Muster der Teststatistiken.

Die deutschsprachige Soziologie scheint allerdings in geringerem Ausmaß durch einen PB betroffen zu sein als die US-amerikanische Forschung (Gerber & Malhotra 2008a,b; 2010). Die Diskrepanz zwischen den deutschsprachigen und den US-amerikanischen Zeitschriften fällt so groß aus, dass sie sicher nicht durch die minimalen Differenzen in der Methodik bedingt ist. Der größere Wettbewerbsdruck in der internationalen Forschung scheint also das Auftreten eines PB zu unterstützen.[49]

Im Mittelpunkt der Ursachenanalyse standen die Manipulations- und Kontrollmöglichkeiten seitens der Forschenden. Mit steigender Komplexität der statistischen Modelle verringert sich das PB-Risiko. Ebenfalls ist eine Tendenz feststellbar, wonach Mehrfachautorenschaften das PB-Risiko reduzieren. In beiden Fällen ist der Effekt nicht im statistischen Sinn signifikant, wichtig ist aber: die Effektstärken sind substanziell (im kleinsten, dem 3%-Caliper, jeweils Marginaleffekte von ca. 15 Prozentpunkten). Gerade diese Aspekte verdienen daher eine wiederholte Betrachtung mit höherer Fallzahl oder anderen Verfahren, die den Einschluss weiterer Studien erlauben.

Kein Einfluss findet sich dagegen für die Zugänglichkeit der Daten (operationalisiert über die Verfügbarkeit der genutzten Daten in Datenarchiven und die *Data Policy* der ZfS, wonach Datensätze und Analysedateien auf Anfrage für Replikationen zur Verfügung gestellt werden müssen; sowie die in der ZfS seit 2009 umgesetzten umfangreicheren Berichtspflichten zu Analysen). Dies könnte in der verwendeten sehr weiten Definition von Datenzugänglichkeit sowie an der völligen Sanktionslosigkeit bei Verstoß gegen die *Data Policy* begründet sein. Trotz Verpflichtungserklärung zur *Data Policy* sind Autoren häufig nicht bereit, Daten für Re-Analysen zur Verfügung zu stellen (Brüderl 2013). Ob die stärkere Sanktion solcher Verstöße, etwa in Form eines *black-listing* wie von Brüderl (2013) gefordert, ihre Wirkung zeigt, bleibt künftigen Analysen vorbehalten. Die transparentere Darstellung der Ergebnisse ist jedenfalls noch zu unverbindlich. So werden immer noch Studien veröffentlicht, bei denen wichtige

[49] Darüber hinaus könnten auch die deutlich strengeren Standards in den US-amerikanischen Zeitschriften ihren Beitrag leisten: So sind dort die zu verwendenden Signifikanzniveaus stärker normiert. Im Falle des Fehlens solcher Standards erscheint es hingegen für Autoren sehr einfach, nur das Signifikanzniveau anzuheben und so „signifikante" Ergebnisse zu erreichen, ohne noch Ergebnisse im Sinne des hier untersuchten PB hintrimmen zu müssen.

methodische Details oder statistische Kennwerte wie Standardfehler oder die für den CT erforderlichen exakten Testwerte (*t*-/z- oder *p*-Werte) nicht berichtet werden. Damit sind die Hürden für Replikationen immer noch sehr hoch gesetzt und zudem ist das Entdeckungsrisiko eines den PB unterstützenden Fehlverhaltens gering.

Trotz der großen Anzahl der erfassten Testwerte leidet die vorliegende Analyse, wenn auch in geringerem Maße als die Vorläuferstudien, unter dem Problem der geringen statistischen Power aufgrund der wenigen in die Caliper eingeschlossenen Testwerte. So verbleiben von 108 verwendbaren Artikeln nur 50 Werte im 3%-Caliper des 5%-Signifikanzniveaus (von ursprünglich 1618). In der vorliegenden Arbeit wurde versucht, die Power des Testverfahrens durch methodische Weiterentwicklungen, wie der Heranziehung von *t*- statt *z*-Werten als genauerer Berechnungsgrundlage und der Beachtung von möglichen Rundungsungenauigkeiten, zu steigern. Für weitergehende Analysen wäre noch wichtiger, die Fallzahlen und zugleich die Abdeckung zu erhöhen, indem von Zeitschriften konsequent Angaben zu den statistischen Testwerten eingefordert werden.[50] So konnten von den quantitativen Arbeiten nur gut ein Drittel der Artikel untersucht werden. Das Ausmaß des PB wird deshalb vermutlich sogar unterschätzt.

Aus diesen Gründen sollte die Forschung zu den Ursachen des PB trotz der hier mehrheitlich berichteten Nullergebnisse fortgeführt werden. Insbesondere müssten die Anreizstrukturen der Autoren noch genauer untersucht werden. Etwa interessiert, ob ein PB mit kritischen Karrierestadien einhergeht. Auch im Hinblick auf ein besseres Verständnis des Teameinflusses wäre es sinnvoll den Karrierestatus von (Ko-)Autoren mit einzubeziehen. Im Zuge weiterer Datenerhebungen wäre zudem zu prüfen, ob Mehrfachautorenschaften im Falle dyadischer und triadischer (oder noch größerer) Teamkonstellationen eine unterschiedliche soziale Kontrollwirkung entfalten. Dies war in den vorliegenden Daten aufgrund des seltenen Vorkommens von mehr als zwei Autoren nicht möglich. Auch wäre eine Verknüpfung mit dem Replikationsexperiment von Brüderl (2013) ertragreich. Sind Akteure, die einer Replikation ihrer Forschungsergebnisse offen gegenüberstehen, in ihrer Arbeit präziser und daher weniger von einem PB betroffen (Feigenbaum & Levy 1993)?

Was lässt sich aus den Ergebnissen und theoretischen Überlegungen für Interventionen jenseits weiterer Ursachenforschung ableiten? Zunächst ist die Diagnose, dass ein PB unter anderem durch entsprechende Manipulationen der Autoren zustande kommt, nicht mit einem moralisch verstandenen Tadel gleichzusetzen. Mit dem CT ist ohnehin nur ein PB-Nachweis auf Aggregatebene und nicht für den Einzelfall möglich, in dem es auch bei ordnungsgemäßen Analysen selbstverständlich hin und wieder zu einem gerade noch signifikanten Ergebnis kommen kann. Zudem legt der hier vorlegte Theorieteil nahe, dass im Prozess des wissenschaftlichen Veröffentlichens ein soziales Dilemma (Dawes 1980; Kollock 1998)

[50] Der Abdruck von lediglich Signifikanzsternchen ist nicht nur in Bezug auf den CT ein Informationsverlust; aufgrund der willkürlichen Signifikanzschwellen wäre es *per se* weitaus informativer, (zusätzlich) die genauen Signifikanzwerte (oder damit assoziierte Testwerte, wie *t*-Statistiken oder Standardfehler) zu berichten.

besteht, aus dem einzelne rationale Akteure nicht einfach ausscheren können, ohne damit Wettbewerbsnachteile hinzunehmen. Wirkungsvoll erscheinen nur Veränderungen der allgemeinen Anreizstrukturen wissenschaftlichen Veröffentlichens, etwa durch die angesprochenen glaubwürdigeren Sanktionen, die Erhöhung der Entdeckungswahrscheinlichkeit von Fehlverhalten sowie durch die stärkere Anerkennung von soliden Replikationsstudien. Wissenschaftsorganisationen, welche die sehr knappen und umkämpften Belohnungen vergeben, also etwa Institutionen der Forschungsförderung oder angesehene Zeitschriften, könnten beispielsweise kontinuierlich eine Zufallsauswahl von Artikeln nachrechnen lassen (Baerlocher et al. 2010: 44; Diekmann 2005: 26).[51] Die dazu nötige technische Infrastruktur wäre mit relativ geringen Mitteln durch diejenigen Akteure, die an der Qualitätssicherung großes Interesse haben müssen, wie die Deutsche Forschungsgemeinschaft (DFG) oder im Fall der Sozialwissenschaft die GESIS, bereitzustellen. Weiterhin müsste der Review-Prozess selbst zum Gegenstand der Reform werden. Etwa sollte überlegt werden, wie sorgfältige und belastbare Reviews stärker belohnt werden können. Denn eines erscheint sicher: Die Selbstheilungskräfte der Wissenschaft scheinen derzeit nicht auszureichen, um einen PB zu verhindern (Stroebe et al. 2012).

[51] Diekmann vergleicht diese Strategie sehr anschaulich mit Kontrolleuren im ÖPNV, deren Kontrollen das Risiko von Schwarzfahrern deutlich reduzieren (2005: 27). Die diesem Artikel zugrunde liegenden Daten und Analysefiles stehen auf folgender Webseite zur Verfügung: http://www.uni-koeln.de/kzfss/materialien/KS-66-4-Auspurg.zip.

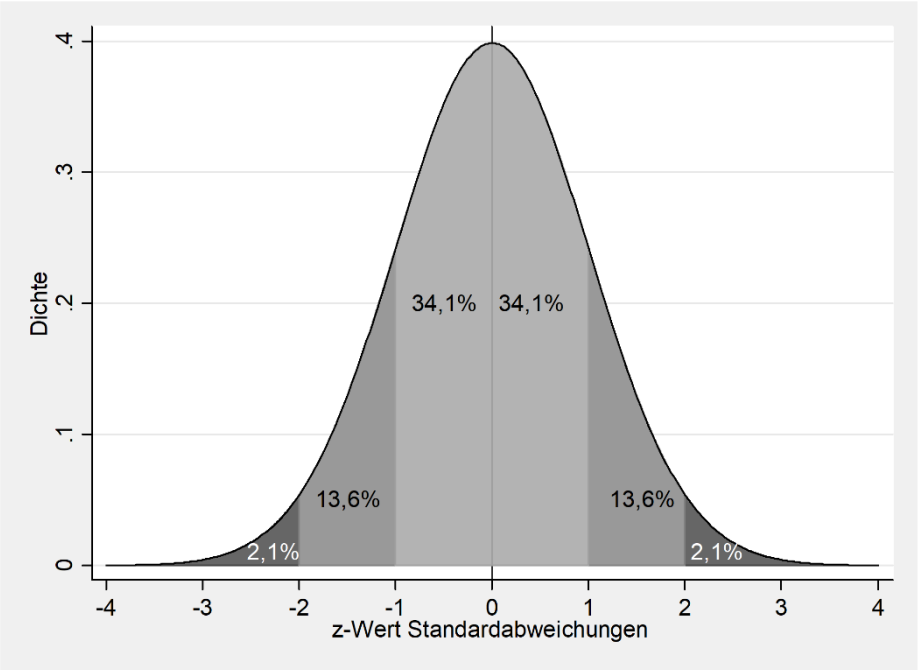
## 2.8. Anhang

### *2.8.1. Methoden & Kodierung*

#### *Logik des Caliper-Tests*

Die Verteilung von *z*- und ebenso *t*-Werten ist bekanntlich glockenförmig, mit einem Maximum bei der wahren Effektstärke (somit bei Gültigkeit der Nullhypothese bei dem Wert Null). Nur betrachtet an diesem einen Maximum ist die Wahrscheinlichkeit von Testwerten, über oder unter dem Wert zu liegen, exakt gleichverteilt; an allen anderen Stellen der Verteilung sind dagegen die Wahrscheinlichkeiten ungleich verteilt – schließlich liegen Stichprobenwerte *immer* wahrscheinlicher nahe am wahren Wert als weiter davon entfernt. Gerade darauf basiert die gesamte Idee des statistischen Testens. Anhand von Abbildung 2-4 lässt sich dies nochmals veranschaulichen, so ist die Wahrscheinlichkeit einen *z*-Wert größer zwei zu erreichen (bei wahrem Nulleffekt, also $z = 0$) nur 4,2% (2*2,1%). Die Annahme einer Gleichverteilung lässt sich in Anbetracht der zum Teil sehr schiefen Verteilung nur für einen sehr kleinen Ausschnitt der *z*- oder *t*-Werte-Verteilung näherungsweise rechtfertigen; bei Betrachtung größerer Abschnitte ist aufgrund der schiefen Verteilung dagegen auch ohne einen PB schon „naturgemäß“ eine Überrepräsentation von Werten über oder unter der betrachteten Schwelle, hier dem Signifikanzniveau, zu erwarten.

**Abbildung 2-4 Dichtefunktion von *z*-Werten unter Gültigkeit der Nullhypothese**




Im vorliegenden Beispiel bedeutet dies konkret, dass mit einer Bandbreite von 0,2 (vergleichbar mit dem 10%-Caliper um das 5%-Signifikanzniveau) um Null eine exakte Gleichverteilung zu erwarten ist, diese wird jedoch bei höheren bzw. niedrigeren Werten immer schiefer. Ist die Wahrscheinlichkeit bei 1 noch 1,22:1 und damit annähernd gleichverteilt, so ist selbst in diesem kleinen Ausschnitt mit einem

z-Wert von 3 die Chance bereits 1,63:1. Im Falle der Betrags-Dichtefunktion der *z*-Werte sind somit niedrigere, und damit nicht-signifikante *z*-Werte unter Geltung der Nullhypothese immer leicht überrepräsentiert. Die Annahme der Gleichverteilung in den Calipern ($E(x_z) = 0{,}5$) um die Signifikanzschwellen ist daher eher konservativ.

Da im vorliegenden Fall jedoch verschiedene getestete Effekte der deutschen Soziologie (heterogene Effekte) zusammengefasst sind, die naturgemäß verschiedene wahre Werte haben, ist keine „natürliche" Verteilung der *z*-Werte ableitbar. Zwar streuen alle getesteten Effekte nach dem oben beschriebenen Muster, jedoch unterscheiden sich die wahren Effekte und damit die Maxima der Verteilungen.

Der CT erlaubt es die angenommene Gleichverteilung um einen Signifikanzschwellenwert mithilfe eines Binomialtests zu testen. Für den CT gibt eine Indikatorvariable $\mathrm{x_z}$ diese Verteilung der Testwerte wieder: *z*-Werte, die größer als das Signifikanzniveau (s) sind fallen bis zur Obergrenze $(\mathrm{s} + \mathrm{cs})$ in den Over-Caliper (OC) ($\mathrm{x}_z = 1$). Werte kleiner gleich s fallen bis zur Untergrenze $(\mathrm{s} - \mathrm{cs})$ in den Under-Caliper (UC) ($\mathrm{x}_z = 0$). Werte, die nicht in diesen kleinen Abschnitt der *z*-Werte-Verteilung liegen fallen aus der Analyse. Der Faktor c ist hierbei der prozentuale Caliper und definiert die Breite des untersuchten Intervalls relativ zu s:

$$\mathrm{x}_z = \begin{cases} 0 \text{ falls } \mathrm{s} - \mathrm{cs} < z \leq \mathrm{s} \\ 1 \text{ falls } \mathrm{s} < z < \mathrm{s} + \mathrm{cs} \end{cases}$$

Am Beispiel des 5%-Calipers um das 5%-Signifikanzniveau (s = 1,96) fallen somit alle Werte größer 1,862 (1,96 - 0,098) bis einschließlich 1,96 in den UC ($\mathrm{x}_z = 0$) und Werte größer 1,96 bis 2,058 (1,96 + 0,098) in den OC ($\mathrm{x}_z = 1$).

Der im Rahmen dieser Studie weiterentwickelte CT berücksichtigt dabei auch die Freiheitsgradabhängigkeit von Signifikanzwerten. Im Falle von kleinen Studien ist somit ein höherer *t*-Wert vonnöten um statistische Signifikanz zu erreichen. Es gibt folglich keine feste Signifikanzschwelle s, sondern eine freiheitsgradabhängige Signifikanzschwelle $\mathrm{s_{df}}$. Dies wirkt sich ebenso auf die Kodierung der Caliper aus. Jeder Koeffizient hat demnach einen von den Freiheitsgraden abhängigen Caliper. Diese höhere Präzision scheint insbesondere im Hinblick auf die größere Instabilität statistischer Modelle bei geringen Freiheitsgraden sinnvoll, bei denen sich durch kleine Modellveränderungen oft schon große Veränderungen in den Signifikanzwerten ergeben.

$$\mathrm{x}_t = \begin{cases} 0 \text{ falls } \mathrm{s_{df}} - \mathrm{cs_{df}} < \mathrm{t} \leq \mathrm{s_{df}} \\ 1 \text{ falls } \mathrm{s_{df}} < \mathrm{t} < \mathrm{s_{df}} + \mathrm{cs_{df}} \end{cases}$$

***Kodierung der Art von Autoren geprüfter Hypothesen***

Die Art der Hypothesenformulierung wurde als Kontrollvariable wie folgt kodiert:

> *Explizite Hypothesen:* Werden direkt im Text benannt und als Hypothesen ausgeflaggt.
>
> Beispiel: „H4: Je mehr in die bisherige Schullaufbahn investiert wurde, desto geringer ist die Schulverweigerung" (Wagner et al. 2004: 461).

*Implizite Hypothesen:* Werden nicht explizit ausgeflaggt und stellen nur eine ungefähre Erwartung an die Effektrichtung dar.

Beispiel: „Dem Ansatz der Humankapitaltheorie folgend müssten am Beginn der Erwerbsphase hohe betriebliche und berufliche Aus- und Weiterbildungsanforderungen die individuellen Einkommenschancen negativ beeinflussen, weil hier zwischen Arbeitnehmer und Arbeitgeber ein Austausch von Qualifizierungsleistung und Einkommensverzicht vorgenommen wird." (Liebeskind 2004: 633).

### *2.8.2. Robustness Checks*

***Ergebnisse des CT mit z-Werten und imputierten Nachkommastellen***

Im Folgenden werden, zusätzlich zu den im Artikel getesteten Calipern basierend auf *t*-Werten, auch die von Gerber & Malhotra (2008b) verwendeten freiheitsgradunabhängigen *z*-Werte berichtet. Darüber hinaus sollen zwei weitere CT aufbauend auf imputierten Nachkommastellen der *t*- und *z*-Werte berichtet werden. Tabelle 2-6 stellt den CT basierend auf *z*-Werten dar und zeigt eine geringfügig höhere Prävalenz von knapp signifikanten Werten als der im Artikel berichtete CT an. Die leicht höher geschätzte PB-Prävalenz im Falle von *z*-Werten zeigt sich auch im Vergleich von Tabelle 2-7 sowie

Tabelle 2-8. Die Verwendung der modelladäquateren *t*-Werte stellt daher eine konservativere Methode dar auf PB zu testen. Die Imputation von Nachkommastellen hingegen verändert das Ergebnis nur minimal. Dies stützt auch die Annahme einer stetigen Testwerteverteilung.

**Tabelle 2-6 *z*-Werte zum 5%-Signifikanzniveau**

| | *N* | OC | UC | % OC | *p*-Wert [a] | k. *p*-Wert [a] |
|---|---|---|---|---|---|---|
| 3%-Caliper | 53 | 36 | 17 | 0,679 | 0,006 | 0,025 |
| 5%-Caliper | 71 | 46 | 25 | 0,648 | 0,008 | 0,025 |
| 10%-Caliper | 134 | 78 | 56 | 0,582 | 0,035 | 0,069 |
| 15%-Caliper | 215 | 114 | 101 | 0,530 | 0,207 | 0,207 |

[a] einseitiger *p*-Wert

**Tabelle 2-7** ***t*-Werte mit imputierten Nachkommastellen zum 5%-Signifikanzniveau**

| | *N* | OC | UC | % OC | *p*-Wert [a] | k. *p*-Wert [a] |
|---|---|---|---|---|---|---|
| 3%-Caliper | 49 | 32 | 17 | 0,653 | 0,022 | 0,089 |
| 5%-Caliper | 72 | 44 | 28 | 0,611 | 0,038 | 0,115 |
| 10%-Caliper | 131 | 75 | 56 | 0,573 | 0,058 | 0,115 |
| 15%-Caliper | 214 | 115 | 99 | 0,537 | 0,153 | 0,153 |

[a] einseitiger *p*-Wert

**Tabelle 2-8** ***z*-Werte mit imputierten Nachkommastellen zum 5%-Signifikanzniveau**

| | *N* | OC | UC | % OC | *p*-Wert [a] | k. *p*-Wert [a] |
|---|---|---|---|---|---|---|
| 3%-Caliper | 51 | 37 | 14 | 0,725 | 0,001 | 0,004 |
| 5%-Caliper | 73 | 47 | 26 | 0,644 | 0,009 | 0,028 |
| 10%-Caliper | 132 | 79 | 53 | 0,598 | 0,015 | 0,029 |
| 15%-Caliper | 210 | 115 | 95 | 0,548 | 0,095 | 0,095 |

[a] einseitiger *p*-Wert

***Vergleich des PB-Risikos zwischen internationaler und deutscher Soziologie***

Tabelle 2-9 stellt die unterschiedliche Prävalenz des PB in der deutschen und US-amerikanischen Soziologie dar ($H_2$). Um einen Vergleich der CT zu ermöglichen wurde, um nicht methodische Artefakte zu berichten, ebenso wie im Vergleichsartikel von Gerber & Malhotra (2008a) der CT auf Basis von *z*-Werten (vgl. Tabelle 2-6) berechnet und zum Vergleich herangezogen.

**Tabelle 2-9 Vergleich der PB-Risiken mit US-amerikanischen Zeitschriften**

| | **Deutsche Zeitschriften (Eigene Auswertungen)** | | **US-Zeitschriften** (Gerber & Malhotra 2008a) | | **Test auf signifikante Unterschiede (zwei Stichproben z-Test)** | | |
|---|---|---|---|---|---|---|---|
| | **% OC** | ***N*** | **% OC** | ***N*** | ***z*** | ***p*-Wert** [a] | **k. p-Wert** [a] |
| **5%-Caliper** | 0,648 | 71 | 0,786 | 56 | 1,697 | 0,045 | 0,045 |
| **10%-Caliper** | 0,582 | 134 | 0,689 | 106 | 1,698 | 0,045 | 0,090 |
| **15%-Caliper** | 0,530 | 217 | 0,731 | 156 | 3,924 | 0,000 | 0,000 |

[a] einseitiger *p*-Wert

***Weitere multivariate Ergebnisse***

Das nachfolgende Modell (Tabelle 2-10) berichtet die Ergebnisse der Modellspezifikation mit der Variable Berichtspflicht ZfS ($H_{4a}$) alternativ zur *Data Policy* der ZfS.

**Tabelle 2-10 Logistische Regression von knapp signifikanten Ergebnissen (OC statt UC) auf Randbedingungen (zum 5%-Signifikanzniveau, unterschiedliche Caliper-Breiten) – mit Berichtspflicht**

| | Variablen | AME | *SE* | *p*-Wert |
|---|---|---|---|---|
| 3%-Caliper | **Anzahl Koeffizienten (log)** | -0,183 | 0,119 | 0,124 |
| | **Berichtspflicht ZfS** | -0,081 | 0,186 | 0,665 |
| | **Datenzugang** | 0,072 | 0,143 | 0,614 |
| | **Mehrfachautorenschaft** | -0,137 | 0,131 | 0,297 |
| | **implizite Hypothesen** | -0,107 | 0,164 | 0,515 |
| | ***N*, Cluster, pseudo $R^2$** | 50 | 34 | 0,063 |
| 5%-Caliper | **Anzahl Koeffizienten (log)** | -0,110 | 0,097 | 0,256 |
| | **Berichtspflicht ZfS** | -0,035 | 0,170 | 0,836 |
| | **Datenzugang** | 0,057 | 0,141 | 0,687 |
| | **Mehrfachautorenschaft** | -0,142 | 0,129 | 0,270 |
| | **implizite Hypothesen** | 0,019 | 0,138 | 0,893 |
| | ***N*, Cluster, pseudo $R^2$** | 71 | 43 | 0,042 |
| 10%-Caliper | **Anzahl Koeffizienten (log)** | -0,071 | 0,062 | 0,251 |
| | **Berichtspflicht ZfS** | -0,055 | 0,094 | 0,564 |
| | **Datenzugang** | -0,025 | 0,082 | 0,758 |
| | **Mehrfachautorenschaft** | -0,092 | 0,080 | 0,247 |
| | **implizite Hypothesen** | -0,052 | 0,094 | 0,577 |
| | ***N*, Cluster, pseudo $R^2$** | 133 | 58 | 0,014 |
| 15%-Caliper | **Anzahl Koeffizienten (log)** | -0,048 | 0,053 | 0,368 |
| | **Berichtspflicht ZfS** | -0,043 | 0,076 | 0,571 |
| | **Datenzugang** | -0,031 | 0,082 | 0,701 |
| | **Mehrfachautorenschaft** | -0,066 | 0,079 | 0,400 |
| | **implizite Hypothesen** | -0,066 | 0,084 | 0,428 |
| | ***N*, Cluster, pseudo $R^2$** | 217 | 73 | 0,011 |

*Anmerkungen: N* = Anzahl Testwerte; Cluster = Anzahl Artikel. Alle *p*-Werte stammen von zweiseitigen Tests.

# 3. Examining publication bias – A simulation-based evaluation of statistical tests on publication bias

A. Schneck. Erschienen in: PeerJ 2017 (5:e4115). https://doi.org/10.7717/peerj.4115

## 3.1. Introduction

All scientific disciplines try to uncover truth by systematically examining their surrounding environment (Descartes 2006: 17). Natural scientists try to observe regularities in nature, whereas social scientists try to uncover patterns in the social behaviour of humans. The success, as well as the reputation of science rests on the accuracy and unbiasedness of scientific results. Publication bias, the publication of only positive results confirming the researcher's hypothesis (cf. Dickersin & Min 1993: 135), threatens this validity. Under publication bias only results showing either statistical significance and/or the desired direction of the effects are published. The published literature in this case is merely a selective (and too optimistic) part of all existing scientific knowledge. Furthermore, science is in the case of publication bias also inefficient as studies that add substantial knowledge to the literature, but contain null-findings remain unpublished.

The study at hand examines the performance of four methods to identify publication bias: Egger's Test/FAT (Egger et al. 1997; Stanley & Doucouliagos 2014), p-uniform (PU; van Assen et al. 2015; van Aert et al. 2016), the test for excess significance (TES; Ioannidis & Trikalinos 2007) and the caliper test (CT; Gerber & Malhotra 2008a,b). In order to compare the performance of these tests, the false positive rate ($\alpha$-error, type I error) and the statistical power (true positive rate) were examined in a Monte Carlo simulation study. This makes it possible to assess the performance of the four tests under different conditions of publication bias (*file-drawer* vs. *p-hacking*), as well as study settings (underlying true effect, effect heterogeneity, number of observations in primary studies and in meta-analyses).

### *3.1.1. The issue of publication bias*

The false positive rate of a test (commonly called *p*-value) is the probability of the estimator rejecting $H_0$ despite this being true. The *p*-value is therefore the probability that the observed estimate is at least as extreme given there is no effect as assumed by $H_0$ (Wasserstein & Lazar 2016). The larger the *p*-value the higher the risk of assuming an effect if none exists in the data. *p*-values below a certain threshold are called statistically significant, whereas values above the threshold are labelled as non-significant. In the empirical sciences the 5%-significance threshold is mostly used (Cohen 1994; Labovitz 1972; Nuzzo 2014). The difference between 0.049 and 0.051 in the error probability is, however, marginal. Nevertheless, from the standpoint of the 5%-significance threshold the first would be a significant effect, whereas the latter would be a non-significant effect. In both of these two cases on average around 1 in 20 null-hypotheses of no difference would be rejected, albeit true. If empirical researchers select their data/models until they find, just by chance, significant evidence that seems worth publishing, publication bias is on the rise, leading to inflated or even artificial effects.

Rosenthal (1979) constructs a worst case scenario in which only the 5% of false positive studies that are "significant" solely by pure chance are published. In this case, misinterpreted results shape the scientific discourse and finally result in (medical or political) interventions. Although Rosenthal's example is extreme, a multitude of evidence for publication bias exists in various disciplines and research fields (e.g. Doucouliagos & Stanley 2009; Jefferson et al. 2012). Godlee (2012) therefore warns that scientific misconduct, under which publication bias is subsumed (Chalmers 1990), may also physically harms patients.

In addition to the societal consequences, publication bias also has severe implications for the evolution of knowledge. Under publication bias no rejection of theories (Popper 1968: 215), on which all scientific progress relies, occur which leads to a state of "undead theory" (Ferguson & Heene 2012: 559) where all existing theories are confirmed irrespective of their truth.

### *3.1.2. Motivation to commit publication bias*

Because statistically significant results stress the originality of research findings (Merton 1957), Both, authors and scientific journals (cf. Coursol & Wagner 1986; Epstein 1990,2004; Mahoney 1977) have large incentives to maximise their significant results to survive in a publish or perish research environment. Authors especially want to increase their publication chances, notably in top-tier journals where low acceptance rates of 5%-10% are quite common (cf. for the political sciences Yoder & Bramlett 2011: 266; for the top interdisciplinary journals Nature 2017; Science 2017). Two distinct strategies to achieve significant results by means of publication bias practices can be pointed out. Firstly, non-significant findings can be suppressed (cf. the classical *file-drawer* effect described by Rosenthal 1979) and significant results are then searched for in another dataset. Secondly, small bits in the data analysis can be changed (e.g. adding covariates, optional stopping, exclusion of outliers, etc.) until a significant result is obtained – this method is known as *p-hacking* (or "researchers degree of freedom" Simmons et al. 2011: 1359; cf. "fishing" Gelman 2013). Whereas the *file-drawer* strategy can be utilised by authors as well as by editors and reviewers, *p-hacking* can only be committed by authors/researchers. Nonetheless, *p-hacking* strategies can be recommended by actors other than authors (e.g. editors, reviewers, etc.). Especially *p-hacking* is almost without any costs, as data analysis tools/packages become increasingly easy to apply (Paldam 2013).

### *3.1.3. Evidence on the prevalence of publication bias*

So far, there are two strategies for identifying publication bias: the first traces studies through the publication process, the second asks authors, reviewers, or editors about their publication practices via surveys. In the first strategy, most of the analyses trace conference papers or ethics committee decisions if those results get published or remain in the file-drawer. Overall previous findings note, that studies with significant results have a substantially higher chance to get published (cf. Callaham et al. 1998; Coursol & Wagner 1986; Dickersin 1990; Easterbrook et al. 1991). Ioannidis (1998) in addition finds that significant studies have, beside their higher publication rate, also a substantially higher publication speed,

meaning a shorter time between the completion of the study and the final publication. This results suggest, that publication bias is a beneficial strategy in order to maximize academic merits.

The second approach asks directly about the publication practices of the involved actors. In a survey of psychologists that used a sensitive question technique up to 50% of the respondents claimed that they exercised publication bias (John et al. 2012: 525). Franco et al. (2014) also note that most non-significant findings go to the file-drawer right after the analysis and are not even written up. Also, other forms of misbehaviour, like optional stopping (stopping data collection when significance is reached) or erroneous rounding of *p*-values to reach significant results, are alarmingly widespread (prevalence rate arround 22.5% John et al. 2012: 525). These results are in line with the survey of Ulrich & Miller (2017: 9), who report that researchers in the field of psychology prefer significant results over non-significant results, and, furthermore, attribute more value to results with smaller *p*-values. These estimates may even be conservative because it is known from the survey literature that sensitive behaviours like scientific misconduct may be underreported (Kreuter et al. 2008: 848). According to the presented research results *file-drawer* and *p-hacking* behaviour is therefore quite widespread.

## 3.2. Methods

### *3.2.1. Publication bias tests in comparison*

So far, the presented detection strategies ask either directly for publication preferences or examine the publication fate of conference papers. Both approaches have the weakness that they either rely on the potentially biased answers of the actors involved or require an immense effort to follow the publication process, while publication bias may have happened before the paper is submitted to a conference. Statistical tests on publication bias circumvent this problem by relying only on the published literature. In the paper at hand the regression-based FAT (Egger et al. 1997; Stanley & Doucouliagos 2014), PU (van Aert et al. 2016; van Assen et al. 2015), an extended version of p-curve (Simonsohn et al. 2014a,b; Simonsohn et al. 2015), the TES (Ioannidis & Trikalinos 2007), and the CT (Gerber & Malhotra 2008a,b) were evaluated (see appendix for an in-depth discussion of the tests).[52]

In order to compare the different publication bias tests, four different criteria have to be established: the assumptions of the test, the measurement level, the sample used, the test method, and its according limitations (see Table 3-1).

[52] Because for Fail-save-N (Rosenthal 1979) only rules of thumbs, instead of a formal statistical test, exist it was not included in the simulation at hand. Although it is still widely applied (Banks et al. 2012: 183; Ferguson & Brannick 2012: 4), this benchmark is not recommended in the *Cochrane Handbook*, a guideline for conducting meta-analyses (Higgins & Green 2008: 321f.).

**Table 3-1 Publication bias tests in comparison**

| Test | Measurement level | Sample | Assumption | Limitation |
|---|---|---|---|---|
| FAT | Continuous [-∞,∞] | All | Cov(es, se) = 0 | Only one-sided publication bias (PB) detectable |
| PU | Continuous [0,1] | $p < 0.05$, effects of same sign | Uniform or right skewed<br>Skewness >= 0 | Only one-sided PB detectable<br>Only on prespecified levels<br>Effect homogeneity (fixed-effect meta-analysis) |
| TES | Dichotomous [0,1] | All | E = O | Only on prespecified levels<br>Effect homogeneity (fixed-effect meta-analysis) |
| CT | Dichotomous [0,1] | Threshold ± caliper width | P(UC) = P(OC) | Only on prespecified levels |

*Note: Table 3-1 compares the four evaluated publication bias tests in respect to four criteria, the measurement level, the sample used by the test, its underlying assumptions and its limitations*

The FAT tests basically the relationship between study's precision and its effect size with all available effect sizes from primary studies. If larger effects are observed for studies with low precision (and low *N*) publication bias is suspected. Nonetheless if alternative reasons may lead to this result: small studies examine specific high risk populations in which treatments may be more effective (Sterne et al. 2011), this effect heterogeneity may lead to the diagnosis of publication bias where none exists (Schwarzer et al. 2002).[53] The FAT has furthermore the disadvantage that only one-sided publication bias either in favour of a positive or negative significant effect can be tested. Alinaghi & Reed (2016: 10) show that if significant results of either sign are searched for, the FAT suffers from massively inflated false positive rates. In the Monte Carlo simulation at hand only the FAT is used because of its better statistical power as shown in prior simulations compared to the similar rank correlation test of Begg & Mazumdar (1994)[54] and the trim and fill technique (Duval & Tweedie 2000).[55]

PU has the assumption that every left skewness in the distribution of *p*-values smaller than the significance threshold (e.g. $p < 0.05$) and conditioned on the underlying observed mean effect (*pp*-value) distribution is caused by publication bias. This assumption is, however, grounded mainly on the fixed-effect estimate of the mean effect, which is very sensitive to effect heterogeneity. PU furthermore limits

[53] For a similar result see Terrin et al. (2003) for the related Trim and Fill technique (Duval & Tweedie 2000).
[54] For simulations see: Hayashino et al. (2005); Kicinski (2014); Macaskill et al. (2001); Sterne et al. (2000).
[55] For simulations see: Bürkner & Doebler (2014); Kicinski (2014); Moreno et al. (2009); Renkewitz & Keiner (2016).

its test value only on significant estimates in the direction where publication bias is suspected (van Aert et al. 2016: 727). Therefore PU is, as the FAT, only able to identify one-sided publication bias.

The TES (Ioannidis & Trikalinos 2007; also called ic-index see Schimmack 2012) in contrast relies only on a dichotomous classifier, testing if the number of expected significant results and the empirically observed number of significant effects differ. Because the TES relies, as PU on the fixed-effect estimate of the mean effect of all included studies it is sensible to effect heterogeneity. A large controversy in the literature is not about the TES itself, but on its application. Francis (2012a,b,c,d,e,2013) used the TES to identify singular articles in order to test if they suffer from publication bias. This may invalidate the assumption of independence (Morey 2013: 181) as well as inflate the false positive rate in a similar manner than in primary research (cp. HARKing Kerr 1998) if the TES is used in such an exploratory manner (Simonsohn 2013: 175). Ioannidis however responds that if the TES is applied on prespecified research questions with a large and independent number of effect sizes, the TES is even a conservative test on publication bias (Ioannidis 2013: 185).

The CT uses the most limited sample of the included tests that includes only estimates slightly over and under the chosen significance level in a distribution of *z*-values. In case of publication bias the assumption of a continuous distribution that results in an approximately even distribution in a narrow interval (caliper) is violated by an overrepresentation of just significant results. The broader the interval is set the more it may deviate from the assumed even distribution caused by the true underlying effect. This restrictive sample has the downside that the exclusion of most available values may drastically reduce the statistical power of the test.

In contrast to the FAT, the other tests are only able to test for publication bias on pre-specified levels (e.g. 0.05). Because the TES and the CT focus only on dichotomous classifiers (significant or not in the case of the TES, slightly over or under the threshold for the CT) also tests on two-sided publication bias are possible.

In previous simulation studies with a low number of included studies as well as observations PU was superior to the TES (van Assen et al. 2015: 303) and the FAT (Renkewitz & Keiner 2016). However no evaluations exists based on a larger number of primary studies. In particular, the newer publication bias tests like PU, the TES, and the CT, are in need of an evaluation under different conditions. For the CT also no studies exist regarding the best caliper width to use. Despite the existence of some simulation studies on publication bias tests, so far no direct comparison exists that evaluates the performance of all four publication bias tests, especially under effect heterogeneity.

### *3.2.2. Simulation setup*

In order to examine the performance of the four publication bias tests, a Monte Carlo simulation approach is used. For the simulation two different processes have to be distinguished: firstly, the data

generation process (DGP), and, secondly, the meta-analytical estimation method (EM). The DGP provides the ground for the hypothetical data used by the simulated actors, as well as the results they report, whereas the EM applies the tests on publication bias reported in the previous section. The central advantage of using Monte Carlo simulations is that controlling the DGP allows to identify which simulated studies suffer from publication bias and which do not. Similar to the case in experiments, different conditions can be defined to ensure a controlled setting. The performance of the estimators can then be examined under the different conditions.

***Data setup of the primary studies and meta-analyses***

The first step of the DGP defines different effect size conditions that underlie the analyses of the simulated actors (see Table 3-2). As a first condition the underlying true effect was specified by a linear relationship with $\beta = 0, 0.5, 1.0, 1.5$. Analogous to a linear regression model this means for $\beta = 0.5$, that an increase of one unit of the independent variable $x$ increases the dependent variable $y$ by 0.5. The specified linear relationship between the dependent variable $y$ and the independent variable $x$ had a normally distributed regression error term of $\varepsilon = N(0,10)$, while the variation of the independent variable was defined as $\sigma_x = 2$ (for a similar setup see Alinaghi & Reed 2016; Paldam 2015). The regression coefficients can also be transformed in the Pearson correlation coefficient yielding approximately $r = 0, 0.1, 0.2, 0.3$. This results are equivalent to low or medium effect sizes in terms of Cohen (1992) and cover about 75% of the empirical observed effects in psychology (Bosco et al. 2015: 436).[56] In addition to the homogenous conditions with a common effect size, a heterogeneous condition was added that assumes no fixed distribution of an underlying effect but a uniform mixture of all four effect sizes, as defined above, plus an additional effect of $\beta = 2.0$ ($r = 0.4$) in order to ensure enough variation.

As the FAT is based on study precision, which is mainly driven by the number of observations ($N$) of the primary studies, $N$ was computed as a second condition by an absolute normal distribution with a mean of 100 (small $N$) or 500 (large $N$) and a standard deviation of 150. In order to ensure an adequate statistical analysis for the primary studies, $N$s equal to or smaller than 30 were excluded. This procedure resulted in a right skewed distribution with a mean $N$ of roughly 500 for the large $N$, and 165 for the small $N$ condition. The small $N$ condition reflects the observed number of observations in leading economics journals (mean: 152, own computations from the publicly available dataset of Brodeur et al. 2016) as well as of typical trials included in Cochrane reviews (mean: 118, Mallett & Clarke 2002: 822). Because both studies refer mainly to an experimental literature, the large $N$ condition reflects the more common number of observations especially in ex-post-facto designs (e.g. survey studies).

[56] 16.6% of the simulated studies were adequately powered with at least 80% power (Cohen 1988: 56). The setting produced by the DGP also reflects the results of (Ioannidis et al. 2017: 245), who report that only 10% of the studies in economics are adequately powered.

**Table 3-2 Data generating process (DGP) of Monte Carlo simulation**

| **Conditions** | **Values** | **Functional form** | ***N* (conditions)** |
|---|---|---|---|
| **Data setup:** | | | |
| 1. True effects $\beta$: | $\beta$ = 0; 0.5; 1; 1.5; Het<br>$\varepsilon$ = NV(0,10)<br>$\sigma_x$ = 2 | y = $\beta$x + $\varepsilon$ | **5** |
| 2. Number of observations *N*: | $\mu_N$ = 100; 500 | \|N($\mu_N$)\| *N* > 30 | **2** |
| 3. Number of studies *K*: | *K* = 100; 1000 | | **2** |
| **Behavioural setup:** | | | |
| 4. Publication bias (PB) | *PB = 0; 0.5; 1.0* | $\beta$ > 0 & *p* < 0.05 | **1+2*2 = 5** |
| 4.1.*File-drawer* | Draw new sample size *N* | *(max. 9 additional samples)* | |
| 4.2.*p-hacking* | Run new analyses with same dataset | y = $\beta$x + $\gamma_j z_j$ + ε<br>z = 0.5x + 0.5y + ε<br>*(max. 3 z's = 7 combinations)* | |
| | | | **5*2*2*5 = 100** |

*Note: The 100 conditions of the Monte Carlo simulations are described. Two different aspects were varied: the underlying data and the publication bias behaviour of the actors. For the underlying data the true effect size, the number of observations (N) and the number of studies included in the meta-analysis (K) were varied. The behavioural component altered the proportions of authors who are willing to commit publication bias and its actual form as either p-hacking or file-drawer*

The heterogeneity of effects in each of the meta-analyses was measured by $I^2$, the share of systematic variation in respect to the overall variation consisting of the systematic and random variation (Higgins & Thompson 2002). In case of the small *N* 68.62% and for a large *N* 86.86% of the variation was systematic in the heterogeneous effect condition. In terms of Higgins & Thompson (2002: 1553), an $I^2$ larger than 50% has to be modelled explicitly in meta-analyses and cannot be ignored.

In addition to the number of observations in the primary studies (*N*) the number of primary studies that were included in the meta-analysis and form the basis of the publication bias tests (*K*) was varied in the third condition. A setting with 100 studies was used as a lower condition, whereas 1000 studies were set as an upper condition. Although on average the number of trials in a meta-analysis is usually much lower

than 100 studies (median 28 studies in the meta-meta-analysis by Elia et al. 2016: 5). 100 studies were chosen because in this setting every publication bias test evaluated is at least partially applicable. In other research areas like economics, where meta-regression models are more widely used to model effect heterogeneity, also higher numbers of included trial-estimates are quite common (e.g. 1474 effect estimates in Doucouliagos & Stanley 2009).

***Behavioural setup of publication bias***

Building on this data setup stage of the DGP the behavioural setup adds publication bias to the simulation in a fourth step (see Table 3-2). Publication bias was defined as the willingness to collect new data or run additional analyses if statistical significance failed ($p \geq 0.05$) or a negative effect occurred. In the simulation only one-sided publication bias was modelled because both, the FAT and PU are not able to model two-sided publication bias that focuses only on significant results irrespective of its sign. It is important to note, that only the intent to commit publication bias was varied in the simulation setup. The actual publication bias depends on the data setup itself: how large is the true effect size ($\beta$) and the number of observations ($N$) in the primary studies? Or, in short: is there already a significant positive result which does not need a publication bias treatment?

Five different publication bias conditions have to be distinguished. Firstly, the condition without publication bias: in this ideal case all estimates ($\beta x$) are estimated by a bivariate ordinary least squares (OLS) model and afterwards published. Publishing in terms of the simulation model means that all estimates enter the final meta-analysis. Therefore, in the condition without publication bias either 100 or 1000 regression results were estimated and enter the meta-analysis.

In the second and third conditions publication bias was present with a 50% probability. That means that 50% of the actors were willing to run additional analyses in order to obtain significant results. These conditions seem closest to the behavioural benchmark of the empirical studies presented.

If a non-significant result was obtained, actors operating under the second condition chose to collect new data in order to obtain significant results that can be published. This second condition therefore modelled publication bias under the *file-drawer* scenario, because the datasets not used remained unpublished. An actor tried to run analyses on the basis of up to nine additional datasets and only stopped earlier if a significant result with a positive sign was obtained. If none of the 10 datasets yielded a significant relationship with a positive sign, the estimate which was closest to the significance threshold has been published. This rule served two purposes: firstly, it seemed plausible that an actor who has tried that many analyses wants to get the results published in the end to compensate for the invested effort and to avoid sunk costs (Thaler 1980). Secondly, from a technical point of view, this allowed to keep the number of observations in a meta-analysis $K$ constant across all simulation conditions.

In the third condition an actor did not try to achieve significant results by running the same bivariate analysis on different samples, but rather tried to run different model specifications on the same data by

including control variables ($z_j$) to achieve statistical significance of the coefficient of interest ($\beta x$). The third condition therefore modelled publication bias as a form of *p-hacking*, because the existing dataset was optimised to receive a significant *p*-value. The actor was able to add three different control variables to the model. The control variables were defined as collider variables that are both an effect of *x* as well as *y*, which biases the effect of interest (Cole et al. 2009; Greenland et al. 1999). The effect of *x* and *y* on $z_j$ was, however, only small ($\gamma = 0.5$). The error term of the equation defining *z* was normally distributed N(0,10). With three available control variables $z_j$ an actor had seven different combinations to improve the research results in order to obtain a significant effect of *x* on *y*.

In contrast to the second and third conditions, where 50% of the actors had the intention to commit publication bias, in the fourth and fifth conditions all actors had the intention to engage in publication bias practices, once again either through *file-drawer* (fourth condition) or *p-hacking* behaviour (fifth condition). Part from the higher intention to engage in publication bias practices the settings remained the same. Although the two conditions where all actors had the intent to engage in publication bias are far too pessimistic, they allow to evaluate the performance of the tests in the most extreme publication bias environment. Tests that are not able to detect publication bias even under such extreme conditions are of low utility to the research community.

The resulting design matrix had 100 different combinations resulting from 20 data setup conditions multiplied by the five publication bias conditions. In order to obtain reliable estimates similarly to in an experiment (Carsey & Harden 2013: 4f.), every single cell of the design matrix had to be replicated multiple times.[57]

The aim of the simulation study at hand was to compare the performance of the four tests in respect of: A) their capability to detect publication bias if present (true positive, statistical power), as well as B) consistent false positive classification ($\alpha$-error). Because the conditions with and without publication bias are known in a simulation study, the power of the tests and the false positive rate is computable (Mooney 1997: 77-79). In a first step, a dummy variable (*s*) was constructed, with the value 1 for a significant test result below the significance threshold (5% significance level; $s = 1 \text{ if } p < 0.05$). The statistical power, was then defined as the proportion of significant results *s* in respect to all *runs* with publication bias ($\sum_{i=1}^{r} s_i / run \;\; if\; PB > 0$). The false positive rate was computed equivalently but in conditions without publication bias ($\sum_{i=1}^{r} s_i / run \;\; if\; PB = 0$).

[57] In order to specify the number of replications that are necessary to achieve a sufficient statistical power of at least 80% (Cohen 1988: 56) a power analysis was conducted for the statistical power, as well as the false positive rate estimates. For the false positive rate a small deviation of 1 percentage point from the set 5%-false positive rate has to be correctly identified with at least an 80% chance. To achieve this goal, every condition without publication bias had to be supported with 3729 runs. As deviations in power are, though important, not as essential as the false positive rate (Cohen 1988: 56) a difference of 3 percentage points is set as acceptable. In order to identify a 3 percentage point deviation from the target power of 80% each of the 80 conditions with existing publication bias needed 1545 runs. In total, 198,080 runs were necessary, resulting in nearly 109 million primary studies that in the case of publication bias contained up to 10 different regression models.

## 3.3. Results

### *3.3.1. Prevalence of publication bias*

Because publication bias in the experimental setup was implemented as the intent to commit publication bias, three variables are useful to address the actual publication bias and its impact on the overall bias. Firstly, the share of actual studies per meta-analysis that suffer from publication bias (if $p < 0.05$ or negative result are obtained as a first result), secondly the share of studies that achieve their goal of a significant positive result by publication bias, and thirdly the impact of publication bias on the *p*-value of a fixed-effect meta-analysis (deflation factor of the *p*-value). Because the heterogeneous effect condition of the simulation does not allow an absolute bias measure the *p*-value deflation factor was used for all conditions.

In a first step the focus is on how the opportunity structures of the simulation conditions shape the committed publication bias. In the first two columns of Table 3-3 the actual committed publication bias is shown dependent on study characteristics like the mean number of observations in the primary studies ($N$) and the underlying effect ($\beta$), including effect heterogeneity. As expected around 50% respectively 100% of the studies committed publication bias in case of an underlying null-effect, because only 2.5% of the results had the right positive sign and were significant just by chance. In case of an underlying effect the share of committed publication bias decreased because an already significant finding made a publication bias treatment unnecessary. For $\beta = 0.5$ in the 50% publication bias condition only 35%; for $\beta = 1$, 15% for $\beta = 1.5$ only 9%; and in the heterogeneous condition 22% of the studies employed publication bias practices. The 100% publication bias condition approximately doubled the prevalence rates of the 50% condition as expected.

Besides the necessity of publication bias to achieve significant results also the success probability in respect to committed publication bias depended on the conditions of the primary studies as shown by the third column of Table 3-3. For small studies ($N = 100$) with an underlying null-effect ($\beta = 0$) the success-probability in respect to the committed publication bias was about 31.3%. Publication bias got more effective if larger studies ($N = 500$) provide the primary study with more statistical power. The success probability of publication bias rose dramatically around 50 percentage points if a specific underlying empirical effect existed. Also in case of effect heterogeneity the success probability increased about 35.8 percentage points. Slight differences could be observed in the effectivity of the publication bias mechanism, as *p-hacking* was with 10 percentage points less effective than the *file-drawer* condition to achieve significant results.

**Table 3-3 Risk factors for publication and its impact on bias in the simulated data (OLS regression)**

| | Publication bias committed (50% intention) | Publication bias committed (100% intention) | Publication bias successful (in relation to committed) | Deflation of *p*-value |
|---|---|---|---|---|
| $N$ = 500 | -0.105*** | -0.211*** | 0.179*** | |
| *(ref. N = 100)* | (0.000) | (0.001) | (0.001) | |
| $\beta$ = 0.5 | -0.196*** | -0.391*** | 0.471*** | |
| (ref. $\beta$ = 0) | (0.001) | (0.001) | (0.001) | |
| $\beta$ = 1 | -0.389*** | -0.777*** | 0.513*** | |
| | (0.001) | (0.001) | (0.001) | |
| $\beta$ = 1.5 | -0.451*** | -0.899*** | 0.503*** | |
| | (0.001) | (0.001) | (0.002) | |
| $\beta$ = heterogeneous | -0.319*** | -0.636*** | 0.358*** | |
| | (0.001) | (0.001) | (0.001) | |
| *p-hacking* | | | -0.100*** | 0.197*** |
| *(ref. file-drawer)* | | | (0.001) | (0.002) |
| Committed PB [+10ppts] | | | | -0.018*** |
| *(ref. mean = 32.5%)* | | | | (0.001) |
| Successful PB [+10ppts] | | | | -0.077*** |
| *(ref. mean = 18.8%)* | | | | (0.001) |
| Constant | 0.541*** | 1.080*** | 0.313*** | 0.225*** |
| | (0.000) | (0.001) | (0.001) | (0.002) |
| Observations | 61,760 | 61,760 | 115,843 | 123,520 |
| $R^2$ | 0.939 | 0.958 | 0.648 | 0.168 |

Standard errors in parentheses; *** p<0.001, ** p<0.01, * p<0.05

*Note: The first two columns in Table 3-3 show that actual committed publication bias behaviour depended largely on the opportunity structure of the underlying data. Despite the defined 50% or 100% willingness of the actors to commit publication bias, only those actors who face insignificant effects (caused by small effects and sample sizes) engaged in publication bias practices. The success of publication bias in terms of significant results is shown in column three, dependent on the opportunity structure and form of publication bias. Conditions under p-hacking were slightly less effective in obtaining significant results than conditions under file-drawer publication bias. Column four shows the deflating impact of publication bias on meta-analytic p-values. For an average publication bias this p-values halved or even quartered.*

As publication bias deflates *p*-values and therefore biases meta-analytical effect estimates the impact of the actual observed publication bias (the share of committed and successful publication bias) on the meta-analytical *p*-value is presented. The fourth column of Table 3-3 shows that with an average proportion of publication bias committed (32.6%) as well successfully implemented (18.8%) in a meta-analysis with 100 studies ($K$ = 100) the *p*-value of the meta-analysis more than quartered. This is further aggravated if the share of committed as well as successful publication bias rose by 10 percentage points. The actual impact of successful publication bias deflated the *p*-values by 7.7 percentage points and was more pronounced than the deflation caused by non-successfully committed publication bias (deflation by 1.8 percentage points). The deflation also was less severe if *p-hacking* procedures, as implemented in the simulation at hand, were used. Nonetheless, the meta-analytical *p*-value in case of *p-hacking* is still less than half the size (42.2%) of the unbiased estimate.

### *3.3.2. False positive rate of publication bias tests*

For the evaluated tests on publication bias consistent false positive rates are most important. In the simulation none of the tests should exceed the prespecified 5% error probability in any condition. The false positive rate of the test was fixed in the simulation setting to 0.05, so all false positive rates should be equal to, or even smaller than, 0.05. Positive deviations from 0.05 point to inflated false positive rates, which lead to more false conclusions than expected.

**Table 3-4 Conditional false positive rates of the publication bias tests (OLS regression)**

| | PU | FAT | TES | 3% CT | 5% CT | 10% CT | 15% CT |
|---|---|---|---|---|---|---|---|
| $K$ = 1000 | -0.005*** | 0.000 | 0.006*** | 0.026*** | 0.023*** | 0.026*** | 0.050*** |
| (ref. $K$ = 100) | (0.001) | (0.002) | (0.001) | (0.001) | (0.001) | (0.001) | (0.002) |
| $I^2$ [+10 percentage points] | -0.003*** | -0.001*** | -0.001*** | 0.000 | 0.000 | -0.001*** | -0.003*** |
| | (0.000) | (0.000) | (0.000) | (0.000) | (0.000) | (0.000) | (0.000) |
| Constant [a] | 0.023*** | 0.049 | 0.010*** | 0.003** | 0.008*** | 0.019*** | 0.028*** |
| | (0.001) | (0.001) | (0.001) | (0.001) | (0.001) | (0.001) | (0.001) |
| Observations | 73,960 | 74,560 | 74,560 | 62,644 | 66,546 | 69,718 | 70,936 |
| $R^2$ | 0.005 | 0.000 | 0.002 | 0.010 | 0.007 | 0.006 | 0.017 |

[a] Test H0: constant = 0.05; Standard errors in parentheses; *** p<0.001, ** p<0.01, * p<0.05

*Note: Table 3-4 displays the false positive rates of the publication bias tests conditional on the number of studies included in the meta-analysis (K) as well as the between study heterogeneity ($I^2$). The FAT had the most consistent false positive rate. The 15% CT missed the 5%-level clearly while the 10% CT showed a large variability and gets close to it. The 10% and 15% CT are therefore problematic because they may suffer from inflated false positive rates.*

Table 3-4 shows the false positive rate in dependence of the number of studies included ($K$ = 100, 1000) and effect heterogeneity measured by $I^2$. In the constant condition of a meta-analysis with $K$ = 100 and no effect heterogeneity none of the tests had larger false positive rates than the expected 0.05. In particular, the TES, the 3% and 5% CTs were very conservative. A larger meta-analytical sample increased the false positive rates for the TES and the CTs. The broadest 15% CT missed the expected significance threshold of 5%, with 7.8% clearly. The false positive rates for PU in contrast were slightly lower. Increasing effect heterogeneity resulted in more conservative false positive rates for PU, the 15% CT, and to a smaller extent also for the FAT, the TES and the 10% CT. The narrower 3% and 5% CTs were unaffected by effect heterogeneity.

The overall influence of the varied conditions on the false positive rate was small, as can be seen by the small share of explained variance ($R^2$ < 1.7%). Looking at the false positive rates by each condition (Table 3-6 in the appendix) only the 10%- and 15%-caliper showed increased false positive rates because the underlying true effect rather than publication bias elicited an overrepresentation of just significant values. Note however, that the 3% and 5% CTs showed now increased false positive rates.

### *3.3.3. Statistical power of publication bias tests*

The following regression model (Table 3-5) addresses the statistical power conditional on the type of publication bias and its occurrence (committed as well as successful publication bias). Starting from the baseline condition of a meta-analysis with $K = 100$, a mean share of publication bias committed (32.6%), as well as successfully applied (18.8%) via a *file-drawer* procedure and no effect heterogeneity, the FAT had a superior power of 56.9%, followed by the TES (51.5%) and the PU (48.3%). The CTs performed worst and yielded only a power of 0.0%-38.6%.

**Table 3-5 Conditional statistical power of the publication bias tests (OLS regression)**

| | PU | FAT | TES | 3% CT | 5% CT | 10% CT | 15% CT |
|---|---|---|---|---|---|---|---|
| $K = 1000$ | 0.165*** | 0.244*** | 0.238*** | 0.573*** | 0.513*** | 0.382*** | 0.307*** |
| *(ref. K = 100)* | (0.002) | (0.002) | (0.002) | (0.002) | (0.002) | (0.002) | (0.002) |
| $I^2$ [+10 percentage points] | -0.065*** | 0.001* | -0.064*** | 0.006*** | 0.005*** | 0.008*** | 0.010*** |
| | (0.000) | (0.000) | (0.000) | (0.000) | (0.000) | (0.000) | (0.000) |
| *p-hacking* | 0.048*** | -0.110*** | 0.075*** | 0.179*** | 0.187*** | 0.186*** | 0.177*** |
| *(ref. file-drawer)* | (0.002) | (0.002) | (0.002) | (0.002) | (0.002) | (0.002) | (0.002) |
| Comitted PB [+10ppts] | 0.051*** | 0.030*** | -0.065*** | -0.035*** | -0.053*** | -0.073*** | -0.084*** |
| *(ref. mean = 32.6%)* | (0.000) | (0.001) | (0.001) | (0.001) | (0.001) | (0.001) | (0.001) |
| Successful PB [+10ppts] | 0.103*** | 0.099*** | 0.221*** | 0.162*** | 0.193*** | 0.224*** | 0.234*** |
| *(ref. mean = 18.8%)* | (0.001) | (0.001) | (0.001) | (0.001) | (0.001) | (0.001) | (0.001) |
| Constant [a] | 0.483*** | 0.569*** | 0.515*** | -0.002 | 0.125*** | 0.300*** | 0.386*** |
| | (0.002) | (0.002) | (0.002) | (0.002) | (0.002) | (0.002) | (0.002) |
| Observations | 123,520 | 123,520 | 123,520 | 107,736 | 111,315 | 115,243 | 117,207 |
| $R^2$ | 0.572 | 0.306 | 0.473 | 0.497 | 0.483 | 0.457 | 0.446 |

[a] Test H0: constant = 0.8; Standard errors in parentheses; *** p<0.001, ** p<0.01, * p<0.05

*Note: Table 3-5 shows the statistical power of the publication bias tests conditional on the number of studies included in the meta-analysis (K) and the between study heterogeneity ($I^2$). In contrast to Table 3-4, also the share of committed as well as successful publication bias and its form as either file-drawer or p-hacking was controlled. Overall the FAT had the largest power but was not able to detect p-hacking as good as the TES. The CTs were underpowered if a low number of studies was included in a meta-analysis but performed well in studies with large Ks. Both, PU and the TES, were not able to detect publication bias under effect heterogeneity.*

The underperformance of the CTs is largely explained by the small number of studies in the meta-analyses. With $K = 100$ hardly any study falls within the small caliper around the significance threshold. This limitation on just significant or non-significant effects also led to missing values, because without observations in the caliper no CT could be performed. The underperformance of the CT changed if 1000 studies were included, which improved the estimated power substantially, by 30.7-57.3 percentage points, while smaller calipers profited most. The FAT, the TES, and the PU, profited moderately from an increased number of studies, by 24.4, 23.8 and 16.5 percentage points, respectively. When focussing

on the influence of heterogeneity in the meta-analyses the PU and the TES showed a drastic drop in power, by 6.5 and 6.4 percentage points, if the heterogeneity measured by $I^2$ rose by 10 percentage points. This decrease in power shows that neither PU nor TES were able to cope with heterogeneity. In contrast, the FAT and the CTs actually showed a slight increased statistical power. Varying the publication bias procedure from *file-drawer* to *p-hacking*, which is less related to the standard error of the effect estimates, increased the power of PU, TES, and the CTs. The CTs profited most, increasing the statistical power by around 18 percentage points. The TES and PU showed a smaller increase of power, by 7.5 and 4.8 percentage points. The FAT, in contrast lost about 11 percentage points of its power under *p-hacking* compared to the *file-drawer* condition. Although the differences in power are dependent especially on the operationalization of the *p-hacking* condition in the simulation, this result points on a weakness of the FAT under non *file-drawer* conditions that are less related to the standard error of the estimate but are still detectable in the distribution of *z*- or *p*-values.

The structural difference between tests based on a continuous effect distribution (FAT, PU) and tests that focus only on a dichotomous classification (TES, CTs)[58] becomes evident looking at the effect of the proportion of studies that underwent a publication bias treatment in the simulation and the proportion of studies that had a successful outcome after publication bias. Increasing the share of studies under publication bias lifted the power by 3.0 (FAT) and 5.1 (PU) percentage points. A 10 percentage point increase in studies successfully applying publication bias increases the power by 9.9 (FAT) and 10.3 percentage points (PU). The TES and the CTs, however, were only able to detect successful publication bias. An increase only in studies committing publication bias (whether successful or not) in contrast reduced the statistical power. Both tests were therefore not able to detect all possible outcomes of publication bias. This is especially problematic as non-successful publication bias may also inflate the overall estimated effect in meta-analyses. All effects presented are statistically significant ($p < 0.05$).

In contrast to the influence of the varied conditions on the false positive rate, the influence on statistical power was substantial, varying from 30.6% in the case of the FAT to 57.2% for the PU. This finding underlines the fact that all publication bias tests have their strengths and weaknesses in specific conditions.

## 3.4. Discussion & Conclusion

In the simulation at hand, the performance of four different tests (PU, FAT, TES, CTs) were evaluated in a Monte Carlo simulation. Different conditions were varied: the underlying true effect size, including effect heterogeneity, the number of observations in the primary studies, the number of studies in the meta-analyses, the degree of publication bias and its form as either *file-drawer* or *p-hacking*.

[58] Significant or not (TES) over- or under-caliper (CTs).

### *3.4.1. Limitations*

In order to compare the tests in a realistic setting that is nonetheless at least from the assumptions of all four tests applicable, four central limitations have to be pointed out:

Firstly the simulation and its according publication bias procedures rest on the assumptions that all the correlation between the study's precision and its effect size is caused by publication bias. In case that studies with larger effects are, for example after a pre-study power analysis (Lau et al. 2006) conducted with a lower number of observation especially the FAT may yield increased false positive rates (Schwarzer et al. 2002).

Secondly the number of observations included in the meta-analyses either set to $K = 100$ or 1000 is large compared to the average meta-analysis (Elia et al. 2016). The results however showed that even in such large meta-analyses and especially in the more realistic condition in which 50% of the actors are willing to commit publication bias, the tests hardly yielded an adequate statistical power under most conditions. Increasing the number of included studies is therefore important to assure an adequately powered test on publication bias.

Thirdly the analysis focused only on one specific form of *p-hacking* that could occur in both small ($N = 100$) or large studies ($N = 500$). Especially for studies where $N$ is small, other strategies like optional stopping may also be applied. Further research on publication bias should therefore focus on the different impact of other *p-hacking* practices.

As a fourth limitation only one-sided publication bias against insignificant or negative results was simulated. By assumption especially PU limits only on the negative or positive signed studies that were supposed to be affected by publication bias. Also, the FAT is not able to detect two-sided publication bias because the funnel in this situation may still be perfectly symmetric. The suggestions for applications if two-sided publication bias is suspected are therefore limited to the TES and the CTs only.

### *3.4.2. Conclusion*

The following five conclusions can be derived from the results: Firstly, for homogenous research settings and with publication bias favouring only effects in one direction (one-sided publication bias) the FAT is recommended due to its most consistent false positive rate as well as its superior statistical power. Secondly, if there are concerns whether there are any correlations between the precision of the study and its effect size for other reasons than publication bias (see first limitation) and if *p-hacking* is suspected, the TES should be preferred to the FAT under effect homogeneity. As the 5% CT offers more relaxed assumptions it is therefore the first alternative for the FAT under effect heterogeneity if a large number of studies is included in the meta-analysis.

Despite the analysis focussed only on one-sided publication bias, also two-sided publication bias, favouring significant results with either sign may also be present. As PU and the FAT are only able to

identify one-sided publication bias the TES and the CTs remain for two-sided publication bias. Therefore, fourthly the TES is recommended under effect homogeneity because of its larger statistical power compared to the CTs. Fifthly, in the case of heterogeneous effect sizes and a sufficient number of observations in the meta-analysis the 5% CT provides the best trade-off between a conservative false positive rate and a decent statistical power.

The 5% CT is therefore best used to identify publication bias in an effect heterogeneous discipline-wide setting which relies per definition on completely different underlying effects but offers enough studies to compensate for the low statistical power. Because the wider 10% and 15% CTs yield inflated false positive rates, at least in some conditions, they are not recommended to identify publication bias.

Identifying publication bias in substantial meta-analyses as well as focussing on publication as a general problem within the scientific domain is necessary in order to establish and retain trust in scientific results. Further research, however, should not only focus on the diagnosis of publication bias just stating a problem that is well known (Morey 2013). Beyond the nonetheless important diagnosis of the scientific "disease" a further examination of the risk factors, either on the side of the involved actors or with regard to the incentive structure within the discipline (see for example Auspurg & Hinz 2011a) seem essential. This includes also the evaluation of possible interventions (e.g. an open data policy). Research on publication bias is inevitable to maintain trust in scientific results and avoid wasted research funds that also limit the efficiency of science at a whole.

Beside the diagnosis of publication bias and its risk factors, also estimators of the unbiased effect, that are beyond the scope of this paper, like the effect estimates provided by PU and the PET (for example the PET/PEESE procedure of Stanley & Doucouliagos 2014) should be evaluated comparatively. This is at most important for meta-analyses with a heterogeneous effect that try to uncover the underlying true effect rather than test for publication bias alone.

## 3.5. Appendix

The online appendix provides further insights into the methodology of the tests evaluated in the Monte Carlo simulation study. Furthermore, the false positive rates as well as the statistical power of each simulated condition are presented.

### *3.5.1. Statistical tests on publication bias in detail*

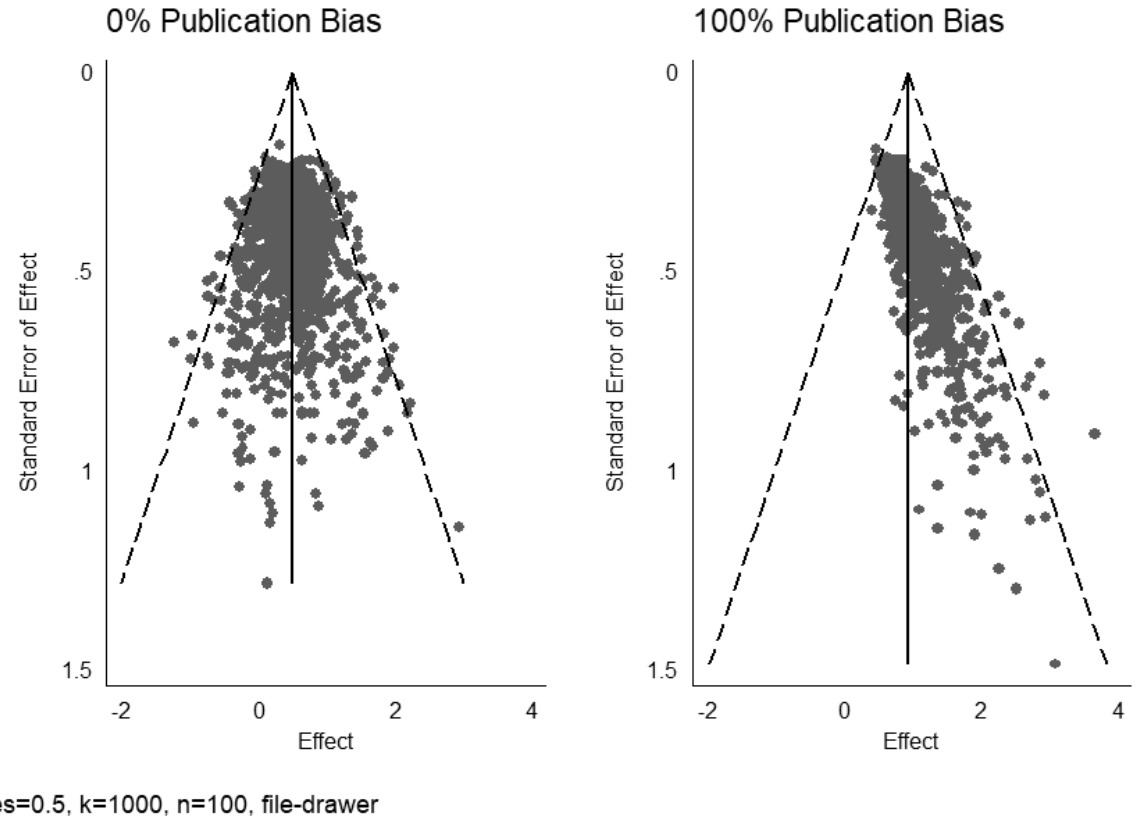
In the following section four publication bias tests, the regression-based FAT (Egger et al. 1997; Stanley & Doucouliagos 2014), PU (van Aert et al. 2016; van Assen et al. 2015), an extended version of p-curve (Simonsohn et al. 2014a,b; Simonsohn et al. 2015), the TES (Ioannidis & Trikalinos 2007) and the CT (Gerber & Malhotra 2008a,b) are discussed in detail.

All of these tests are applied mostly in a discipline-specific context: The FAT is routinely used in classical meta-analyses across all disciplines (cf. the Cochrae Handbook Higgins & Green 2008: 314), PU (for applications see Blázquez et al. 2017; Head et al. 2015; Simmons & Simonsohn 2017), as well as the TES (for applications see Francis 2012a,b,c,d,e,2013) are more widely used in psychology. The CT is in contrast mostly implemented in the general social sciences (for further applications in Sociology and Political Science see Auspurg & Hinz 2011a; Auspurg et al. 2014; Berning & Weiß 2015; Gerber & Malhotra 2008a,b; in Psychology see Hartgerink et al. 2016; Kühberger et al. 2014). The discipline-specific use of the tests is therefore to a certain degree path dependent on the practices involved in testing publication bias in the specific fields.

#### *Funnel asymmetry test (FAT)*

The first class of tests makes it possible to address publication bias by the association of the effect sizes and their variance. Because the variance *($se^2$)* of an effect size in a primary study *(es)* is strongly related to the sample size, small studies with a low number of observations (*N*) show an increased variation of effects around the unobserved true effect. The larger the *N*, the smaller the variation and thus the more precise is the effect size of the study. Under publication bias small non-significant studies are mostly omitted, whereas small but precise effects with a large *N* still remain in the analysis. When this pattern for a small positive effect is represented through a scatterplot graph a typical inverted funnel-shaped pattern can be observed (called "funnel plot" Light & Pillemer 1984: 63-69). In the exemplary Figure 3-1 on the right, studies in the lower left side are missing because of publication bias with a preference for significant positive effects. On the left side, in contrast, a symmetric funnel with no publication bias is shown.

**Figure 3-1 Funnel plot**




*Note: Exemplary funnel plot showing a symmetric funnel in the unbiased left graph and an asymmetric funnel in the right graph with an asymmetry towards positive effects.*

Relying only on subjective graphical information, as provided by funnel plots, might be misleading (Tang & Liu 2000). Begg & Mazumdar (1994: 1089) examine the rank correlation of the standardised effect ($t = es/se$) and its variance ($se^2$). A similar approach by Egger et al. (1997)[59] regresses $t$ on the inverse standard error ($1/se$). $t$ is chosen as the dependent variable in order to account for the unequal variance across the effects (heteroscedasticity) by weighting each observation by the inverse of its variance. Compared to the regression of *se* on *es* this changes the interpretation.

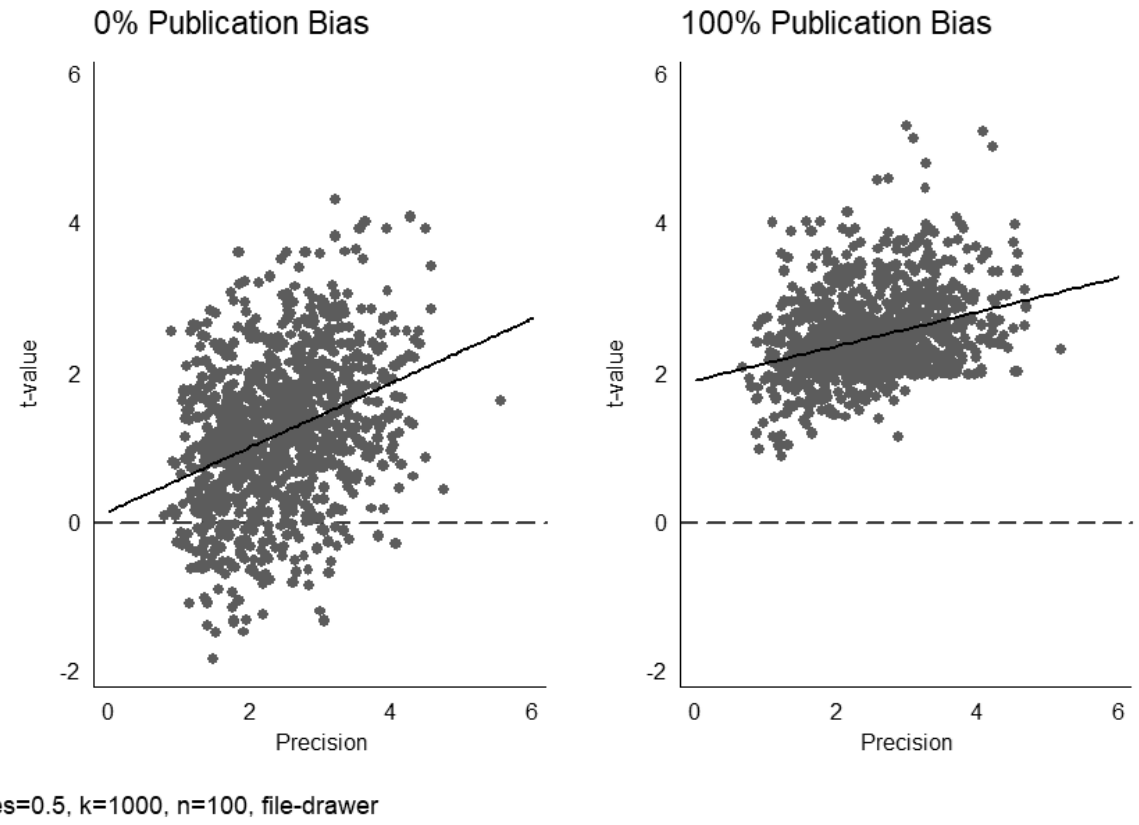
$$t_i = \beta_0 + \beta_1 \frac{1}{se_i} + \varepsilon_i$$

The constant $\beta_0$ is the test on publication bias (FAT stating publication bias if $\beta_0 \neq 0$), whereas $\beta_1$ makes it possible to identify a true empirical effect controlling for publication bias (Egger et al. 1997: 632). In the left graph of Figure 3-2 a primary study (depicted as one dot), with almost no precision, would not

[59] This estimator is equivalent to the bivariate FAT-PET recommended by Stanley & Doucouliagos (2014). The FAT-PET furthermore makes it possible to also include "potential effect modifiers" (Deeks et al. 2008: 284) in a meta-regression model. This is especially necessary if the literature being studied has, besides its theoretical meaningful overall effect, systematic differences (e.g. different implementations of an experimental stimulus, different experimental populations, etc.).

able to find an effect ($H_0$: $\beta_0 = 0$ could not be rejected). In contrast, in the right graph under publication bias a study with no precision would also find a substantial effect.

**Figure 3-2 Funnel asymmetry test (FAT)**




*Note: Exemplary graphical example of the FAT indicating no publication bias in the left (intercept through the origin) and publication bias in the right graph (positive intercept).*

Despite its strengths, the central weaknesses of the FAT lies in its low statistical power in a setting with only a small number of primary studies (Macaskill et al. 2001 simmulated the performace only based on 20 primary studies).[60]

***p-uniform (PU)***

The tests discussed so far focus on the empirical effect sizes, whereas the p-curve method, proposed by Simonsohn et al. (2014b), and the similar PU, a method proposed by van Assen et al. (2015), focus entirely on the distribution of significant *p*-values. All non-significant values are therefore dropped from the analysis. The sample is, furthermore, restricted to the direction of suspected publication bias: that means only positive or negative effects are examined (Simonsohn et al. 2014a: 677). In the first step,

[60] In addition to the performance of the FAT, multiple simulation studies (Alinaghi & Reed 2016; Paldam 2015; Reed 2015) also examine the unbiasedness of the effect estimate (PET – the estimated underlying effect size corrected on publication bias) which is not of interest in the study at hand. The PET is especially threatened by an increased false positive rate under effect heterogeneity (Deeks et al. 2005; Stanley 2017), the properties of the FAT in these conditions have not yet been examined.

the *p*-value of the estimate in the primary study is rescaled in respect to the significance threshold. For the present study the 5%-significance threshold ($p = 0.05$) rescales the *pp*-values to the range [0,1]. This *p*-value of *p*-values (*pp*-value) reflects the probability under the null hypothesis of a non-existing effect that a *p*-value would be as small as, or even smaller than, the observed one.[61]

$$pp_i = \frac{p_i}{0.05} = \frac{1 - \Phi\left(\frac{es_i}{se_i}\right)}{0.05} \quad if\ p_i < 0.05$$

In a second step the skewness of the *pp*-distribution is tested (Simonsohn et al. 2015: 1149). Right skewness shows an overrepresentation of findings with a substantial statistical significance and indicates a genuine empirical effect. Left skewness, in contrast, shows an overrepresentation of just significant estimates that barely pass the significance threshold (in this case 5%) and indicates publication bias under the null hypothesis (Simonsohn et al. 2014b: 536).

Whereas p-curve by Simonsohn et al. (2014b) only allows to identify publication bias under a true underlying null effect, PU (van Assen et al. 2015) allows to also identify publication bias under an empirically observed effect. This seems essential in order to distinguish between an underlying true effect and publication bias as criticised by Bruns & Ioannidis (2016) for p-curve. For PU as a first step the underlying effect has to be estimated empirically by a fixed-effect meta-analysis (FE-MA)[62] with all primary studies. In a second step, and equivalent to p-curve, only *k* estimates with $p < 0.05$ and the direction of the suspected publication bias remain in the analysis (van Aert et al. 2016: 727). By adjusting on the existing underlying effect, the fixed-effect estimate $\mu$, it is possible to test the skewness of the distribution conditional on the underlying empirical effect (van Assen et al. 2015). In the case of an underlying null-effect, p-curve is therefore a special case of PU. In the numerator, the effect size estimate is conditioned on the underlying effect ($\mu$), similar to a one-sample *z*-test. The denominator of the *pp*-value is not fixed to 0.05 as in p-curve, but is also conditioned on the underlying effect ($\mu$), which is subtracted from the effect threshold (*et*) an effect has to reach to become statistically significant given its standard error (*se*).

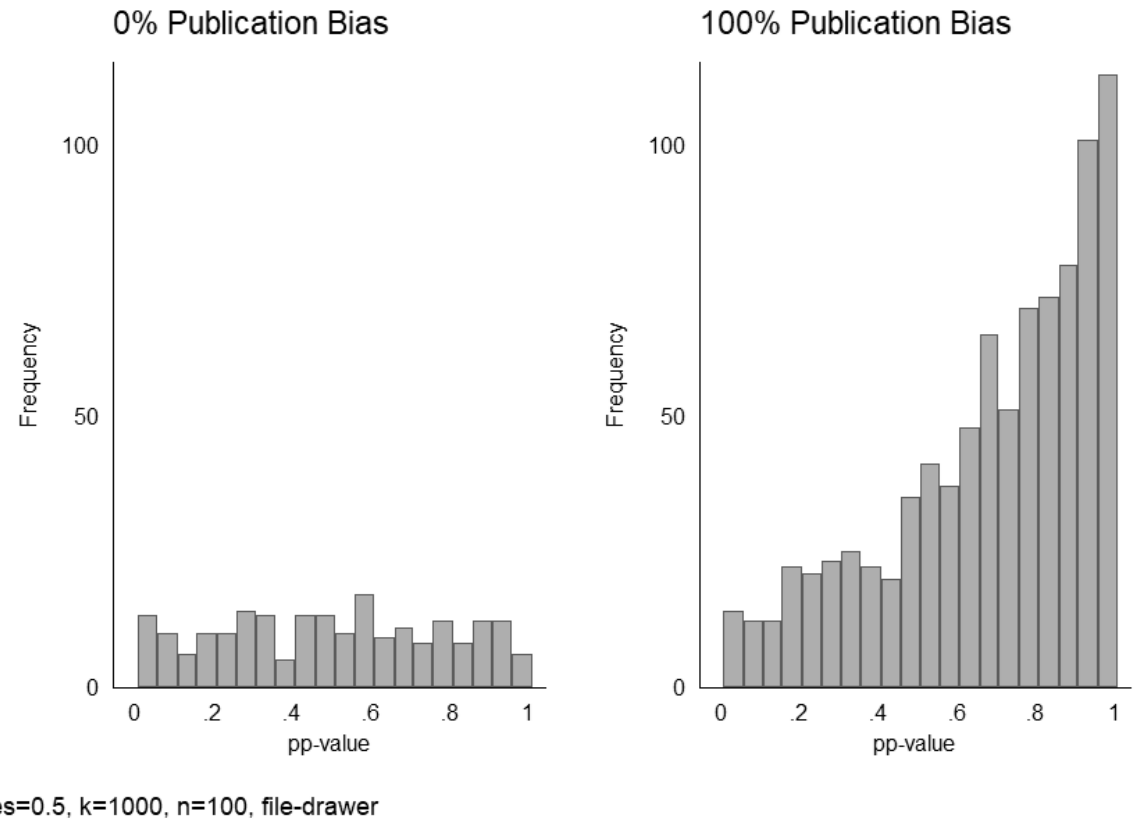
$$pp_i^{\mu} = \frac{1 - \Phi\left(\frac{es_i - \mu}{se_i}\right)}{1 - \Phi\left(\frac{et_i - \mu}{se_i}\right)} \quad if\ p_i < 0.05$$

The test statistic is gamma-distributed with *k* degrees of freedom.[63] Because the skewness is now conditional on the underlying empirical effect left skewness observed by PU identifies publication bias across all underlying empirical effects, as depicted in Figure 3-3.

[61] $\Phi$ represents the standard normal distribution.
[62] Mean effect size across all included studies weighted by the inverse study variance.
[63] $p = \Gamma\left(k, -\sum_{i=1}^{k} \log(pp_i^{\mu})\right)$

Because PU rests on the average effect size estimated by a fixed-effects meta-analysis it may sensible to effect heterogeneity. The degree of heterogeneity which invalidates the publication bias test is, however, unclear for PU.[64]

**Figure 3-3 p-uniform (PU)**




*Note: Graphical example of the PU indicating no publication bias in the left (uniform distribution of pp-values) and publication bias in the right graph (left skewed distribution of pp-values).*

van Assen et al. (2015) evaluate the performance of PU, the TES (a publication bias test, discussed in the next section), as well as trim-and-fill, and conclude that PU has a greater statistical power than the other methods (van Assen et al. 2015: 303). Also, Renkewitz & Keiner (2016) evaluate the PU publication bias test and observe its slightly better performance compared to the FAT and the TES. However, in both studies the number of studies in the meta-analyses (max. 160), as well as the number of observations (max. 80) in the primary studies, is relatively small.[65]

[64] Simonsohn et al. (2014a: 680) state that p-curve is able to estimate the average true effect of the observed significant studies correctly, whereas van Aert et al. (2016: 718) note the sensitivity towards heterogeneity of PU referring to the true underlying effect of all studies, which is mostly of concern in meta-analyses.

[65] Similar to the FAT-PET, evaluations of PU center mainly on the estimated overall effect. While van Assen et al. (2015) show a good coverage of the estimated overall effect, McShane et al. (2016) state, in contrast, that while "p-curve and p-uniform approaches have increased awareness about the consequences of publication bias in meta-analysis, they fail to improve upon, and indeed are inferior to, methods proposed decades ago" (McShane et al. 2016: 744).

***Test for excess significance (TES)***

The TES builds on the observed power of every single study to uncover the true total effect. This true effect is estimated by a fixed-effect meta-analysis, as in PU. Observed power analyses make it possible to compute the post hoc power ($pw_i$) of a study. This allows to specify the expected number of significant effects *E*, given the average effect as well as the significance threshold (in this case $\alpha = 0.05$).[66]

$$E = \sum_{i=1}^{k} (pw_i)$$

*E* may even be a conservative estimate of the expected number of significant studies because it heavily relies on the fixed-effect estimate, which suffers from an eventual publication bias. In relation to *O*, the empirically observed number of significant studies ($p_i < 0.05$) the TES tests whether more significant results than expected are reported in the literature. To test whether the share of observed positive outcomes $\left(\frac{O}{K}\right)$ is larger than the share of expected positive outcomes $\left(\frac{E}{K}\right)$ a one-sided binomial test is used (Ioannidis & Trikalinos 2007: 246).

On exemplary datasets the TES performs considerably better under moderate effect heterogeneity in large meta-analyses, where the FAT in particular failed to uncover publication bias (Ioannidis & Trikalinos 2007: 248). Nevertheless, Johnson & Yuan (2007: 254) ask if the TES makes it possible to dissect between publication bias and study-heterogeneity accurately. Therefore, the authors of the *Cochrane Handbook* (Higgins & Green 2008: 323) express the need for further evaluations.

***Caliper test (CT)***

In contrast to the aforementioned three tests, the CT, developed by Gerber and Malhotra (2008a,b) ignores most of the information provided by the studies included and looks only at a narrow interval (caliper = *c*) around the significance threshold (*th*) in a distribution of absolute *z*-values. In case of a continuous distribution of *z*-values, studies in the interval below the significance threshold (in the so-called over-caliper; $x_z = 1$) should be as likely as just non-significant studies (in the so-called under-caliper; $x_z = 0$).

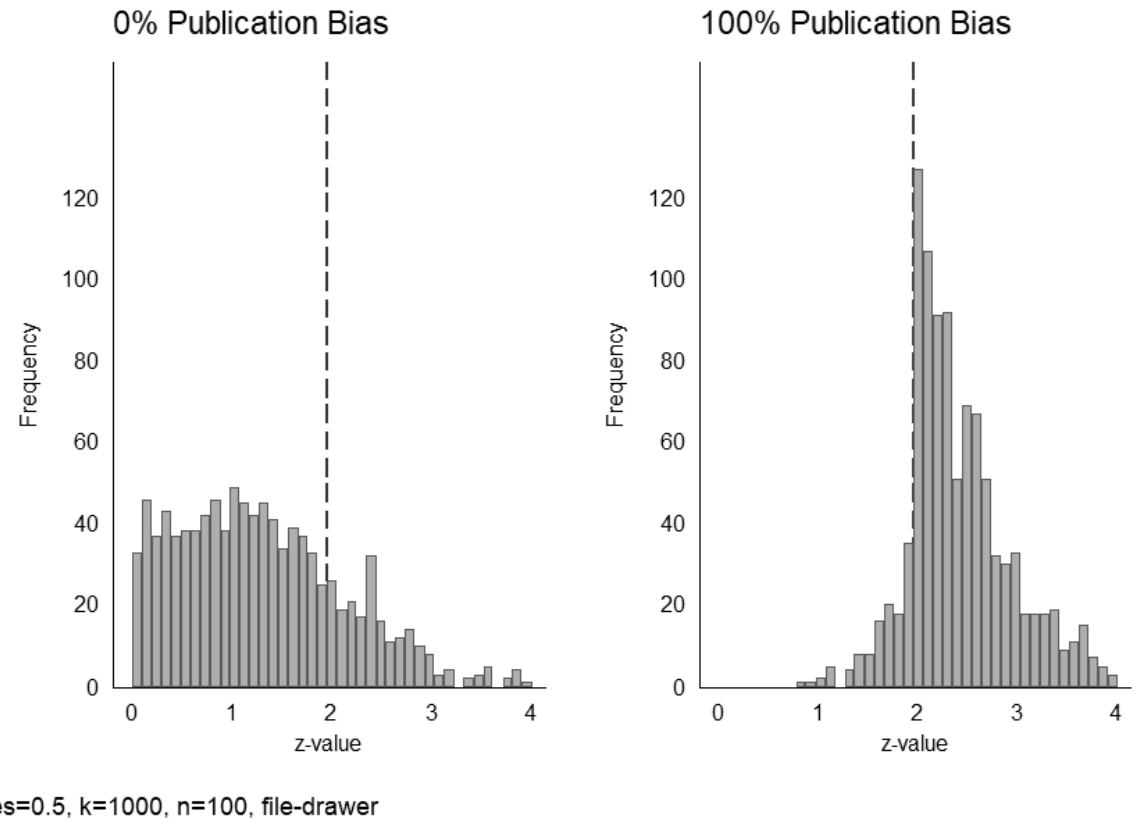
$$x_z = \begin{cases} 0 \; if \; th - c * th < z \leq th \\ 1 \; if \; th < z < th + c * th \end{cases}$$

Gerber and Malhotra (2008a,b) use a 5%, 10%, 15% and 20% interval (*c*) proportional to the significance threshold (*th*). In particular, the widest 20% caliper may be too wide because the 10%-significance level that could be another target threshold for publication bias is fully overlapped. The higher the overrepresentation in the over-caliper, the higher the likelihood of publication bias. This is also shown in Figure 3-4: in the left graph with no publication bias no discontinuities are seen around the arbitrary

[66] Although Hoenig & Heisey (2001) criticise the application of post-hoc power analyses in primary studies for the good reason that the observed power estimate may be biased, meta-analyses circumvent this critique because a distribution of power estimates allows to infer more accurately the power of a set of studies.

5% significance threshold (dashed line), whereas in the right graph a stepwise increase of just significant results indicates publication bias. As with the TES, a one-sided binomial test is used to test the equal distribution of *z*-values in the over- and under-caliper.[67]

**Figure 3-4 Caliper test (CT with 5% caliper)**




*Note: Graphical example of the CT indicating no publication bias in the left (no jump point around the significance threshold visualized by the dashed line) and publication bias in the right graph (jump point at the 5% significance level).*

### *3.5.2. Results in detail by simulation conditions*

The following section presents the results of the false positive rates by each simulation condition. Besides the statistical power (Table 3-7-Table 3-10) of the evaluated publication bias tests also the actual committed as well as successful publication bias is reported along the results. As in the regression analysis in the article also the impact of publication bias on the meta-analytical *p*-value is reported.

[67] Masicampo & Lalande (2012) and Leggett et al. (2013) test the deviance of values around the significance threshold from a fitted exponential curve on *p*-values in a broader range from 0.1 – 0.10 to counter the huge loss of observations in the CT. This may be problematic, because a single distributive function may not be able to describe the pattern well enough across the suspected jump points (cf. Lakens 2015). In the case of substantial effect heterogeneity this problem would be aggravated even further.

***False positive rates***

Table 3-6 shows the false positive rates of the publication bias tests across all simulated conditions. Inflated false positive rates are highlighted in bold. Over all conditions the FAT, PU, the TES, as well as the narrower CTs (3%, 5%), had a consistent false positive rate. The FAT was closest to the expected 5% error rate. PU and the TES, as well as the 3% and 5% CTs, in contrast, were in most cases very conservative because they fall far below 0.05. This over-conservatism may be problematic in respect to a decreased statistical power, a matter which is discussed later on. The wider 10% and 15% CTs suffered under inflated false positive rates because, due to the large caliper width, the assumption of a uniform distribution in both calipers was violated.[68] For the 10% CT the specified false positive rate doubles to more than 10%, whereas in case of the 15% CT it more than quadruples.

[68] This means that an asymmetry between over- and under-caliper is not caused by publication bias rather than by an underlying effect distribution that is skewed in the caliper width.

**Table 3-6 False positive rates by each simulation condition**

| 0% FD/PH | PU | FAT | TES | 3% CT | 5% CT | 10% CT | 15% CT |
|---|---|---|---|---|---|---|---|
| N100/K100 | | | | | | | |
| 0.0 | 0.045 | 0.043 | 0.024 | 0.001 | 0.001 | 0.003 | 0.002 |
| 0.5 | 0.039 | 0.045 | 0.004 | 0.004 | 0.011 | 0.013 | 0.012 |
| 1.0 | 0.014 | 0.056 | 0.005 | 0.005 | 0.017 | 0.033 | 0.040 |
| 1.5 | 0.001 | 0.047 | 0.010 | 0.000 | 0.005 | 0.026 | 0.041 |
| Het | 0.000 | 0.042 | 0.001 | 0.002 | 0.012 | 0.021 | 0.025 |
| N100/K1000 | | | | | | | |
| 0.0 | 0.032 | 0.051 | 0.036 | 0.020 | 0.012 | 0.000 | 0.000 |
| 0.5 | 0.020 | 0.046 | 0.005 | 0.031 | 0.023 | 0.007 | 0.001 |
| 1.0 | 0.008 | 0.048 | 0.003 | 0.043 | 0.049 | **0.067** | **0.092** |
| 1.5 | 0.002 | 0.046 | 0.013 | 0.040 | 0.049 | **0.101** | **0.204** |
| Het | 0.000 | 0.047 | 0.000 | 0.032 | 0.032 | 0.028 | 0.030 |
| N500/K100 | | | | | | | |
| 0.0 | 0.051 | 0.051 | 0.024 | 0.000 | 0.002 | 0.003 | 0.002 |
| 0.5 | 0.025 | 0.050 | 0.002 | 0.010 | 0.019 | 0.039 | 0.043 |
| 1.0 | 0.000 | 0.045 | 0.007 | 0.000 | 0.000 | 0.001 | 0.010 |
| 1.5 | 0.000 | 0.047 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| Het | 0.000 | 0.037 | 0.000 | 0.000 | 0.002 | 0.011 | 0.018 |
| N500/K1000 | | | | | | | |
| 0.0 | 0.043 | 0.052 | 0.037 | 0.019 | 0.009 | 0.001 | 0.000 |
| 0.5 | 0.024 | 0.042 | 0.004 | 0.039 | 0.045 | **0.070** | **0.104** |
| 1.0 | 0.000 | 0.054 | 0.033 | 0.018 | 0.043 | **0.108** | **0.244** |
| 1.5 | 0.000 | 0.048 | 0.003 | 0.000 | 0.000 | 0.002 | 0.007 |
| Het | 0.000 | 0.035 | 0.000 | 0.031 | 0.036 | 0.038 | 0.035 |

Bold numbers > 0.05 at $p < 0.05$

*Note: False positive rates of the seven evaluated tests by each condition. The 10% and 15% CT show an increased false positive rate (highlighted in bold).*

***Statistical power***

Looking at conditions with 50% publication bias in the *file-drawer* condition (see Table 3-7), the FAT had a superior power compared to other tests in 14 of 20 conditions, as indicated by the underlined numbers. The FAT is, however, closely followed by the TES, which had a larger number of conditions with a satisfactory power (> 0.8) compared to the FAT (7 vs. 6). In the first condition with $N = 100$ as well as $K = 100$ the TES was superior in the case of an underlying small or moderate effect ($\beta = 0.5; 1; 1.5$). The large variability of the primary study effect, which was caused by the low-*N* and low-*K* in the meta-analyses, resulted in an overall minor statistical power. A sufficient power (highlighted in bold) was only reached in conditions with a low or moderate underlying true effect ($\beta = 0.5, 1$). This is caused by high prevalence of committed publication bias (PB com) that is also successful (PB suc – meaning $p < 0.05$). None of the CTs yielded a sufficient power. This picture changes if more studies were included in the meta-analysis. With $K = 1000$ most of the tests yielded a sufficient power. In particular, the FAT had a statistical power close to 100%, also under effect heterogeneity. The PU and the TES failed to uncover *file-drawer* behaviour under effect heterogeneity, but performed well under homogeneity. PU was only able to discover *file-drawer* behaviour under low underlying true effects. The CTs profited the most from an increased *K*, the wider caliper (10, 15%) had a larger statistical power than the narrower ones but also had inflated false positive rates (see Table 3-6) that might invalidate the conclusions (grey shaded area). The narrower caliper had a sufficient power only in studies with no or small underlying effects ($\beta = 0; 0.5$). $K = 100$ and $N = 500$ decreased the power of all tests drastically. In this condition the FAT had the largest, but still not satisfactory power. With $K = 1000$ a sufficient power is yielded in conditions with a low overall effect ($\beta = 0; 0.5$).

The statistical power of the tests increased if the intent to engage in *file-drawer* behaviour is set to 100% (see Table 3-8). Overall, more publication bias tests achieved a satisfactory statistical power to detect publication bias. Also, in these conditions, the FAT dominated in 13 of 20 conditions. As before, neither the TES nor the PU were able to detect publication bias under effect heterogeneity. The TES was, furthermore, not able to detect publication bias with an underlying null effect, despite publication bias was successfully applied by 21.3% of the cases. Similar to the 50% *file-drawer* condition, the CTs showed a drastically decreased power in conditions with $K = 100$.

The dominance of the FAT weakened when looking at the 50% *p-hacking* condition (see Table 3-9). Instead, the TES was besides the 15% CT superior under most conditions but had the advantage that its false positive rate was not inflated. The overall pattern was, however, quite similar: both PU and TES had almost no power to detect *p-hacking* under effect heterogeneity. Also, the statistical power was only satisfactory for PU when $K = 100$. With a large number of included studies, however, the power of the CT was close to, or even outperformed, the FAT, PU and the TES.

In the 100% *p-hacking* condition (see Table 3-10) the FAT caught up with the TES and yielded an increased power, especially in the case of $K = 100$. Despite the dominance of the 15% CT, the TES and

the FAT closely followed. The CT had a similar strength to that demonstrated in the earlier conditions under effect heterogeneity and $K = 1000$. The underperformance of all tests in the condition with $N = 500$ and moderate underlying effects ($\beta = 1; 1.5$) is caused by the already existing significance of most results in this condition.

Overall, the FAT dominated under the *file-drawer* condition. The TES, in contrast, had a slightly higher statistical power than the FAT under the *p-hacking* condition without effect heterogeneity. However, the differences between both tests were quite small. The CTs performed well under the *file-drawer* as well as *p-hacking* condition with heterogeneous effect sizes and large numbers of studies included ($K = 1000$). Although the 10% and 15% caliper had the highest power to detect *p-hacking* these tests should not be applied due to their increased false positive rate.

**Table 3-7 Statistical power by each simulation condition for 50% file-drawer publication bias**

| 50% FD | PU | FAT | TES | 3% CT | 5% CT | 10% CT | 15% CT | PB com | PB suc | *p* defl. |
|---|---|---|---|---|---|---|---|---|---|---|
| N100/K100 | | | | | | | | | | |
| 0.0 | 0.179 | 0.662 | 0.148 | 0.007 | 0.013 | 0.015 | 0.005 | 0.488 | 0.215 | 0.154 |
| 0.5 | 0.691 | **0.822** | **0.912** | 0.108 | 0.220 | 0.416 | 0.563 | 0.379 | 0.843 | 0.070 |
| 1.0 | 0.348 | **0.823** | **0.881** | 0.052 | 0.149 | 0.415 | 0.594 | 0.184 | 0.977 | 0.068 |
| 1.5 | 0.034 | 0.457 | 0.537 | 0.007 | 0.032 | 0.164 | 0.285 | 0.073 | 0.997 | 0.260 |
| Het | 0.000 | 0.370 | 0.042 | 0.032 | 0.082 | 0.220 | 0.321 | 0.223 | 0.746 | 0.360 |
| N100/K1000 | | | | | | | | | | |
| 0.0 | 0.720 | **1.000** | 0.737 | 0.029 | 0.019 | 0.001 | 0.000 | 0.487 | 0.215 | 0.000 |
| 0.5 | **1.000** | **1.000** | **1.000** | **0.894** | **0.981** | **1.000** | **1.000** | 0.379 | 0.843 | 0.000 |
| 1.0 | **1.000** | **1.000** | **1.000** | **0.859** | **0.976** | **0.999** | **1.000** | 0.185 | 0.978 | 0.000 |
| 1.5 | 0.521 | **0.999** | **1.000** | 0.530 | 0.763 | **0.981** | **0.999** | 0.074 | 0.998 | 0.000 |
| Het | 0.000 | **0.997** | 0.125 | 0.639 | **0.839** | **0.977** | **0.996** | 0.224 | 0.746 | 0.002 |
| N500/K100 | | | | | | | | | | |
| 0.0 | 0.238 | 0.245 | 0.104 | 0.007 | 0.010 | 0.013 | 0.003 | 0.488 | 0.207 | 0.466 |
| 0.5 | 0.580 | 0.499 | 0.736 | 0.080 | 0.201 | 0.442 | 0.671 | 0.204 | 0.993 | 0.268 |
| 1.0 | 0.001 | 0.110 | 0.039 | 0.000 | 0.001 | 0.003 | 0.021 | 0.013 | 0.998 | 0.752 |
| 1.5 | 0.000 | 0.056 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.001 | 0.994 | 0.944 |
| Het | 0.000 | 0.058 | 0.000 | 0.005 | 0.029 | 0.095 | 0.166 | 0.116 | 0.748 | 0.836 |
| N500/K1000 | | | | | | | | | | |
| 0.0 | **0.905** | **0.950** | 0.544 | 0.043 | 0.028 | 0.001 | 0.000 | 0.487 | 0.207 | 0.019 |
| 0.5 | **1.000** | **0.999** | **1.000** | **0.911** | **0.987** | **1.000** | **1.000** | 0.205 | 0.992 | 0.000 |
| 1.0 | 0.004 | 0.396 | **0.874** | 0.068 | 0.165 | 0.529 | **0.826** | 0.013 | 0.998 | 0.328 |
| 1.5 | 0.001 | 0.064 | 0.019 | 0.000 | 0.000 | 0.005 | 0.019 | 0.001 | 1.000 | 0.887 |
| Het | 0.000 | 0.214 | 0.000 | 0.373 | 0.569 | **0.855** | **0.950** | 0.116 | 0.745 | 0.506 |
| Best / Satisfac-tory | 3/ 4 | 14/ 6 | 8/ 7 | 0/ 3 | 0/ 4 | 2/ 6 | 3/ 7 | | | |

Bold numbers: > 0.8 p < 0.05. Underlined: best estimator. Grey shaded: inflated false positive rate, cf. Table 3-6

*Note: PB com displays the share of studies committing publication bias. PB suc describes the share of studies successfully committing publication bias. p defl. shows the the deflation of the meta-analytical p-value by publication bias.*

**Table 3-8 Statistical power by each simulation condition for 100% file-drawer publication bias**

| 100% FD | PU | FAT | TES | 3% CT | 5% CT | 10% CT | 15% CT | PB com | PB suc | *p* defl. |
|---|---|---|---|---|---|---|---|---|---|---|
| N100/K100 | | | | | | | | | | |
| 0.0 | 0.756 | **1.000** | 0.000 | 0.013 | 0.016 | 0.012 | 0.005 | 0.974 | 0.213 | 0.000 |
| 0.5 | **1.000** | **1.000** | **1.000** | 0.328 | 0.569 | **0.891** | **0.981** | 0.759 | 0.843 | 0.000 |
| 1.0 | **0.958** | **1.000** | **1.000** | 0.222 | 0.618 | **0.962** | **0.999** | 0.371 | 0.978 | 0.000 |
| 1.5 | 0.177 | **0.975** | **1.000** | 0.028 | 0.138 | 0.621 | **0.898** | 0.149 | 0.998 | 0.012 |
| Het | 0.000 | **0.962** | **0.882** | 0.124 | 0.278 | 0.595 | 0.790 | 0.447 | 0.746 | 0.015 |
| N100/K1000 | | | | | | | | | | |
| 0.0 | **1.000** | **1.000** | 0.000 | 0.047 | 0.021 | 0.002 | 0.000 | 0.975 | 0.215 | 0.000 |
| 0.5 | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | 0.759 | 0.843 | 0.000 |
| 1.0 | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | 0.369 | 0.978 | 0.000 |
| 1.5 | **0.999** | **1.000** | **1.000** | **0.999** | **1.000** | **1.000** | **1.000** | 0.149 | 0.998 | 0.000 |
| Het | 0.000 | **1.000** | **1.000** | **0.981** | **0.999** | **1.000** | **1.000** | 0.448 | 0.745 | 0.000 |
| N500/K100 | | | | | | | | | | |
| 0.0 | **0.888** | **0.990** | 0.000 | 0.011 | 0.012 | 0.015 | 0.003 | 0.975 | 0.208 | 0.005 |
| 0.5 | **1.000** | **0.999** | **1.000** | 0.351 | 0.755 | **0.992** | **1.000** | 0.410 | 0.992 | 0.001 |
| 1.0 | 0.001 | 0.221 | 0.235 | 0.000 | 0.000 | 0.006 | 0.047 | 0.026 | 0.998 | 0.527 |
| 1.5 | 0.001 | 0.059 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.002 | 0.997 | 0.931 |
| Het | 0.000 | 0.129 | 0.000 | 0.026 | 0.092 | 0.290 | 0.473 | 0.230 | 0.741 | 0.666 |
| N500/K1000 | | | | | | | | | | |
| 0.0 | **1.000** | **1.000** | 0.000 | 0.039 | 0.021 | 0.003 | 0.000 | 0.975 | 0.206 | 0.000 |
| 0.5 | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | 0.409 | 0.992 | 0.000 |
| 1.0 | 0.093 | **0.898** | **1.000** | 0.233 | 0.669 | **0.995** | **1.000** | 0.026 | 0.998 | 0.042 |
| 1.5 | 0.000 | 0.108 | 0.145 | 0.000 | 0.000 | 0.009 | 0.041 | 0.002 | 1.000 | 0.757 |
| Het | 0.000 | 0.628 | 0.000 | **0.829** | **0.957** | **1.000** | **1.000** | 0.231 | 0.743 | 0.153 |
| Best / Satisfac-tory | 7/ 10 | 13/ 15 | 12/ 11 | 3/ 6 | 4/ 6 | 6/ 10 | 9/ 11 | | | |

Bold numbers: > 0.8 p < 0.05. Underlined: best estimator. Grey shaded: inflated false positive rate, cf. Table 3-6

*Note: PB com displays the share of studies committing publication bias. PB suc describes the share of studies successfully committing publication bias. p defl. shows the the deflation of the meta-analytical p-value by publication bias.*

**Table 3-9 Statistical power by each simulation condition for 50% p-hacking publication bias**

| 50% PH | PU | FAT | TES | 3% CT | 5% CT | 10% CT | 15% CT | PB com | PB suc | *p* defl. |
|---|---|---|---|---|---|---|---|---|---|---|
| N100/K100 | | | | | | | | | | |
| 0.0 | 0.764 | 0.006 | 0.598 | 0.077 | 0.139 | 0.196 | 0.166 | 0.489 | 0.305 | 1.331 |
| 0.5 | **0.870** | 0.371 | 0.490 | 0.152 | 0.288 | 0.465 | 0.527 | 0.379 | 0.535 | 0.357 |
| 1.0 | 0.321 | 0.395 | 0.422 | 0.079 | 0.168 | 0.396 | 0.528 | 0.184 | 0.598 | 0.354 |
| 1.5 | 0.018 | 0.129 | 0.166 | 0.015 | 0.058 | 0.154 | 0.259 | 0.074 | 0.602 | 0.693 |
| Het | 0.000 | 0.369 | 0.020 | 0.068 | 0.139 | 0.292 | 0.380 | 0.223 | 0.507 | 0.356 |
| N100/K1000 | | | | | | | | | | |
| 0.0 | **1.000** | 0.000 | **1.000** | 0.767 | **0.874** | **0.929** | **0.846** | 0.488 | 0.304 | 1.872 |
| 0.5 | **1.000** | **0.992** | **1.000** | **0.973** | **0.995** | **1.000** | **1.000** | 0.379 | 0.539 | 0.003 |
| 1.0 | **0.997** | **0.997** | **1.000** | **0.879** | **0.968** | **0.999** | **1.000** | 0.184 | 0.597 | 0.003 |
| 1.5 | 0.175 | 0.503 | **0.962** | 0.505 | 0.733 | **0.958** | **0.994** | 0.074 | 0.598 | 0.233 |
| Het | 0.000 | **0.994** | 0.007 | 0.797 | **0.942** | **0.997** | **0.999** | 0.224 | 0.507 | 0.004 |
| N500/K100 | | | | | | | | | | |
| 0.0 | **0.958** | 0.000 | **1.000** | 0.211 | 0.394 | 0.659 | 0.769 | 0.491 | 0.684 | 1.969 |
| 0.5 | 0.806 | 0.437 | **0.843** | 0.112 | 0.285 | 0.602 | 0.784 | 0.206 | 0.925 | 0.319 |
| 1.0 | 0.000 | 0.066 | 0.028 | 0.000 | 0.001 | 0.013 | 0.046 | 0.013 | 0.894 | 0.874 |
| 1.5 | 0.001 | 0.058 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.001 | 0.736 | 0.969 |
| Het | 0.000 | 0.408 | 0.000 | 0.015 | 0.067 | 0.233 | 0.383 | 0.115 | 0.843 | 0.310 |
| N500/K1000 | | | | | | | | | | |
| 0.0 | **1.000** | 0.000 | **1.000** | **0.995** | **1.000** | **1.000** | **1.000** | 0.488 | 0.685 | 2.011 |
| 0.5 | **1.000** | **0.999** | **1.000** | **0.966** | **0.999** | **1.000** | **1.000** | 0.204 | 0.923 | 0.001 |
| 1.0 | 0.004 | 0.159 | 0.775 | 0.116 | 0.271 | 0.676 | **0.908** | 0.013 | 0.886 | 0.645 |
| 1.5 | 0.000 | 0.046 | 0.012 | 0.002 | 0.002 | 0.007 | 0.026 | 0.001 | 0.749 | 0.972 |
| Het | 0.000 | **0.997** | 0.000 | 0.772 | **0.935** | **0.998** | **1.000** | 0.115 | 0.846 | 0.001 |
| Best / Satisfactory | 6/ 7 | 4/ 5 | 7/ 8 | 0/ 4 | 1/ 7 | 3/ 8 | 11/ 8 | | | |

Bold numbers: > 0.8 p < 0.05. Underlined: best estimator. Grey shaded: inflated false positive rate, cf. Table 3-6

*Note: PB com displays the share of studies committing publication bias. PB suc describes the share of studies successfully committing publication bias. p defl. shows the the deflation of the meta-analytical p-value by publication bias.*

**Table 3-10 Statistical power by each simulation condition for 100% p-hacking publication bias**

| 100% PH | PU | FAT | TES | 3% CT | 5% CT | 10% CT | 15% CT | PB com | PB suc | *p* defl. |
|---|---|---|---|---|---|---|---|---|---|---|
| N100/K100 | | | | | | | | | | |
| 0.0 | **0.999** | 0.808 | 0.212 | 0.203 | 0.331 | 0.477 | 0.497 | 0.975 | 0.305 | 0.079 |
| 0.5 | **1.000** | **0.997** | **0.992** | 0.481 | 0.727 | **0.918** | **0.964** | 0.759 | 0.536 | 0.002 |
| 1.0 | **0.854** | **0.916** | **0.985** | 0.286 | 0.518 | **0.835** | **0.947** | 0.368 | 0.596 | 0.032 |
| 1.5 | 0.089 | 0.293 | 0.679 | 0.051 | 0.165 | 0.443 | 0.648 | 0.149 | 0.606 | 0.419 |
| Het | 0.000 | **0.903** | 0.390 | 0.235 | 0.436 | 0.724 | **0.847** | 0.450 | 0.506 | 0.039 |
| N100/K1000 | | | | | | | | | | |
| 0.0 | **1.000** | **1.000** | **0.999** | **0.984** | **0.999** | **1.000** | **1.000** | 0.975 | 0.305 | 0.000 |
| 0.5 | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | 0.758 | 0.538 | 0.000 |
| 1.0 | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | 0.369 | 0.595 | 0.000 |
| 1.5 | **0.887** | **0.976** | **1.000** | **0.957** | **0.997** | **1.000** | **1.000** | 0.148 | 0.599 | 0.009 |
| Het | 0.000 | **1.000** | **0.999** | **0.997** | **1.000** | **1.000** | **1.000** | 0.447 | 0.507 | 0.000 |
| N500/K100 | | | | | | | | | | |
| 0.0 | **1.000** | **0.979** | **1.000** | 0.561 | 0.791 | **0.977** | **0.994** | 0.975 | 0.685 | 0.010 |
| 0.5 | **0.999** | **0.997** | **1.000** | 0.525 | **0.847** | **0.995** | **1.000** | 0.411 | 0.923 | 0.001 |
| 1.0 | 0.001 | 0.106 | 0.138 | 0.000 | 0.002 | 0.036 | 0.119 | 0.026 | 0.897 | 0.741 |
| 1.5 | 0.000 | 0.051 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.002 | 0.746 | 0.976 |
| Het | 0.000 | **0.916** | 0.003 | 0.099 | 0.328 | 0.758 | **0.917** | 0.231 | 0.847 | 0.032 |
| N500/K1000 | | | | | | | | | | |
| 0.0 | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | 0.975 | 0.685 | 0.000 |
| 0.5 | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | 0.409 | 0.923 | 0.000 |
| 1.0 | 0.028 | 0.405 | **1.000** | 0.438 | 0.804 | **0.996** | **1.000** | 0.026 | 0.885 | 0.310 |
| 1.5 | 0.000 | 0.061 | 0.028 | 0.000 | 0.002 | 0.022 | 0.072 | 0.002 | 0.739 | 0.953 |
| Het | 0.000 | **1.000** | 0.000 | **0.998** | **1.000** | **1.000** | **1.000** | 0.231 | 0.845 | 0.000 |
| Best / Satisfactory | 8/ 11 | 7/ 14 | 9/ 12 | 4/ 8 | 6/ 9 | 8/ 13 | 12/ 15 | | | |

Bold numbers: > 0.8 p < 0.05. Underlined: best estimator. Grey shaded: inflated false positive rate, cf. Table 3-6

*Note: PB com displays the share of studies committing publication bias. PB suc describes the share of studies successfully committing publication bias. p defl. shows the the deflation of the meta-analytical p-value by publication bias.*

# 4. Are really most of our research findings false? An empirical estimation of trends in statistical power, publication bias and the false discovery rate in psychological journals (1975-2017)

A. Schneck. Unpubliziertes Workingpaper

## 4.1. Introduction

Scientific integrity has been shaken by concerns about misconduct that threatens the validity of scientific findings but also the integrity of science in society as a whole (Titus et al. 2008). Clear misconduct like the fraudulent makeup of whole datasets (Markowitz & Hancock 2014) as well as minor forms like (deliberately) erroneous rounding results (Nuijten et al. 2016) or the suppression of research results, i.e. publication bias (Hartgerink et al. 2016), contribute to this problem. Because minor forms of scientific misconduct are much more common than fraud (Fanelli 2009), they enfold a larger leverage on the erosion of scientific integrity in society, shape the public discourse and finally drive (political or medical) interventions that might do more harm than good (Godlee 2012). Monitoring scientific integrity is therefore crucial to describe the salience of the problem but also to offer the possibility of monitoring interventions that strengthen good research practice.

The paper at hand shows the development of statistical power and publication bias – as one form of scientific misconduct (for such a definition see Chalmers 1990) that is widespread among researchers (John et al. 2012: 525) – in papers published in psychology in the period from 1975-2017. Building on the two measures of statistical power and publication bias it is furthermore possible to estimate the share of statistical artefacts on all significant findings (false discovery rate – FDR) and test the claim of John Ioannidis "why most published research findings are false" (Ioannidis 2005) empirically.

## 4.2. Problem

Statistical power,[69] the probability to find a significant effect given a true effect is present (Cohen 1988), is clearly a quality criterion of good science. Prior to the data collection, researchers should determine their sample size adequately in order to keep the statistical power as high as possible but also to keep the included number of cases as small as possible (e.g. to minimize harm to participants or fit budget constraints). From a resource focused perspective low statistical power is inefficient, because with a low sample size random variation in the data disguises the underlying patterns. Since the pioneering study by Cohen (1962: 150f.) there has been no increase in statistical power which is inadequate for most of the studies till these days (Smaldino & McElreath 2016).

Besides being inefficient, low statistical power has also a deterrent side effect: because only a low share of existing effects can be rightly detected the noise of the falsely detected, estimates as defined by the

[69] Also known under true positive rate and 1-β, the complimentary probability of the type II error of failing to detect an effect despite its presence in truth.

significance threshold (e.g. a false positive rate, FPR = 5%)[70] gains leverage. The false discovery rate (FDR) estimates this problem denominating the share of significant effects that are due to false positives and acts as a measure of trustworthiness of found significant results. Jager & Leek (2014) report a quite modest FDR of 14% with no clear time trend. This analysis however suffers from sample selectivity because only *p*-values reported in the abstract of the papers under study are used (for a critique see Ioannidis 2014).

Publication bias, where the publication of research results depend solely on their outcome either in direction (e.g. favouring positive outcomes) or significance (Dickersin & Min 1993: 135), further aggravates this problem by inflating the FPR (for evidence on publication bias see e.g. Hartgerink et al. 2016; Head et al. 2015).[71] Under publication bias the 5%-significance threshold gets inflated drastically up to 60% (cp. the simmulation study by Simmons et al. 2011: 1361). This also increases the FDR dramatically (for theoretical considerations see Ioannidis 2005). In this situation the paradigm of inferential statistics comes to its worst producing just statistical artefacts (Nuzzo 2014). Instead of rejecting theoretical predictions, "undead theories" with no evidential value survive (Ferguson & Heene 2012) and inhibit any scientific progress.

## 4.3. Data & Methods

The dataset at hand builds on all published articles in journals edited by the APA (*American Psychological Association*) from 1975-2017 that are tagged as empirical in the bibliographical database *PsycArticles*. This assures that the strict reporting guideline that first had been released by the APA in 1974 were followed. In total 39,218 articles containing 648,578 test values reported in the main texts could be utilized for the study.

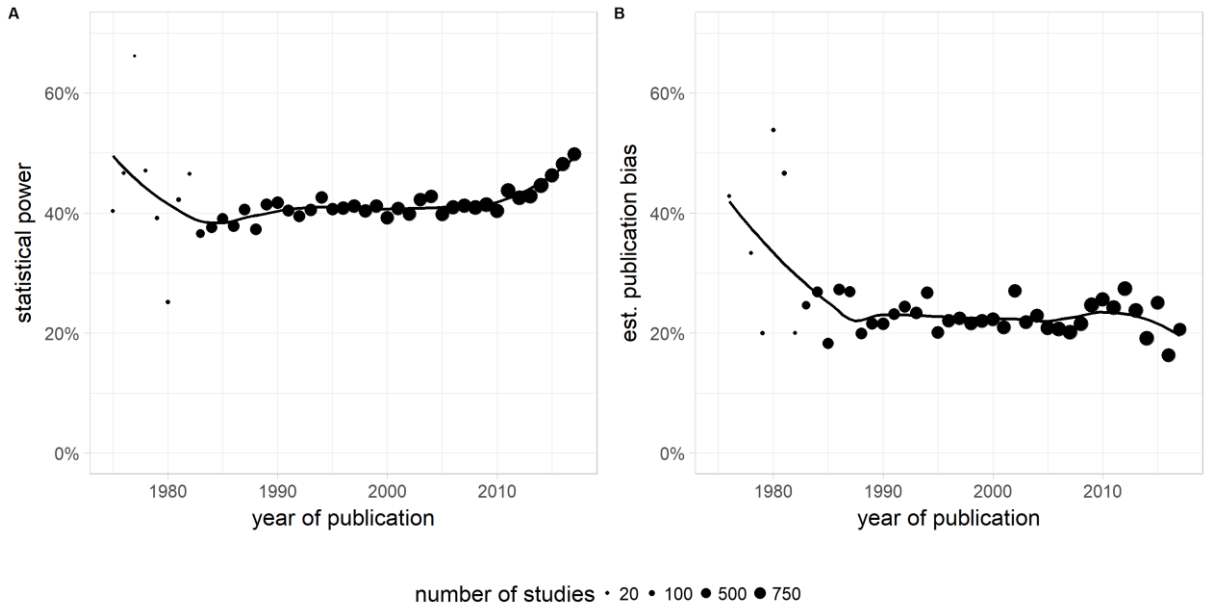
In order to compute the statistical power the true mean effect in each psychological subfield indicated in the included articles was computed. The statistical power was then estimated on the basis of the estimated mean effect for the respective subfield given the variability of the reported effects in the articles. Publication bias was assessed using the caliper test (CT, Gerber & Malhotra 2008b) that builds up on the assumption that in a narrow interval around the significance threshold just significant results (OC) should be as likely as just non-significant results. The publication bias estimate (ρ) is then defined by the share of studies that deviate from the expected uniform distribution in the narrow interval (ρ = OC/0.5-1).[72]

[70] Also known under α or type I error.
[71] The term publication bias is used both to describe the outcome as well as process of selective reporting (see discussion on *p-hacking* and *file-drawer* in the appendix) in order to assure optimal readability.
[72] For a detailed description of the data and methods used see the appendix.

## 4.4. Results

On average, the statistical power over the examined years from 1975-2017 was with 41.7% surprisingly low (Figure 4-1 A).[73] This means that only less than half of the existing effects could be detected with the research designs implemented in the primary studies. However, in the recent years since 2010 there was an upward trend in the statistical power resulting in an average statistical power of 49.8% in 2017. The large variations in power estimates, especially before 1983 is caused mainly by the small number of included studies (<20, denominated by point size in Figure 4-1 A). Overall, only 13.9% of the studies were adequately powered (>80%).

**Figure 4-1 Statistical power (A) and publication bias (B) over time**



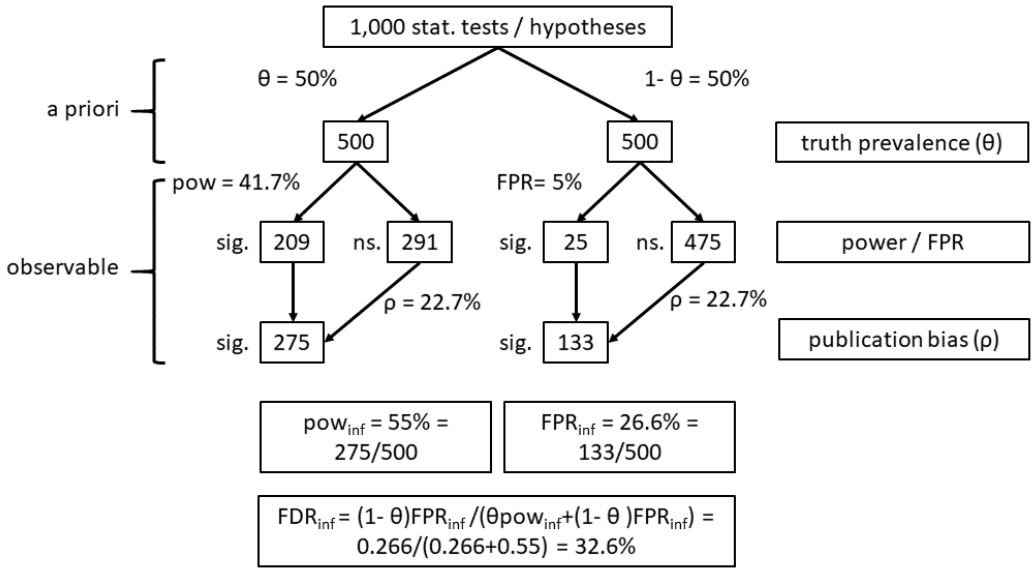

Low statistical power has the deterrent side effect to foster also publication bias practices. Existing effects that could not been found due to a lack of statistical power are repeatedly tested until they are found statistically significant. A first hinge on publication bias was the large share of statistically significant findings (72.9% on the 5% level) in the examined literature, especially in context of the low statistical power. The overall estimated publication bias in the literature was 22.7%, meaning that roughly a quarter of the results that were non-significant in the first place were turned significant. Unlike the statistical power, publication bias varied only in the early years before 1985 (Figure 4-1 B). The larger publication bias before 1985 however could be caused by the low number of estimates. In the recent years there is only a small but unstable trend for a declining publication bias.

More interesting than the isolated description of the statistical power and publication bias is its joint measure, the FDR denominating the share of significant results that are actually false. The FDR is cal-

[73] All estimates are averaged over time and weighted with the number of studies in the respective year.

culated based on the aforementioned two measures, but furthermore requires an assumption on the prevalence of true effects θ, that means how many of the tested hypothesis are true (Vidgen & Yasseri 2016: 2). Because in the examined psychological literature a strong theoretical tradition exists that increases the *a priori* probability of a true effect (Diekmann 2011: 631), θ = 50% is chosen (Figure 4-2).

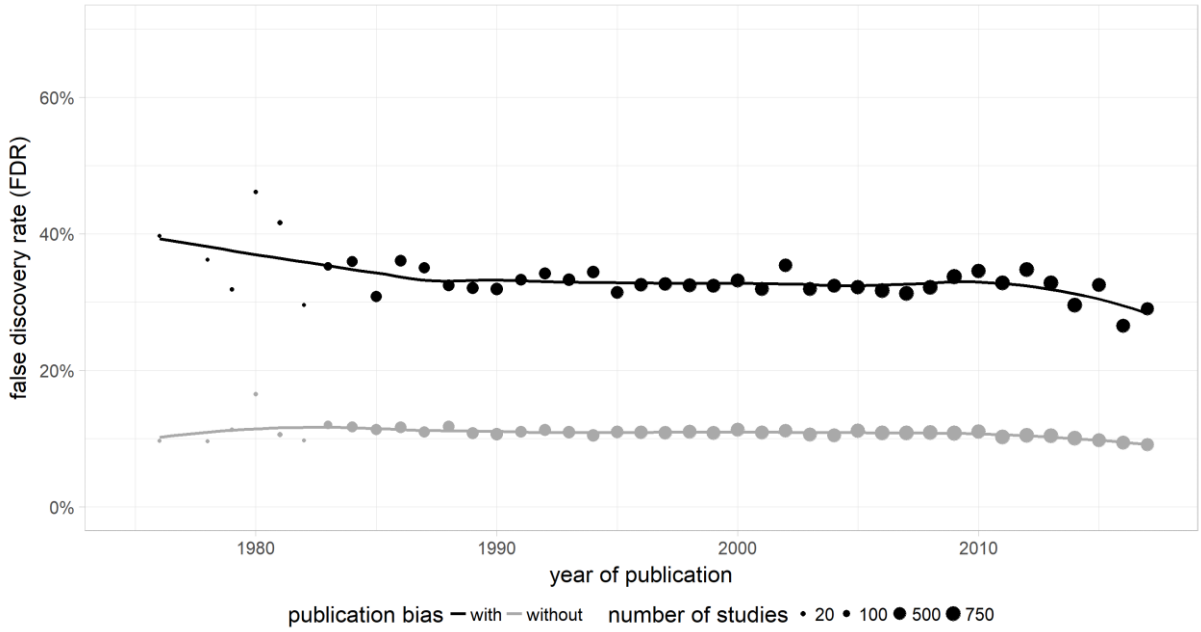
**Figure 4-2 Calculation of the FDR using the calculated statistical power and publication bias**




Based on these theoretically derived assumption, the estimated statistical power is able to identify an existing effect in 41.7%, while 58.3% of the true effects remain undetected. In case of no underlying true effect, following the *a priori* set false positive rate (FPR), 5% of the estimates become significant just by chance while 95% of the estimates correctly find no significant result. In this third step publication bias (ρ) comes into play converting 22.7% of the non-significant into significant results. Publication bias increases thereby both, the statistical power and the FPR. The statistical power increased roughly by one third to 55% ($pow_{inf}$), as more true effects can be detected. The FPR however rose from 5% up to 26.6% ($FPR_{inf}$) by more than factor five. Given the *a priori* probability of 50% of testing a true hypothesis, the FDR inflated by publication bias ($FDR_{inf}$) is defined as the share of the inflated false positive rate ($FPR_{inf}$) on both, $FPR_{inf}$ and the inflated statistical power ($pow_{inf}$). In total, nearly one-third (32.6%) of the significant effects were wrongly detected. This result means that roughly one-third of the studies reporting statistically significant findings are just statistical artefacts, rather substantial findings. Choosing a smaller *a priori* probability θ increases the FDR even further.[74]

[74] For θ = 10% & θ = 20% see the appendix.

**Figure 4-3 False discovery rate (FDR) over time with and without publication bias**




This alarmingly high $FDR_{inf}$ (Figure 4-3 black line) was mostly driven by publication bias as the FDR without publication bias (Figure 4-3 grey line) is on average only 10.7%. In this ideal case, against a common misconception, the FDR is not fixed to 5% but heavily depends on the statistical power, which is in our case way too low. The time trend for both the FDR and $FDR_{inf}$ mirrors the trend of the statistical power and publication bias showing a slight downward trend over time.

## 4.5. Discussion & Conclusion

The article presented evidence on the development of three measures of scientific integrity: statistical power, publication bias and the FDR. Despite a positive development in the recent years, the overall statistical power is still in deficit as well as publication bias is still present with hardly any decline. The resulting $FDR_{inf}$ illustrates the consequences as 32.6% of all significant results are just statistical artefacts.

As Auspurg & Hinz (2017) note, researchers face a social dilemma, because maximizing their own rewards (citations, prizes, etc.) with questionable but significant results benefit their careers as individuals but inhibit scientific progress. To overcome this dilemma Stroebe et al. (2012) argue against the self-correcting nature of science but for clear (institutional) procedures that assure an adequate quality control.

Three specific interventions should be emphasized, that may serve as the abovementioned normative rules: firstly, mandatory power analyses before the study, secondly, preregistration of the research design along with a complete model specification, and thirdly, clear-cut reporting guidelines on how to present statistical tests. While the first intervention focuses on maximizing the statistical power, this may only be effective if real pre-study power analyses are conducted instead of post-hoc power analysis

that justify the chosen setting (Hoenig & Heisey 2001). This can be done especially in in a pre-registration of planned analyses. Pre-registration makes it more difficult to commit publication bias via model selection. Furthermore, publication bias in respect of omitted publications is more likely to be visible in this case (Trinquart et al. 2012). While the first two interventions help to increase power or decrease publication bias, the third intervention enables the monitoring of these indicators. Diagnosing such a scientific crisis, in the case at hand in psychology, was only possible because of existing reporting guidelines that allow for a structured analysis. Other disciplines like economics, medicine or sociology simply miss such guidelines, which are necessary for the analysis at hand, rather than a proof that such problems are limited to psychology and especially APA journals alone.

The recommendations presented are far from new, however the analysis at hand allows to test possible future interventions on their effectivity as the dataset at hand can be updated quite easily due to its automated data collection. Beyond such a merely descriptive monitoring of scientific integrity, future research should also focus on the mechanisms behind potential threats, like author composition or third-party funding that might challenge scientific integrity.

## 4.6. Appendix

The appendix provides further information about the data used and its underlying statistical concepts. Furthermore, parametric regression models for the nonparametric graphical models presented in the article are supplied.

### *4.6.1. Data*

The data was obtained from the database *PsycArticles* and includes all empirical articles published between 1975 and 2017. The year 1975 was set as a starting point because in the year before, 1974, the second edition of the APA (*American Psychological Association*) publication manual was published.[75] For the first time this publication manual offered clear guidelines (in the following called APA-style) on how to report statistical test values (American Psychological Association 1974). The APA-style requires that all test values had to be reported with its respective degrees of freedom. These are used to recalculate the *p*-value on which the statistical conclusions are based on. The APA-style was adopted over time by many journals also beyond the APA in the general social sciences and medicine. To assure that all exported articles fall under this reporting guideline, only articles that were published in APA journals and therefore explicitly fall under the reporting guideline were analysed. Despite this limitation on one major publisher (the APA) a broad variety of different subfields in psychology, starting from clinical psychology (more closely related to medicine) to social psychology (more similar to the social sciences) were covered. This large coverage of different subfields of psychology mirrors the 93 journals that were published or affiliated to the APA in 2018 (cp. American Psychological Association 2018).

In the first step the bibliographical information of all relevant articles was exported. In total, it was possible to retrieve the information on 54,115 articles. In the second step the full texts belonging to the exported bibliographic information were identified via the DOI (Digital Object Identifier). The export routine, written in Python 2.7, was able to export both, html and pdf versions of the full-texts via web-scraping. It was possible to obtain full-texts for 53,861 articles (99.5%). The small loss of articles was caused by either duplicates in the exported biographical data, or non-functional links to the full text. If available, html texts were preferred because the underlying text structure was readily accessible. The remaining texts in pdf format were already converted to text files via OCR (optical character recognition), therefore the time intense image to text conversion was not necessary (Patel et al. 2012). The pdf format however has the problem of misrecognized characters, as the plain text file is converted to an image (pdf) and back to text again in the OCR procedure. This might obscure and bias the text corpus especially if characters are visually very similar or do not exist in the encoding set (e.g. $\chi$ in ASCII).

[75] The introduction of the APA-style can be traced in the percentage of articles that include at least one exportable test value. For the year 1974, the year before the introduction of the 2nd edition of the APA-style, only 7.3% of all articles contain test values whereas this share rises drastically to 33.9% in 1975 although far from the nearly 80% covered articles in the complete time period (cp. Table 4-1).

Despite the problems with automatic extraction in pdf full texts, searching pdf texts is inevitable to assure a constant coverage of texts over time. Until the year 1986 all texts retrieved were only available in pdf format. In the following years this share drastically drops nearly to zero. An exception are the most recent years 2016 (42.2%) and 2017 (10.6%) where a substantial part of the texts could only be accessed in pdf format. In the third step the statistical test values reported in the articles were exported. Here the central advantage of using APA journals comes into play. Due to the strict reporting guidelines, all reported results had to be mentioned in the text along with its complementing test statistic. This could be done directly in the text or in accompanying tables (American Psychological Association 1974,1983,1994,2001,2010). Although the in-text presentation of test values is strongly recommended as a primary option also tables are allowed in the APA-style, especially for more complex designs (American Psychological Association 1974: 39). In the 6th edition of the APA-style tables and figures are recommended for articles reporting a larger number of results (>4 and >20 respectively American Psychological Association 2010: 116). In contrast to the in-text reporting it was not possible to extract the results reported in tables due to diverse reporting styles (e.g. standard errors or *p*-values below or beside the reported estimates).[76] The in-text presentation of central results in contrast is consistent in the APA-style and varies only slightly over time (e.g. *N* reported along $\chi^2$). It was therefore possible to export the most common test statistics ($F$, $\chi^2$, $r$, $t$, $z$).

The reported test statistics usually follow the same reporting pattern: in the first section the test statistic itself is denominated, followed by the respective degrees of freedom and the resulting test value (e.g. $F(1, 4) = 3.25$). An exception of this reporting style is the mandatory sample size (*N)* after the degrees of freedom for the $\chi^2$-test until the 6th revision of the APA-style (American Psychological Association 1974,1983,1994,2001). To extract these text patterns all articles were searched for strings matching the reporting pattern of the APA-style using regular expressions. In order to be as error tolerant as possible, small deviations from the publication manual were also allowed. As mentioned above, the pdf texts bear the problem of misrecognized characters due to the image to text conversion (OCR). This concerns italic characters used for the denomination of the test statistic and especially, because of their similar typographical shape, *t* and *F*. In order to overcome this limitation, the most frequent mistakes were detected and corrected in the export routine (e.g. $\chi^2$: x2, x², X², X2, …). This approach allowed to extract as much valid test statistics as possible while minimizing random noise that only follows a similar pattern. In total 736,835 test values could be retrieved via the text mining procedure (cp. Table 4-1).

[76] For robust results on the missed tabular results see chapter 4.6.4.

**Table 4-1 Dropouts during data cleaning**

| | Estimates | Percent left (%) | Articles | Percent left (%) |
|---|---|---|---|---|
| Export | 736,835 | 100.00 | 53,861 | 100,00 |
| *z*-value computable | 736,596 | 98.39 | 42,234 | 78.41 |
| Effect size *d* computable | 651,048 | 86.96 | 39,266 | 72.90 |
| Other exclusions | 648,578 | 86.63 | 39,218 | 72.81 |

### *4.6.2. Operationalisation*

The statistical test values that had been extracted using the abovementioned export routine had to be transformed into two different metrics: a common test statistic in order to estimate publication bias and a common substantially interpretable effect size metric that is needed to calculate the statistical power. As the effect size metric that is needed for the computation of the statistical power builds on the test statistic upon which the estimate of publication bias is based, publication bias is introduced first.

***Publication Bias***

Publication bias is the selective publication of research results in favour of effect sign or significance (Dickersin & Min 1993: 135). This means that statistically significant results have a larger probability of getting published compared to non-significant results. Publication bias is mostly examined in meta-analyses concerning only one single effect (e.g. Doucouliagos & Stanley 2009 on the effect of minimum wage on employment). This is however not possible for research areas with no or only small cumulative research. The caliper test (CT) developed by Gerber and Malhotra (2008a,b) allows to detect publication bias under heterogenous effects using only the *z*-values (*z*) in a small band (called caliper, *c*) around a prespecified significance threshold (*th*).

In order to transform all reported test values into one common test statistic (*z*) all test values were first transformed into their according *p*-value and afterwards into the *z*-value. Although the APA-style demands a strict reporting format, not all test values were reported accurately. In total, for 736,596 (98.39%) of the test values a *p*-value could be computed. The small number of estimates/studies that drop from the analysis was caused either by missing values on the test statistic or its according degree(s) of freedom, or by infinite test values. This is the case especially for implausible large test values, which result in a virtually zero *p*-value and therefore an infinite *z*-value. In order to compare two-sided (*t*, *z*, *r*) and one-sided test statistics ($F$, $\chi^2$), two-sided *p*-values were only computed for all two-sided tests. All *p*-values were then transformed to *z*-values to fulfil the requirements of the CT.

As the CT ignores *z*-values outside the narrow range (*th*), it is not affected by effect heterogeneity that occurs when examining the whole test value distribution (for simulations see Schneck 2017). Effect heterogeneity in this context means that the studies differ not only by random variation but by substantial

differences that may be explained by different research topics or research designs (Higgins & Thompson 2002). In the intervals around *th* the upper interval that contains results that are just statistically significant (over-caliper, $x_z = 1$) should be as likely as results that slightly miss statistical significance (under-caliper, $x_z = 0$). The CT therefore follows the logic of a regression discontinuity design (Lee & Lemieux 2010) that allows to examine a treatment effect (in this case publication) that is caused by an assignment variable (in this case statistical significance). A limitation on the narrow range is necessary because the functional form of the entire distribution is unknown. The only characteristic of the *z*-statistic that is known is its continuity over its complete range, meaning there should not be any abrupt jumps in the distribution. The narrower the bandwidth of the examined interval *c* around *th* is set the more likely the assumption of an equal distribution due to continuity is met.

$$x_z = \begin{cases} 0 \; if \; th - c * th < z \leq th \\ 1 \; if \; th < z < th + c * th \end{cases}$$

Gerber and Malhotra (2008a,b) use a 5%, 10%, 15% and 20% bandwidth *c*. Especially the largest bandwidths may be too large, because in the 20%-CT, the 10% significance threshold is overlapped and publication bias around the 5%- and 10%- significance level may cancel each other out. In simulations comparing the aforementioned alternative tests on publication bias a bandwidth of 5% the size of *th* (0.098 in case of the 5% significance threshold *th* = 1.96) proved to be adequate to absorb fluctuations in the underlying effect distribution but also include as much test values as possible in order to assure an adequate statistical power of the CT (Schneck 2017). Although the statistical power of the CT is not a great concern in the analysis at hand, because a large number of *z*-values could be examined, another reason to choose wider intervals is the reporting accuracy of the test values reported in the primary studies that are usually reported with two decimal places and add further noise to the distribution. In the main analysis *c* is therefore set to 5%, while the results for the wider 10%-CT are reported in the robustness section below.

The CT presented so far only allows to test for publication bias. When estimating the prevalence rate of publication bias it is important to keep the limitation of the test on just a small bandwidth of test values in mind. For this limited set of just significant or non-significant effects the prevalence estimate of publication bias (ρ) is the quotient of the observed share of values in the over caliper $P(x_z = 1)$ and the expected over-caliper of 0.5 (uniformly distributed). If an over-caliper is slightly overrepresented with 0.6 a bias of 0.2 or 20% is present.

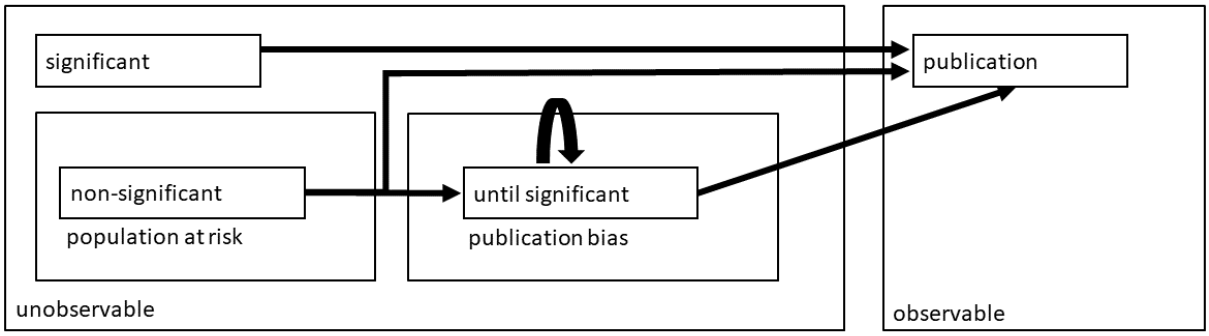
$$\rho = \left(\frac{P(x_z = 1)}{0.5}\right) - 1$$

Before inferring from the small set of values included in the CT on the whole test value distribution and therefore to the general prevalence of publication bias, a few conditions have to be clarified. First, it should be considered that the risk of publication bias only comes into play if a researcher is confronted with a non-significant result in the first place (cp. Figure 4-4). If already a significant result was obtained

there is simply no need for publication bias practices. In other words, the researcher is not at risk committing publication bias. In contrast if there is a non-significant result a researcher faces two options, either try to publish the result irrespective of its outcome or start searching for significant results. This search for significant results by publication bias practices can be done either by data mining (this includes e.g. fishing Humphreys et al. 2013; HARKING Kerr 1998; p-hacking Simmons et al. 2011), or via a completely new data collection (also called file drawer Rosenthal 1979).

All publication bias procedures are however not a guarantee for success (if no severe data manipulation comes into play). Under publication bias several rounds of reanalysis or recollection of data may be necessary until a significant result is reached. It may therefore be possible that researcher decide to publish even an undesired non-significant result because the search or production costs of a significant result outrun its returns. It may also be possible that a researcher engages in publication bias practices and achieves a lower *p*-value that meets a softer significance threshold (e.g. 10%). Committed publication bias therefore may produce both, the desired significant result, but it is also possible that non-significant results survive that suffer nonetheless from publication bias in the sense of an inflated effect size. All the aforementioned variables are unobservable except the published research results. The estimate of the publication bias prevalence as introduced above is therefore only able to determine the probability that a researcher with a non-significant result (population at risk) successfully improves his or her research results by the means of publication bias. Although several rounds of publication bias may be possible and are important for the specific bias in the literature, only the prevalence of publication bias in the literature is estimable. Therefore, the chosen operationalisation is rather a conservative than inflated measure.

**Figure 4-4 Process of publication bias**



***Statistical Power***

Beside the prevalence of publication bias also the statistical power of the estimates had to be computed. Before computing the statistical power it was necessary to estimate an adequate effect size upon which the power calculations are based on. The effect size *d* was preferred to the alternative correlation coefficient (*r*) because it is positively unbound [0,∞] and therefore more appropriate to estimate the mean effect than the bounded effect size *r* [0,1]. In this second step after the calculation of *z*-value some test values could not be transformed into *d* as a common effect size metric because sample sizes were not reported along with the test value (some $\chi^2$- and all *z*-values). For the transformation from the remaining test statistics to *d* the following formulas were used in Table 4-2 (cp. Friedman 1982: 524).

**Table 4-2 Transformation formulas for test statistics in Cohen's d**

| **Test value** | **Effect size *d*** | **Standard error of d ($\sigma_i$)** |
|---|---|---|
| $\chi^2 \; if \; df = 1$ | $2\sqrt{\frac{\chi^2}{N}}$ | $d/z$ |
| $\chi^2 \; if \; df > 1$ | $2\sqrt{\frac{\chi^2}{N-\chi^2}}$ | „ |
| $F$ | $2\sqrt{\frac{df_1 F}{df_2}}$ | „ |
| $r$ | $\left\lvert\sqrt{\frac{4r^2}{1-r^2}}\right\rvert$ | „ |
| $t$ | $\left\lvert\frac{2t}{\sqrt{df}}\right\rvert$ | „ |

Whereas *t*-tests and correlation coefficients (*r*) may yield negative values of *d* this is not the case for *F*- and $\chi^2$-value, as those values are by definition only positive. Furthermore, the direction of the effect may differ by different definition of test and control group for two-sided tests. Therefore, the absolute effect size was computed for all test statistics.

In total it was possible to calculate a power estimate for 86.96% (651,048) of all exported test values covering 72.90% (39,266) of the articles. In a third step all infinite values on *z*, *d* and implausible values[77] in the extracted test values were excluded. Finally, the analysis is based on 648,578 test values (86,63%) reported in 39,218 articles (72.81%). All of these estimates were used for the main and supplemental analyses.

On the basis of the converted effect size estimates it is possible to compute the statistical power. The statistical power (Cohen 1988: 1) is defined as the share of truly (significant) detected effects on the overall number of true effects (cp. first column Table 4-3). In research practice the statistical power is

[77] Test values with more than 6 digits (e.g. 9999.99), as well as a larger first degree of freedom than the second degree of freedom for *F*-tests.

limited because a large statistical power near 100% given a false positive rate of 5% demands exorbitant sample sizes, a statistical power of 80% is therefore suggested by Cohen (1988: 56) as a trade-off.[78] The significance-level is set by the researcher in advance. The most commonly used significance level is at FPR = 0.05 or 5% (Cohen 1994; Labovitz 1968). A 5% significance threshold therefore allows 5% significant results just by chance if no effect is present in truth (cp. Table 4-3 second column).

**Table 4-3 Truth table**

| | | **Truth** | | |
|---|---|---|---|---|
| | | Effect | No effect | |
| **Estimator** | Effect | True positive (TP) | False positive (FP) | False discovery rate (**FDR**) = FP / (TP + FP) |
| | No effect | False negative (FN) | True negative (TN) | |
| | | statistical power (**pow**) = TP / (TP + FN) | False positive rate (**FPR**) = FP / (FP + TN) | |

To estimate the statistical power three parameters are necessary: the significance threshold, the underlying true effect ($\mu$) and the variability of the effect size ($\sigma_i$). The statistical power for a *z*-test (based on a normal distribution, Φ) is then defined by the probability of a significant finding detecting the true effect μ at a pre-specified level (e.g. FPR = 0.05) given the variability $\sigma_i$ of the effect size in the specific study (*i*). In the study at hand, the statistical power of a two-sided test is computed. Two thresholds, one on the left hand and one on the right-hand side are necessary. The right-hand side defines the probability of finding a significant result with a negative sign (where $z < -1.96$), determined by the inverse normal distribution $\Phi^{-1}(0.025)$. The left-hand part of the equation defines in contrast the probability of finding a significant result with a positive sign (where $z > 1.96$, determined by the inverse normal distribution $\Phi^{-1}(0.975)$). As a consequence of the *a priori* specified significance the threshold of 5% determines the minimum power of a study result.

[78] The exact opposite situation is desirable in case of a false positive (right lower corner in Table 4-3). In this situation an effect is suggested by the estimator while none is present. The false positive rate is defined by the share of detected significant tests on the overall number cases in which no effect is present (cp. Table 4-3 second column).

$$pow_i = \Phi\left(\Phi^{-1}(0.025) - \left(\frac{\mu}{\sigma_i}\right)\right) + \left(1 - \Phi\left(\Phi^{-1}(0.975) - \left(\frac{\mu}{\sigma_i}\right)\right)\right)$$

For the equation above it is crucial to estimate both parameters $\sigma_i$ and μ. The first is defined straightforward as the standard error of the effect in each study (as defined in Table 4-2). The underlying true effect μ however is not directly estimable. The central problem of a true effect is that it is unknown and therefore examined in the study. There is however the possibility to compute an expected value of the underlying true effect by using the mean effect size of similar studies. All available estimates that share certain characteristics are then pooled together in a meta-analysis (Borenstein 2011; see Ioannidis & Trikalinos 2007: 246 for a similar method to calculate the statistical power). The most crucial step is to determine a set of studies that is homogenous enough to pool effect sizes.

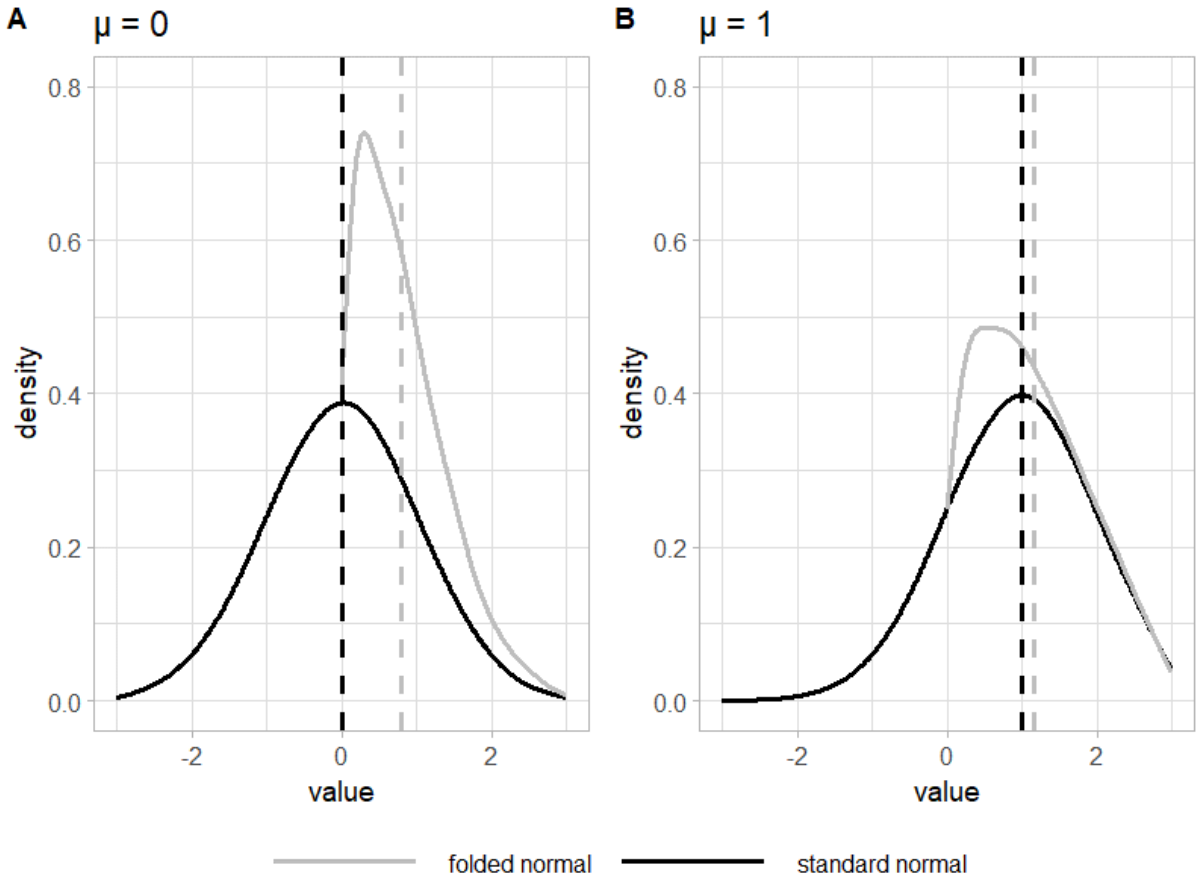
Because of excess heterogeneity of effect sizes across the different articles, subgrouping the articles is inevitable to resolve at least some of the effect heterogeneity. The data used poses however two different problems in calculating the mean effect size: As always in meta-analysis the mean effect has to be justified by theoretical reasons to avoid a too large effect heterogeneity as the coined saying of comparing apples and oranges (Sharpe 1997) suggests. In the used bibliographic database, the PsycINFO Classification Code (PIC) was used that classifies the included literature in up to 157 categories.[79] Although far from a perfectly homogenous meta-analysis that examines only one single effect this approach should reduce the effect heterogeneity yielding a more accurate mean effect estimate for each subfield in psychology as indicated by the PIC.

Estimating just the mean underlying effect was however not possible, because only positive effects sizes (*d*) occur in the data.[80] Existing negative effect sizes are therefore missing, leading to an upwardly biased mean due to the artefactually skewed distribution. Whereas a skewed effect distribution would not be a problem per se if it mirrors the factual effect distribution in the present case it is only an artefact of the implemented transformation formulas leading to a positively biased true mean effect.

[79] The categories are stratified in three levels containing 22, 84 and 51 categories in descending order from the most general category (e.g. human experimental psychology) to the middle category (e.g. sensory perception) to the narrowest category (e.g. visual perception). In order to maintain as detailed categories as possible the existing codes were used instead of pooling the sub-levels to the next level.

[80] Transforming *F*- and $\chi^2$-values in *d* does not yield a negative sign as the tests itself cover only a positive range. Nonetheless, negative effects may occur but could not be recovered without any information on an effect metric.

**Figure 4-5 Bias of folded normal distribution (dependent on the true mean μ)**




These properties are characteristic of a folded normal distribution that is defined as the absolute value of the underlying (mother) normal distribution. Figure 4-5 A shows the normal distribution (solid black lines) with a mean of zero (dashed vertical black line) as well as its accordingly folded distribution (grey lines).[81] Because the mother normal distribution centres symmetrically around zero, one half of the distribution changes its sign in case of the folded normal distribution from a negative value to a positive one. The result of this fold is a biased mean estimate, while the true mean is zero, the mean of the folded normal distribution is 0.80. The folded normal distribution therefore clearly produces an upwardly biased estimate of a true effect of zero ($\mu = 0$). If the true mean differs from zero (Figure 4-5 B), a smaller fraction of the distribution is affected by the fold. In this situation the folded normal distribution approximates the normal distribution (Tsagris et al. 2014: 20). In case of an underlying effect of one ($\mu = 1$) the bias is lessened (1 vs. 1.16) but still present. In psychology, the field under study, the first scenario, where a large share of values is affected by the fold seems plausible because in the literature low to medium effect sizes are common (Bosco et al. 2015).

In order to recover the true mean two different solutions were used: on the one hand the folded normal distribution and its according transformation rules (Leone et al. 1961), on the other hand a weighted

[81] This special case in which the distribution is folded at its mean is also called half normal distribution (Tsagris et al. 2014).

least square (WLS) regression approach weighting by the inverse variance of the study effect size. The first strategy uses the fast computational routine by Chatterjee & Chakraborty (2016). In a two-equation system the true variance as well as the true mean were recovered. The approach however has the problem that it is very sensible to outliers that invalidate the estimation strategy by inflating the estimated variance of the folded distribution. Therefore, an outlier correction (*dffits*<1) was used in order to avoid the estimation problem. The second approach, WLS, gives estimates with a high precision (small variance) a higher weight. This approach is standardly known in the meta-analyses literature as the fixed-effect model (Borenstein 2011). Originally, this approach is used to tackle the problem of heteroscedasticity, as more precise studies exhibit less random variation around the true mean. In the commonly used data however no folded distribution occurs, meaning that the variation with positive and negative sign chancels each other out leading only to less precise than biased estimates. The WLS model also accounts for the bias introduced by the folded distribution: More extreme estimates that occur for less precise studies account for most of the bias and were therefore downweighted. The WLS model has also the advantage of using the full sample, whereas the approach of Chatterjee & Chakraborty (2016) used only an outlier corrected sample.

Both estimated mean effects of a respective field of study had to be recovered for each PIC area. As some of the articles were assigned up to two PIC areas, each study estimate was weighted with the inverse of the number of assigned PIC areas. An estimate with two assigned PIC areas therefore received a weight of 0.5 in each of the two PIC areas.

The mean effect size (*d*) that represents the underlying true effect size for the power analysis was computable for 150 of the possible 154 PIC areas present in the data.[82] The effect sizes (reported in *d*) show large variation from 0 to 1.5 in case of the recovered mean of the mother normal distribution and 0.1 to 1 in case of the WLS estimate. Despite the large variation in effect sizes the mean and median effects of the respective PIC areas could be classified as medium in size (folded normal: 0.56, 0.55; WLS: 0.43, 0.47) according to the effect size terminology of Cohen (1992).

***False Discovery Rate***

The false discovery rate (FDR) defines the share of significant false effects on all significant effects (first row in Table 4-3). In a first step, the consequences of publication bias on the false positive rate as well as the statistical power have to be modelled. Assuming a significance level (FPR = 0.05), publication bias (ρ) transforms some of the true negative results (TN) to false positive ones (FP) leading to an inflated false positive rate ($FPR_{inf}$).

$$FPR_{inf} = FPR + \rho * (1 - FPR)$$

[82] In case of the WLS estimate, the importance of the calculation of a PIC specific mean can also be shown in a meta-regression model with PIC dummies, again weighting with the inverse variance. In this specification the model explains about 8.7% of the variance $F(153, 712680) > 100, p < 0.001$ and therefore allows to separate different underlying effects in different scientific disciplines.

The same is also true for the statistical power, as non-significant results that truly exists become significant by publication bias. Therefore, also the share of detected true effects increases with publication bias ($pow_{inf}$).

$$pow_{inf} = pow + \rho * (1 - pow)$$

The FDR is then defined as the relative frequency of false positive results (FP) compared to the overall number of positive results (cp. the first row in Table 4-3).

$$FDR = \frac{FP}{FP + TP}$$

As the absolute cell frequencies in Table 4-3 are unknown, the absolute number of TP and FP have to be derived by the given column percentages (FPR, pow). In addition, an assumption has to be made *a priori* in order to specify the truth prevalence, meaning the probability that a tested hypothesis is true (θ). The FDR is then recovered by weighting with the *a priori* information on θ. If an equal share of the hypotheses is true or false, FPR and pow receive equal weights. If, however the share of true hypotheses gets larger, pow receives more weight leading to a low FDR. The opposite is true for a small share of true hypotheses. In this case, the FPR is weighted more strongly leading to a higher FDR. $FDR_{inf}$, the FDR inflated by publication bias, is computed with $FPR_{inf}$ and $pow_{inf}$ respectively.

$$FDR = \frac{(1 - \theta)FPR}{\theta pow + (1 - \theta)FPR}$$

The *a priori* set truth rate θ depends heavily on the scientific discipline, in genetics the truth rate is perceived very low due to the exploratory nature of most research (cp. Ioannidis 2005). In theoretically founded disciplines like in psychology there should be a higher *a priori* probability that the hypothesis building on this theoretical tradition is true. In the study at hand a 50% truth rate is chosen in the main article, however the results of 10% and 20% are reported in the sections below.

### *4.6.3. Methods*

The main aim of the article at hand is to estimate time trends in publication bias, statistical power, and the FDR. To this end, the data had to be averaged for each publication year. With this aggregated data it is possible to identify changes in all three measures over time. Especially for the publication bias as well as the FDR, this would not be possible using estimate level data because these results are only estimable at the aggregate level. Therefore, only 43 observations, each for one year (1975-2017), remained. The sample of analysis was furthermore restricted because each estimate should at least include five publications (*K*). This was unproblematic for the estimates of the statistical power because all reported estimates in the respective publication year could be used. However, for the estimate of publication bias this was a problem, from the smallest 5%-CT two years (1975 & 1977), and in the 10%-CT the year 1977 had to be excluded from the analysis.

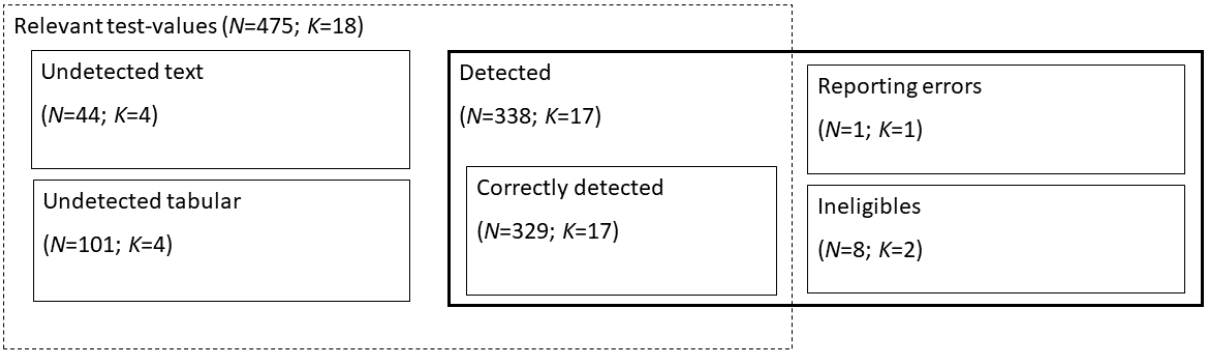
The results in the main article were presented graphically via a LOESS regression (Cleveland 1979). The time trend was additionally modelled in a parametric linear regression, reported in the robustness section, including either a linear time trend or a more flexible form using different time periods (<1985, 1985-1994, 1995-2004, 2005-2015, 2015-2017). In order to account for the greater variance of the estimates in years with only a small number of included publications, *K* was used as weight in a WLS model.

### *4.6.4. Robustness checks*

#### *Comparison with manual data extraction*

In order to test the robustness of the automatically exported results, 20 randomly selected articles were manually coded by the author. In a first step it was checked if the export algorithm missed any APA-style conform test values or detected them wrong. In total, 338 test values were identified in 17 articles, while 3 articles reported no APA-style conform results (cp. bold box in Figure 4-6). Although all these results follow the APA-style not all results are reported correctly which leads to falsely detected test values (outside the dashed box in Figure 4-6). Those can be split up in two categories, erroneous reported results (for one test value, a comma was used as a thousands separator) as well as ineligible results. Results are ineligible if only the least significant among many results is reported (e.g. minimal $t(11) = 2.73$). The reported minimal test values may therefore introduce downward bias and should be excluded ($N = 8$, in $K = 2$), nonetheless the report of all estimates including the unreported larger test values in the primary articles would be the best solution. In total only 9 estimates (1 false export, 8 ineligibles) that make up 2.66% of all detected estimate were false positive detections due to reporting inconsistencies in the primary articles.

**Figure 4-6 Robustness analysis data extraction process**




Despite the importance to reduce false positive detections, it is as well important to detect as many relevant test values as possible. Although the export of test values was limited on reported test values that follow the APA-style, also results reported in a non-APA-style format are equally important for the analysis, but hard to extract automatically. Those non-APA-style compliant results split up into two categories, results that were reported in the main text of the article but in a non-APA-style format and results that are reported in tables that follow no clear reporting style. 44 test values reported in the main

text of 4 articles could not be detected via the export algorithm (e.g. the report of multiple test values with common degrees of freedom: *Fs*(1, 29) = 4.01 and 4.03). Taking these results into account, 88.20% of the relevant in-text test values could be exported by the algorithm.

The limitation on test values reported in the main text was however only a practical solution in order to allow an automatic export procedure, as tabular results follow no clear reporting guideline. Nonetheless, tabular results are also of importance as the conclusions of the paper at hand should nevertheless hold to all reported test values in the examined psychological journals. In total, 101 test values presented in 4 articles used a tabular presentation and could therefore not be exported. The distribution of tabular result was quite skewed and one single article provided up to 83.25% (84 estimates) of all tabular results. This article in total reported with 118 test values by far the largest number of estimates. 69.41% of all estimates, reported either in text or in tables, could be extracted by the algorithm. The used export routine has therefore two advantages: a very low false positive rate of 2.66% as well as a high true detection rate (69.41% for all test values or 88.20% for test values reported in-text only).

Albeit a high share of truly detected estimates reduces the risk of bias introduced by a selective test value sample, the differences between the *p*-value distribution of the included and missing test values that are reported either in text or in tables are examined. The *p*-value distribution was favoured to the *z*-value distribution because of its bounded nature between 0 and 1, that is not prone to extreme values. A direct comparison of the *p*-value distribution of truly detected values vs. falsely non-detected values (tabular or in text) was however not possible because the articles differ in their different research questions that pose different *a priori* probabilities of the hypothesis being true. Therefore, the difference between truly algorithmically detected and falsely non-detected test values is only estimable within each primary article. Holding the study characteristics like the *a priori* probability of a hypothesis being true constant, there should not be any difference between the detected and not detected test values left over. To this end, a fixed-effects linear regression model was estimated that builds upon 7 studies with variation in respect of detected and not detected test values. There was no significant difference between the detected results and the falsely non-detected results reported in text that were missed by the algorithm $t(266)=-0.948$; $p=0.344$. A slight difference that was also statistically significant on the 10% level was however observed in the *p*-value distributions of detected and not detected tabular results $t(266)=-1.716$; $p=0.087$. When looking additionally at the share of significant values, that is especially of interest for the test on publication bias, no differences were observed for the in-text $t(266)=0.338$; $p=0.736$ as well as tabular results $t(266)=1.226$;$p=0.221$.

***Comparison with existing export algorithm***

The abovementioned study by Hartgerink et al. (2016, for the data documentation see Hartgerink 2016) used a similar approach to export test values from psychological journals following the APA-style using the R package *statcheck* (Epskamp & Nuijten 2018). Therefore, it is possible to compare the number of exported results of both algorithms in order to minimize the risk of missing relevant text values. In

contrast to the algorithm in the study at hand, *statcheck* exports also *p*-values (that should be reported along with the test values according to the APA-style). This has the advantage that also misreported *p*-values that could be identified by a mismatch between the reported *p*-value and test value can be detected (Nuijten et al. 2016). A downside of this approach however is the increased error probability in the export as another structured text part has to be extracted in addition to the test value.

In order to keep the following comparison as similar as possible, only studies from 1985-2016 that are included in the data of Hartgerink (2016) were compared. For the algorithm used in this study it was possible to export 695,376 test values and transform it into its according *p*-values.[83] This exceed the export of Hartgerink (2016) by around 33%. But besides only exporting just more test values from the articles, it was even possible to include also 8% more articles which leads to a more holistic picture. Albeit the average *p*-value of the exported results was close (0.097 vs. 0.098), differences in the distribution between exported and missed test values may distort the results. Besides *statcheck* has a clear value in uncovering misreporting, it is too restrictive in the export and misses too many test values if the reported *p*-value is not of interest as in the study of Hartgerink et al. (2016).

***Results***

In the following section the graphical models reported in the main article are supported with its according parametric models. In order to allow a more flexible modelling of the time trend, both, a categorical specification and a linear time trend, was estimated. The categorical specification divides the data in five categories containing ten years. The only exception is the reference category (2015-2017) that covers only the three most recent years. In order to make the interpretation of the regression intercept as comparable as possible across the two specifications, the study year was set to zero for 2017. The intercept can be therefore interpreted as the current state of the statistical power, publication bias or FDR respectively.

The graphical findings from the main articles that suggested an increasing statistical power over time, especially in the recent years were confirmed by the parametric regression models in both the categorical (Table 4-4) and the linear specification (Table 4-5). The overall level of the statistical power however was still remarkably low with below 50% in the recent years. If the folded normal distribution is used calculating the statistical power instead of the WLS-model (Table 4-4 model 2) the overall level of the statistical power increases by around 10 percentage points (ppts) up to 60%. The increasing time trend shows exactly the same pattern in both the WLS and folded normal specification. Besides the average level of the statistical power also the share of adequately powered studies (>80%) was modelled over time (Table 4-4 & Table 4-5 models 3 & 4). In the WLS model for the recent years only up to 20% of the studies are powered adequately. Despite the pattern of models 3 & 4 is strikingly similar to models 1 & 2, model 4 that builds on the folded normal distribution showed up to 38% adequately powered

[83] In order to assure comparability also *z*-values were included in the comparison of *p*-values.

studies. Although differences in the baseline could be observed, the increasing trend of the statistical power could be confirmed in all models. Despite substantial differences in the share of adequately powered studies, the statistical power in all models was far too low comparing it to the benchmark of 80% as proposed by Cohen (1988: 56).

Before turning to the results on publication bias the share of significant findings reported in the literature (Table 4-4 & Table 4-5 model 5) was examined, as the literature reports an increasing share of significant findings that are interpreted as a first hint on publication bias (Fanelli 2010; Sterling 1959; Sterling et al. 1995). In the dataset at hand the share of significant results was around 66% in the recent years. If the time trend however is considered, the share of significant results rises up to 77% in the 1980s with a decline starting at the 2000s. In contrast to the literature, no trend of increasingly significant results in the literature was found.

**Table 4-4 Statistical power and significant effects on categorical study year**

| | | | **Statistical Power** | | **Significant effects** |
|---|---|---|---|---|---|
| | WLS | Fold | Adequate (WLS) | Adequate (Fold) | $p < 0.05$ |
| | (1) | (2) | (3) | (4) | (5) |
| Constant | 0.480*** | 0.596*** | 0.200*** | 0.375*** | 0.663*** |
| (ref. 2015-2017) | (0.008) | (0.010) | (0.008) | (0.011) | (0.007) |
| 1975-1984 | -0.100*** | -0.154*** | -0.091*** | -0.158*** | 0.112*** |
| | (0.016) | (0.021) | (0.016) | (0.022) | (0.013) |
| 1985-1994 | -0.079*** | -0.118*** | -0.077*** | -0.120*** | 0.113*** |
| | (0.010) | (0.012) | (0.009) | (0.013) | (0.008) |
| 1995-2004 | -0.071*** | -0.085*** | -0.065*** | -0.094*** | 0.076*** |
| | (0.009) | (0.012) | (0.009) | (0.012) | (0.008) |
| 2005-2014 | -0.061*** | -0.065*** | -0.063*** | -0.077*** | 0.045*** |
| | (0.009) | (0.011) | (0.009) | (0.012) | (0.007) |
| *N* | 43 | 43 | 43 | 43 | 43 |
| $R^2$ | 0.685 | 0.757 | 0.668 | 0.731 | 0.876 |

*Note:* *p<0.05 **p<0.01***p<0.001; standard errors in parentheses

**Table 4-5 Statistical power and significant effects on linear study year**

| | Statistical Power | | | | Significant effects |
|---|---|---|---|---|---|
| | WLS | Fold | Adequate (WLS) | Adequate (Fold) | $p < 0.05$ |
| | (1) | (2) | (3) | (4) | (5) |
| Constant | 0.446*** | 0.570*** | 0.165*** | 0.339*** | 0.676*** |
| (ref. 2017) | (0.006) | (0.006) | (0.006) | (0.007) | (0.003) |
| + 1 Year | 0.002*** | 0.004*** | 0.002*** | 0.003*** | -0.004*** |
| | (0.0003) | (0.0003) | (0.0003) | (0.0004) | (0.0002) |
| *N* | 43 | 43 | 43 | 43 | 43 |
| $R^2$ | 0.480 | 0.740 | 0.393 | 0.661 | 0.894 |

*Note:* *p<0.05 **p<0.01***p<0.001; standard errors in parentheses

Despite a large share of significant values may be a first hint for publication bias, a formal test is necessary. In the following, the bias estimates based on the 5%- and 10%-CT are presented. The results in Table 4-6 (categorical) and Table 4-7 (linear) show that around 20% of the non-significant results were omitted and replaced by significant results due to publication bias. This 20% publication bias estimate has a huge impact especially on the false positive rate. Instead being set on the conventional 5%-significant level, the level is drastically inflated, as 20% of the 95% non-significant findings were altered towards a significant result and resulted in an inflated false positive rate of 25% or an inflation factor of 5. Both CTs produced similar results while the larger 10%-CT showed slightly stronger evidence on publication bias. In contrast to the results on the statistical power, the results on publication bias show no decisive time trend on neither the categorical nor the linear specification. An exception was the slightly higher bias in the earliest period from 1975-1984 for the 5%-CT and the declined publication bias in the most recent period in the 10%-CT. All in all, there is robust evidence for a mostly time invariant substantial publication bias in the psychological literature that shows at most only slight signs of decline.

**Table 4-6 Publication bias on categorical study year**

| | 5%-CT | 10%-CT |
|---|---|---|
| | (1) | (2) |
| Constant | 0.206*** | 0.239*** |
| (ref. 2015-2017) | (0.015) | (0.013) |
| 1975-1984 | 0.069* | 0.065* |
| | (0.031) | (0.027) |
| 1985-1994 | 0.027 | 0.040* |
| | (0.018) | (0.016) |
| 1995-2004 | 0.017 | 0.025 |
| | (0.017) | (0.015) |
| 2005-2014 | 0.022 | 0.031* |
| | (0.016) | (0.015) |
| *N* | 41 | 42 |
| $R^2$ | 0.141 | 0.203 |

*Note:* *p<0.05 **p<0.01***p<0.001; standard errors in parentheses

**Table 4-7 Publication bias on linear study year**

| | 5%-CT | 10%-CT |
|---|---|---|
| | (1) | (2) |
| Constant | 0.220*** | 0.257*** |
| (ref. 2017) | (0.008) | (0.008) |
| + 1 Year | -0.001 | -0.001 |
| | (0.0005) | (0.0004) |
| *N* | 41 | 42 |
| $R^2$ | 0.028 | 0.072 |

*Note:* *p<0.05 **p<0.01***p<0.001; standard errors in parentheses

So far, the two measures, statistical power and publication bias, were looked at separately. Table 4-8 (categorical) and Table 4-9 (linear) show the time trend of the FDR without the impact of publication bias. Mirroring the trend of the statistical power, the FDR decreases over time, resulting in a FDR of less than 10% in models 1 & 2. This picture changes drastically if the *a priori* share of true hypotheses changed: for 10% *a priori* probability the FDR rises up to around 50% (models 3 & 4) and for 20% true hypotheses to more than 25% (models 5 & 6). All models of the WLS and folded normal power calculation mirror the differences of the power results. These differences seem nonetheless marginal compared to the difference caused by the changes in the priori probability of true hypotheses.

**Table 4-8 False discovery rate (w/o publication bias) on categorical study year**

| | 50% True Hypotheses | | 10% True Hypotheses | | 20% True Hypotheses | |
|---|---|---|---|---|---|---|
| | WLS | Fold | WLS | Fold | WLS | Fold |
| | (1) | (2) | (3) | (4) | (5) | (6) |
| Constant | 0.094*** | 0.078*** | 0.484*** | 0.431*** | 0.294*** | 0.252*** |
| (ref. 2015-2017) | (0.002) | (0.002) | (0.005) | (0.005) | (0.004) | (0.004) |
| 1975-1984 | 0.023*** | 0.024*** | 0.061*** | 0.074*** | 0.053*** | 0.060*** |
| | (0.004) | (0.003) | (0.010) | (0.010) | (0.009) | (0.009) |
| 1985-1994 | 0.016*** | 0.017*** | 0.044*** | 0.054*** | 0.038*** | 0.043*** |
| | (0.002) | (0.002) | (0.006) | (0.006) | (0.005) | (0.005) |
| 1995-2004 | 0.015*** | 0.012*** | 0.040*** | 0.038*** | 0.034*** | 0.030*** |
| | (0.002) | (0.002) | (0.005) | (0.006) | (0.005) | (0.005) |
| 2005-2014 | 0.012*** | 0.009*** | 0.034*** | 0.029*** | 0.029*** | 0.022*** |
| | (0.002) | (0.002) | (0.005) | (0.006) | (0.005) | (0.005) |
| *N* | 43 | 43 | 43 | 43 | 43 | 43 |
| $R^2$ | 0.652 | 0.730 | 0.675 | 0.740 | 0.665 | 0.736 |

*Note:* *p<0.05 **p<0.01***p<0.001; standard errors in parentheses

**Table 4-9 False discovery rate (w/o publication bias) on linear study year**

| | 50% True Hypotheses | | 10% True Hypotheses | | 20% True Hypotheses | |
|---|---|---|---|---|---|---|
| | WLS | Fold | WLS | Fold | WLS | Fold |
| | (1) | (2) | (3) | (4) | (5) | (6) |
| Constant | 0.101*** | 0.080*** | 0.503*** | 0.441*** | 0.310*** | 0.259*** |
| (ref. 2017) | (0.001) | (0.001) | (0.003) | (0.003) | (0.003) | (0.002) |
| + 1 Year | -0.0004*** | -0.001*** | -0.001*** | -0.002*** | -0.001*** | -0.001*** |
| | (0.0001) | (0.00005) | (0.0002) | (0.0002) | (0.0002) | (0.0001) |
| *N* | 43 | 43 | 43 | 43 | 43 | 43 |
| $R^2$ | 0.506 | 0.746 | 0.503 | 0.747 | 0.505 | 0.747 |

*Note:* *p<0.05 **p<0.01***p<0.001; standard errors in parentheses

So far, the development of the FDR over time was presented without the impact of publication bias. Taking publication bias into account (Table 4-10 & Table 4-11) increased the FDR drastically because of the inflated false positive rate. In the 50% true hypotheses scenario around 30% of all discovered significant effects are in fact null effects. For the 10% scenario the share of false discoveries rises up to almost 80% while also in the 20% scenario the FDR was around 60%. Therefore, in the 20% scenario more than half of the claimed significant effects in the literature do not hold up. As for the FDR not affected by publication bias, the decreasing trend over time caused by an increased statistical power is very robust across all model specifications. Compared to the *a priori* share of true hypotheses the impact of different modelling strategies of the statistical power as well as publication bias (5%- and 10%-CT) and parametric specification of the time trend had minor impact.

**Table 4-10 False discovery rate (w/ publication bias) on categorical study year**

| | 50% True Hypotheses | | | | 10% True Hypotheses | | | | 20% True Hypotheses | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | WLS | | Fold | | WLS | | Fold | | WLS | | Fold | |
| | 5%-CT | 10%-CT | 5%-CT | 10%-CT | 5%-CT | 10%-CT | 5%-CT | 10%-CT | 5%-CT | 10%-CT | 5%-CT | 10%-CT |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) | (10) | (11) | (12) |
| Con-stant | 0.294*** | 0.314*** | 0.265*** | 0.285*** | 0.788*** | 0.804*** | 0.763*** | 0.782*** | 0.623*** | 0.646*** | 0.589*** | 0.614*** |
| (ref. 2015-2017) | (0.008) | (0.006) | (0.009) | (0.007) | (0.006) | (0.004) | (0.007) | (0.005) | (0.009) | (0.006) | (0.010) | (0.007) |
| 1975-1984 | 0.067*** | 0.058*** | 0.076*** | 0.068*** | 0.047*** | 0.038*** | 0.060*** | 0.049*** | 0.069*** | 0.057*** | 0.084*** | 0.071*** |
| | (0.017) | (0.013) | (0.018) | (0.014) | (0.012) | (0.008) | (0.015) | (0.010) | (0.018) | (0.012) | (0.021) | (0.014) |
| 1985-1994 | 0.040*** | 0.042*** | 0.046*** | 0.050*** | 0.030*** | 0.029*** | 0.039*** | 0.037*** | 0.043*** | 0.043*** | 0.054*** | 0.054*** |
| | (0.010) | (0.007) | (0.011) | (0.008) | (0.007) | (0.005) | (0.008) | (0.006) | (0.010) | (0.007) | (0.012) | (0.008) |
| 1995-2004 | 0.033** | 0.033*** | 0.032** | 0.034*** | 0.025*** | 0.023*** | 0.029*** | 0.027*** | 0.036*** | 0.034*** | 0.039** | 0.038*** |
| | (0.009) | (0.007) | (0.010) | (0.008) | (0.007) | (0.004) | (0.008) | (0.005) | (0.010) | (0.007) | (0.011) | (0.008) |
| 2005-2014 | 0.032** | 0.034*** | 0.030** | 0.033*** | 0.024*** | 0.023*** | 0.026** | 0.026*** | 0.035*** | 0.034*** | 0.036** | 0.036*** |
| | (0.009) | (0.007) | (0.010) | (0.007) | (0.007) | (0.004) | (0.008) | (0.005) | (0.010) | (0.007) | (0.011) | (0.008) |
| *N* | 41 | 42 | 41 | 42 | 41 | 42 | 41 | 42 | 41 | 42 | 41 | 42 |
| $R^2$ | 0.382 | 0.516 | 0.428 | 0.560 | 0.398 | 0.540 | 0.438 | 0.585 | 0.393 | 0.533 | 0.435 | 0.578 |

*Note:* *p<0.05 **p<0.01***p<0.001; standard errors in parentheses

**Table 4-11 False discovery rate (w/ publication bias) on linear study year**

| | 50% True Hypotheses | | | | 10% True Hypotheses | | | | 20% True Hypotheses | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | WLS | | Fold | | WLS | | Fold | | WLS | | Fold | |
| | 5%-C | 10%-C | 5%-C | 10%-C | 5%-C | 10%-C | 5%-C | 10%-C | 5%-C | 10%-C | 5%-C | 10%-C |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) | (10) | (11) | (12) |
| Con-stant | 0.313*** | 0.333*** | 0.280*** | 0.302*** | 0.803*** | 0.818*** | 0.777*** | 0.795*** | 0.645*** | 0.666*** | 0.608*** | 0.633*** |
| (ref. 2017) | (0.005) | (0.004) | (0.005) | (0.004) | (0.004) | (0.003) | (0.004) | (0.003) | (0.005) | (0.004) | (0.006) | (0.004) |
| + 1 Year | -0.001** | -0.001*** | -0.001*** | -0.001*** | -0.001** | -0.001*** | -0.001*** | -0.001*** | -0.001** | -0.001*** | -0.001*** | -0.001*** |
| | (0.0003) | (0.0002) | (0.0003) | (0.0002) | (0.0002) | (0.0002) | (0.0002) | (0.0002) | (0.0003) | (0.0002) | (0.0003) | (0.0003) |
| *N* | 41 | 42 | 41 | 42 | 41 | 42 | 41 | 42 | 41 | 42 | 41 | 42 |
| $R^2$ | 0.184 | 0.274 | 0.275 | 0.388 | 0.186 | 0.269 | 0.276 | 0.383 | 0.186 | 0.271 | 0.276 | 0.386 |

*Note:* *p<0.05 **p<0.01***p<0.001; standard errors in parentheses

# 5. Berufungsverfahren als Turniere: Berufungschancen von Wissenschaftlerinnen und Wissenschaftlern

Auspurg, K., T. Hinz und A. Schneck. Erschienen in: Zeitschrift für Soziologie. 2017 (46):283-302. https://doi.org/10.1515/zfsoz-2017-1016

## 5.1. Einleitung

Nach welchen Kriterien werden in der Wissenschaft Ressourcen, Professuren oder Forschungsmittel vergeben? Inwieweit wird hierbei das Prinzip der Leistungsgerechtigkeit befolgt? Diese Fragen sorgen für anhaltende wissenschaftspolitische Diskussionen und haben bereits etliche Studien angeregt (s. z. B. das 2013 publizierte Nature-Special "Women in Science").[84] Die mögliche Benachteiligung von Wissenschaftlerinnen in Berufungsverfahren wird dabei als wesentliche Ursache für die starke Unterrepräsentanz von Professorinnen angesehen (für die USA: Long et al. 1993; für Deutschland: Färber & Spangenberg 2008). Studien jüngeren Datums zum deutschen Wissenschaftssystem finden jedoch genau das gegenteilige Ergebnis einer Bevorzugung von Frauen (Jungbauer-Gans & Gross 2013; Lutter & Schröder 2016). Wenden sich hier mögliche Erfolge gleichstellungspolitischer Bemühungen nun in ihr Gegenteil?

Die Grundthese des vorliegenden Beitrages ist, dass das heterogene Bild von anhaltender Benachteiligung bis hin zu einer Bevorzugung von Wissenschaftlerinnen einem zu oberflächlichen Blick geschuldet ist, der die einzelnen Berufungsverfahren und ihre Verfahrenslogik nicht ausreichend würdigt. Die bestehenden Studien fokussieren fast ausschließlich auf das Endergebnis von Auswahlprozessen und berichten Analysen auf der Aggregatebene von Fächern, Universitäten und Ländern. Berufungsverfahren stellen allerdings eine Art *winner-takes-all* Turnier mit verschiedenen Auswahlrunden von der Bewerbung bis zur Ruferteilung dar, bei denen es neben den Qualifikationen der einzelnen Bewerber/innen auch darauf ankommt, wer mit wem um welche Stelle konkurriert. So unterscheiden sich die Stellenanforderungen selbst innerhalb von einzelnen Disziplinen stark und je nach Ausschreibung ist mit einer anderen Konkurrenzsituation zu rechnen. Werden diese Wettbewerbsstrukturen nicht berücksichtigt, sind Ergebnisse kaum sinnvoll zu interpretieren. Dieser Aspekt wurde für den allgemeinen Arbeitsmarkt eindrücklich aufgedeckt (eindrücklich z. B. Fernandez & Mors 2008; Fernandez & Sosa 2005), erstaunlicherweise aber für Auswahlverfahren im Wissenschaftssystem bislang ausgeblendet.

Im vorliegenden Beitrag kann dazu auf einen umfangreichen Datensatz zurückgegriffen werden. Für eine mittelgroße deutsche Universität liegen prozessproduzierte Daten zu mehr als 230 Berufungsverfahren in den Jahren 2001 bis 2013 vor. Über den Verfahrensverlauf, der aus *fünf Stufen* besteht (Bewerbung, Erstauswahl, Vorstellungsvortrag, Aufnahme in den Berufungsvorschlag, erster Listenplatz), lässt sich feststellen, ob und bei welchen Verfahrensstufen Veränderungen der Frauenanteile auftreten.

[84] http://www.nature.com/news/specials/women/index.html.

Für knapp 450 Listenplatzierte aus 158 Verfahren liegen zudem umfassende Publikations- und Lebenslaufdaten vor, die es erlauben, die besonders strittige Frage vergleichbarer Leistungen zu adressieren.

## 5.2. Geschlechtsspezifische Auswahlverfahren in der Wissenschaft?

### *5.2.1. Selbstselektion, Diskriminierung und Stereotype*

Mit welchen theoretischen Konzepten kann eine mögliche Geschlechterspezifität von Auswahlverfahren in der Wissenschaft begründet werden, wenn doch weitgehend Konsens darüber besteht, dass Auswahlentscheidungen primär an der universellen Norm der Leistungsgerechtigkeit (Meritokratie) orientiert sein sollten (Merton 1973)? Zunächst könnten tatsächliche *Geschlechterunterschiede in der Produktivität* bestehen (Allgemeine Überblicke über den Forschungsstand bieten z. B. Findeisen 2011; Gross & Jungbauer-Gans 2007; Leemann 2005; Lind 2004). Ursächlich zu denken ist hier etwa an die mögliche stärkere Belastung von Wissenschaftlerinnen mit Haus- und Familienarbeit. Weitere diskutierte Gründe für eine niedrigere Produktivität sind die geringere Spezialisierung von Wissenschaftlerinnen, die es erschwert, Artikel oder Anträge auf Forschungsmittel in kurzer Zeit zu verfassen (Leahey 2007; Leahey et al. 2008; Leahey et al. 2010), sowie ihre schlechtere Unterstützung durch Mentoren und eine Beschäftigung auf ungünstigeren Positionen (Lind 2004). Diese Aspekte sind für die Erklärung von Ungleichheiten relevant, im vorliegenden Beitrag interessiert aber primär, ob es trotz gleicher Leistungen noch zu ungleichen Berufungschancen kommt.

Zu Beginn des Berufungsverfahrens können bereits Leistungsunterschiede vorhanden sein, wenn sich primär Wissenschaftlerinnen mit überdurchschnittlichen Leistungen bewerben (Selbstselektion). Befragungsdaten zufolge gehen Wissenschaftlerinnen oftmals davon aus, schlechtere Auswahlchancen gegenüber männlichen Mitbewerbern zu haben (s. z. B. für den Zugang zu Stipendien: Deutsche Forschungsgemeinschaft 2009). Diese Einschätzung wird sicherlich auch von Gleichstellungspolitiken gestärkt, die Frauen vor Diskriminierungen warnen (Hirschauer 2016). Bei einer Antizipation von vergleichsweise schlechteren Chancen lohnt es dann weniger, in Bewerbungsaktivitäten zu investieren, bzw. müssen für realistische (subjektive) Erfolgschancen die (vermeintlichen) Benachteiligungen durch entsprechend überdurchschnittliche Leistungen kompensiert werden (Engels et al. 2012; Duch et al. 2012; Färber & Spangenberg 2008; Lee 2016). Jedenfalls ist schon allein aufgrund der Tatsache, dass Frauen überdurchschnittlich häufig das Wissenschaftssystem verlassen, von einem *survivor bias* in der Form auszugehen, dass nur besonders exzellente Wissenschaftlerinnen im Bewerberpool für Professuren verbleiben (Lutter & Schröder 2016). Berücksichtigt man dies nicht, kann der ungerechtfertigte Eindruck einer Bevorzugung von Wissenschaftlerinnen entstehen (Lee 2016).

Wie lassen sich weitere, im Berufungsverfahren selbst ansetzende Auswahlprozesse theoretisch fassen? Bei einer pauschalen Ablehnung von Wissenschaftlerinnen aufgrund Abneigungen würde man von präferenzbasierter Diskriminierung (Becker 1971) sprechen. In der Wissenschaft wird dabei – in Anleh-

nung an den gut erforschten Matthäuseffekt einer zunehmenden Kumulation von Ressourcen bei wenigen Personen – oft auch von einem „Matilda-Effekt“ gesprochen (Lincoln et al. 2012). Eine rein präferenzbasierte Ablehnung oder Bevorzugung ist aus verschiedenen Gründen jedoch unwahrscheinlich.[85]

Häufiger wird in der Wissenschaft eine *statistische Diskriminierung* vermutet. Diese Theorie unterstellt rationale Akteure, die Probleme unvollständiger (oder nur sehr kostspielig zu erwerbender) Information zu lösen haben (Arrow 1971; Phelps 1972). Bei fehlenden Informationen zur exakten Leistungsfähigkeit greifen Entscheidungsträger auf bekannte Durchschnittswerte der jeweiligen Gruppe zurück, was zu einer Diskriminierung von Forschenden mit überdurchschnittlicher Leistung führen kann. Eine statistische Diskriminierung von Wissenschaftlerinnen erscheint in frühen Karrierestadien, wie etwa bei der Auswahl von Doktorand/innen, plausibel, da Frauen im Durchschnitt häufiger das Wissenschaftssystem verlassen, womit sich Investitionen in ihre Karrieren langfristig weniger auszahlen (Konsortium Bundesbericht Wissenschaftlicher Nachwuchs 2013: 141; Löther 2015). Bewerbungen auf Professuren erfolgen aber in der Regel erst in Karrierestadien, in denen bereits so viel Zeit und Energie investiert wurde, dass gewollte Ausstiege unwahrscheinlich sind. In weiteren Varianten statistischer Diskriminierung könnte es rational sein, Gruppen zu diskriminieren, bei denen die Leistungsfähigkeit weniger reliabel prognostizierbar ist (Aigner & Cain 1977; Borjas & Goldberg 1978; Fang & Moro 2011). Einigen Studien zufolge publizieren Frauen im Vergleich zu Männern eher erst zu späteren Karrierestadien hochrangig, in denen sie sich schon fest auf Professuren etabliert haben (s. z. B. Long 1992). Dann wäre ihr Leistungspotenzial zum Bewerbungszeitraum womöglich tatsächlich schwieriger prognostizierbar. Unabhängig auf welchem Mechanismus sie basiert, sollte statistische Diskriminierung definitionsgemäß jedenfalls zurückgehen, wenn mehr Informationen über die Bewerber/innen vorliegen.

Im Gegensatz zur statistischen Diskriminierung kann bei sozialpsychologischen Theorien zu Stereotypen das Geschlecht selbst bei vollständigen Informationen ein verzerrter Proxy für die wissenschaftliche Leistungsfähigkeit sein. Nach *reward expectation* Theorien folgen nicht nur Gratifikationen der Leistungsfähigkeit, sondern wird umgekehrt auch aus Belohnungen (Ressourcen) auf die Leistungsfähigkeit geschlossen (Berger et al. 1985; Berger & Murray 2006). Verfügt eine Gruppe, wie etwa männliche Wissenschaftler, im Allgemeinen über mehr Ressourcen als eine andere Gruppe, kann es dazu kommen, dass das askriptive Gruppenmerkmal, hier also Geschlecht, mit einer allgemein höheren Kompetenz in Verbindung gebracht wird (s. dazu auch die status construction theory von Ridgeway 1991,2014). Der Hauptunterschied zur statistischen Diskriminierung liegt darin, dass die kognitiv verankerten Stereotype einer höheren Leistungsfähigkeit nicht mit der wahren Leistung der Gruppen übereinstimmen müssen;

[85] Die pauschale und direkte Diskriminierung sollte ökonomischen Theorien zufolge aufgrund der ineffizienten Ausnutzung von Talent auf Märkten mit hohem Wettbewerbsdruck nicht dauerhaft bestehen (Becker 1971). Weiterhin bestehen starke normative Vorgaben, deren Einhaltung in Kollegialorganen überwacht wird. Zudem bräuchte es überhaupt erst einmal eines Argumentes, warum Präferenzen (von Männern) gegen weibliche Kolleginnen bestehen.

und sie sich nicht einfach durch weitere Informationen revidieren lassen (Ridgeway 2014; Correl & Benard 2006).

Einen Mechanismus für die Persistenz solcher Stereotype liefert die Theorie doppelter Leistungsstandards (Foschi 1996; Foschi et al. 1994). Demnach würde die kognitive Prägung durch Stereotype bereits zu einer verzerrten Wahrnehmung der Leistungen führen: Erwartungskonträre, nicht den gängigen Klischees entsprechende überdurchschnittliche (unterdurchschnittliche) Leistungen von Wissenschaftlerinnen (Wissenschaftlern) werden tendenziell eher als Messfehler gewertet. Zusammengenommen führt dies dazu, dass striktere Maßstäbe an die Leistungen von Frauen angelegt werden (daher auch die Bezeichnung *double standards*).

Ein weiterer Mechanismus, warum Stereotype Wissenschaftlerinnen zu schaffen machen könnten, ist der sog. *stereotype threat* (Steele 1997; Steele & Aronson 1995), wonach Stereotype einer gruppenspezifisch geringeren Leistungsfähigkeit zu einer stärkeren Nervosität und Angst in Leistungssituationen führen, die dann in einer Art sich selbst erfüllende Prophezeiung zur Bestätigung des Stereotyps führen. Eine solche stereotypenaktivierende Situation könnte in Berufungsverfahren der Vorstellungsvortrag sein, da bei diesem im Gegensatz zu den übrigen Verfahrensstufen die Auswahl nicht rein aktenbasiert erfolgt, sondern den Bewerber/innen eine aktive Leistung abverlangt wird.

### *5.2.2. Kontextfaktoren: Stellenvakanzen und Labor Queues*

Der vorliegende Aufsatz plädiert dafür, die Vakanz (also die zu besetzende Professur) und die Wettbewerbssituation um die jeweilige Vakanz stärker zu berücksichtigen. Wie unterschiedlich wissenschaftliche Leistungen nach Fächern gewertet werden, veranschaulicht Igor Podlubny, demzufolge "one citation in mathematics roughly corresponds to 15 citations in chemistry, 19 citations in physics, and 78 citations in clinical medicine" (2005: 98). Bislang wurde in der Wissenschaftsforschung oft ausgeblendet, dass es ebenso starke Variationen *innerhalb* von Disziplinen geben kann. So ist etwa in der quantitativen Soziologie die Orientierung an bibliometrischen Maßen stärker etabliert als in der Theoriearbeit (siehe die Evaluation des Faches Soziologie durch den Wissenschaftsrat).[86] Hinzu kommt, dass in den einzelnen Forschungsfeldern unterschiedlich viele Wissenschaftler/innen aktiv sind. Allein schon mit der Größe des jeweiligen Forschungsfeldes steigt die Chance, zitiert zu werden. Weiter entscheidet die Anzahl an Publikationsorganen (welche wiederum mit dem Alter von Forschungsfeldern korreliert, in jungen Forschungsfeldern gibt es weniger etablierte Journale) über Publikations- und Zitationschancen. Wählt man hier keine adäquaten Referenzstandards, drohen Leistungsmessungen zu dem sprichwörtlichen Vergleich von „Äpfeln mit Birnen" zu verkommen (Bornmann et al. 2008). Aufgrund geschlechtsspezifischer Spezialisierung mag die Diagnose vermeintlicher Diskriminierung naheliegen, etwa wenn

[86]http://www.wissenschaftsrat.de/download/Forschungsrating/Dokumente/Grundlegende%20Dokumente%20zum%20Forschungsrating/8422-08.pdf

übersehen wird, dass Wissenschaftlerinnen stärker in Feldern aktiv sind, in denen der wissenschaftliche *impact* generell (geschlechtsunabhängig) geringer ist (Leahey et al. 2010).

Für Berufungsverfahren gibt es allerdings Argumente, dass die Berücksichtigung der Fächer allein nicht hinreichend wäre. So dürften die Quantität und Qualität des Bewerberfeldes auch stark mit der Denomination der Professur und damit fachintern zwischen Verfahren variieren. So werden etwa in einzelnen Feldern unterschiedlich viele Wissenschaftler/innen ausgebildet, oder es kann unterschiedlich viele Vakanzen oder alternative Joboptionen (außerhalb des akademischen Wissenschaftsbetriebs) geben.

Systematischer wurde die Idee, dass Jobvakanzen und der mit ihnen verbundene Wettbewerb die Auswahlchancen strukturieren, in der Arbeitsmarktsoziologie mit „Turniermodellen" und speziell Theorien zu *labor queues* gefasst (Fernandez & Mors 2008; Fernandez-Mateo & Fernandez 2013; Reskin 1991). Die zentrale Überlegung lautet, dass unterschiedliche Platzierungschancen erst dann umfassend zu verstehen sind, wenn berücksichtigt wird, welche (spezifischen) Bewerber/innen zu einer bestimmten Zeit um welche offenen Vakanzen konkurrieren, und mit welcher Sortierlogik diese dann in eine Rangfolge gebracht werden. Lässt man die unterschiedliche Chancenstruktur außer Acht, kann fälschlicherweise der Eindruck einer gruppenspezifischen Benachteiligung entstehen, indem man etwa konstatiert, dass Männer mehr Publikationen für einen Ruf benötigen; dabei aber übersieht, dass dies nur im Aggregat *zwischen* und nicht *innerhalb* der einzelnen Verfahren bzw. Turniere gilt.[87]

### *5.2.3. Überblick über die diskutierten Annahmen*

Wie die bisherigen Überlegungen gezeigt haben, werden aktuell sehr kontroverse und zum Teil widersprüchliche Erwartungen diskutiert; zugleich führen häufig unterschiedliche Theorieannahmen zu ähnlichen Vorhersagen für den *Ausgang* von Verfahren (globale Erwartungen über bessere oder schlechtere Chancen von Frauen). Der Blick auf den *Verlauf* der Verfahren mit ihren unterschiedlichen Selektionskriterien ermöglicht es, die verschiedenen Mechanismen besser zu unterscheiden – so kann man etwa nach den Theorien statistischer Diskriminierung annehmen, dass eine Benachteiligung von Frauen insbesondere auf den ersten Verfahrensstufen besteht, bei denen noch vergleichsweise wenige Informationen über die Bewerber/innen vorliegen. Führen dagegen *stereotype threats* zu Chancennachteilen von Frauen, sollte sich das ausschließlich auf der Stufe bemerkbar machen, auf der eine aktive Performance von Bewerberinnen verlangt wird. Dies sind vor allem die Vorstellungsvorträge. Die nachfolgende Tabelle 5-1 zeigt die unterschiedlichen Annahmen in der Übersicht – sie sind hier bewusst zugespitzt, um empirische Unterscheidungen zu ermöglichen.

[87] Bei Unterschieden, die ausschließlich zwischen Verfahren bestehen, könnte man wohl allenfalls eine institutionelle Diskriminierung in dem Sinne konstatieren, dass allgemeine Verfahrensregeln stärker die eine oder andere Gruppe bevorteilen (Gomolla & Radtke 2009: 50).

**Tabelle 5-1 Übersicht zu Annahmen und Erwartungen in Berufungsverfahren**

| Theorie | Annahmen | Erwartungen |
|---|---|---|
| Chancen von Bewerber/innen und Leistungsbewertungen | | |
| Selbstselektion/Antizipierte Diskriminierung | Wissenschaftlerinnen bewerben sich zurückhaltender als Wissenschaftler. | $P_{frau}(Bew.) < P_{mann}(Bew.)$ |
| | Bewerberinnen haben aufgrund einer Positivselektion (höhere Leistungen) *bessere* Chancen in Berufungsverfahren als Bewerber. | $X_{frau} > 0$ |
| | Dies gilt nicht mehr, wenn Leistungen kontrolliert sind. | $X'_{frau} < X_{frau}$ (wobei $X'_{frau}$ den Geschlechtereffekt unter Kontrolle von Leistung darstellt.) |
| Statistische Diskriminierung / *reward expectations* | Bewerberinnen haben *geringere* Chancen in Berufungsverfahren als männliche Bewerber. | $X_{frau} < 0$ |
| Statistische Diskriminierung | Dies gilt primär auf den *ersten* Verfahrensstufen, in denen vergleichsweise wenig Information vorliegt. | $X_{frau} * Z_{verfahrensstufe} > 0$ |
| *reward expectations* / *double standards* | Leistungen von Männern werden eher anerkannt. | $X_{leistung} > 0$<br>$X_{leistung} * X_{frau} < 0$ |
| *stereotype threats* | Bewerberinnen haben speziell beim *Bewerbungsvortrag* geringere Chancen als Bewerber. | $X_{frau} * Z_{vortrag} < 0$ |
| Wettbewerbsstruktur | | |
| *labor queues* | Nach Kontrolle für die einzelnen Verfahren verringern sich Geschlechterunterschiede. | $\|X°_{frau}\| < \|X_{frau}\|$ (wobei X° den Effekt darstellt, der unter Kontrolle der einzelnen Verfahren geschätzt ist.) |

*Anmerkung*: P bezeichnet einen Anteilswert. Mit X werden Effekte von Bewerber/innenvariablen auf die Auswahlchance bezeichnet, mit Z Effekte von Variablen, welche die Eigenschaften der Verfahren beschreiben. Bew. = Abkürzung für Bewerbungen.

## 5.3. Forschungsstand

### *5.3.1. Forschung zum Wissenschaftssystem allgemein*

Forschungen zur Selbstselektion stützen sich auf Vergleiche der Frauenanteile beim Pool möglicher Bewerber/innen und bei tatsächlichen Bewerbungen und sprechen bislang insgesamt dafür, dass sich Frauen zurückhaltender als Männer bewerben (für Anträge auf Stipendien in den Niederlanden: Brouns 2000; für DFG-Einzelanträge: Hinz et al. 2008).

Im Hinblick auf Erfolgschancen liegt umfangreiche Forschung inzwischen vor allem zu geschlechtsspezifischen Akzeptanzquoten von Zeitschriftenartikeln und der Vergabe von Forschungsmitteln und Stipendien vor (für ausführliche Darstellungen siehe z. B. Bornmann 2011; Gross & Jungbauer-Gans 2007; Lind 2004). Frühere Arbeiten haben dabei oft recht spektakuläre Benachteiligungen von Frauen aufgezeigt, so findet etwa die stark beachtete Studie von Wold und Wennerås zur Vergabe von Stipendien durch das *Swedish Medical Research Council* (MRC), dass Wissenschaftlerinnen für die gleiche Leistungsbewertung deutlich mehr Publikationen vorweisen mussten; der Unterschied betrug "approximately three extra papers in *Nature* or *Scienc*e (…), or 20 extra papers in a journal with an impact factor of around 3" (1997: 342). Dieses beunruhigende Ergebnis konnte jedoch weder für das MRC noch andere Institutionen repliziert werden (Sandström & Hällsten 2008), zudem gab es methodische Kritik an der Originalstudie (Ceci & Williams 2011). In den letzten Jahren scheint sich der Forschungsstand auf eine recht gut abgesicherte Erkenntnis einer inzwischen weitgehenden Gleichbehandlung der Geschlechter einzupendeln (Ceci & Williams 2011). So findet die bislang umfassendste und methodisch sehr versierte Meta-Analyse zu Drittmittelbewerbungen insgesamt keinen signifikanten Geschlechtereffekt (Marsh et al. 2009), bei zugleich aber einer sehr großen Heterogenität der Einzelbefunde. Letztere gibt Anlass zu Spekulationen, ob es nicht doch in einzelnen Fachdisziplinen zu geschlechtsspezifischen Benachteiligungen kommt (Brouns 2000; Sanz-Menéndez et al. 2013). So berichten einzelne Autoren, dass faire Urteile erst bei einem gewissen Mindestanteil an Bewerberinnen auftreten (Sackett et al. 1991), und auch die Relevanz von Fachdisziplinen gibt noch Rätsel auf. Beispielsweise finden Studien, dass in den stärker standardisierten Naturwissenschaften die Benachteiligung von Wissenschaftlerinnen geringer ist als in den Geistes- und Sozialwissenschaften oder Wissenschaftlerinnen dort einen Bonus erhalten; andere Studien berichten das Gegenteil (für Stipendienanträge in den Niederlanden: Brouns 2000; für DFG-Einzelanträge: Auspurg & Hinz 2010).[88]

Weiteren Anlass zur Skepsis, ob geschlechtsneutral begutachtet wird, geben sozialpsychologische Studien, die bei Standardisierung der Leistungsinformationen im Experiment Evidenz für strengere Bewertungsstandards gegenüber Wissenschaftlerinnen finden, womit die Theorie der *double standards* unterstützt wird (s. z. B. Knobloch-Westerwick et al. 2013; Lee 2016; Steinpreis et al. 1999). Allerdings stellt sich hier die Frage, ob ein *publication bias* in Form einer Unterdrückung von erwartungskonträren Ergebnissen zu einer Überschätzung der Effekte führt (allgemein zum Publication Bias: Dickersin & Min 1993; für aktuelle Evidenz in der Psychologie: Kühberger et al. 2014). Experimente mit einem *double blind* Verfahren sowie natürliche Experimente mit Umstellung des Begutachtungsmodus auf einen *blind review* finden keine Effekte (Lee et al. 2013). Insgesamt kann man festhalten, dass gemittelt über alle Fachdisziplinen die geschlechterspezifischen Verzerrungen wohl gering ausfallen. Jedoch handelt es

[88] Die Studie von Auspurg & Hinz (2010) beruht auf Vollerhebungen prozessproduzierter Daten zu etwa 79.000 DFG-Einzelanträgen und damit sehr hohen Fallzahlen, die Zufallsschwankungen eigentlich ausschließen. Gleichwohl gelang es nicht ein einheitliches Muster von Zusammenhängen der fachspezifischen Förderquoten mit Frauenanteilen unter Bewerbungen, Fachdisziplinen oder auch Zusammensetzungen der Gutachtenden auszumachen (s. auch Findeisen et al. 2010).

sich bei der Besetzung von Professuren im deutschen Wissenschaftssystem um vergleichsweise personalisierte Verfahren (schließlich geht es um die Auswahl künftiger Kolleg/innen), sodass sich die angeführten Ergebnisse zur Drittmittelvergabe oder Zeitschriftenbegutachtungen nur bedingt übertragen lassen. Auffallend ist weiterhin, dass sich viele Studien zur Forschungsförderung lediglich mit dem Endergebnis beschäftigen (Förderung ja/nein), ohne die ablaufenden Prozesse zu analysieren.

### *5.3.2. Forschung zu Berufungsverfahren*

Forschungsbedarf besteht insbesondere noch hinsichtlich der Kriterien, die für Berufungen auf Professuren angelegt werden. Eine erste Studie mit Einbezug von Leistungsmerkmalen wurde in den USA von Long et al. (1993) vorgelegt. Es fanden sich geringere Auswahlchancen für Wissenschaftlerinnen (um 23 Prozent geringere Chance, zum *associate professor* berufen zu werden; für *full professors* ist die Chance um 40 Prozent geringer), wobei die Unterschiede etwa zur Hälfte durch geringere Publikationsleistungen erklärbar waren. Dabei zeigten sich zunächst Hinweise auf *double standards* zu Gunsten von weiblichen *assistant professors*. Die in dem Aufsatz vorbildlich genauen Analysen demonstrieren dann aber, dass dieser Aspekt im Wesentlichen durch wenige außergewöhnlich produktive Wissenschaftlerinnen bedingt war; nach Ausschluss von Extremwerten verschwand die Evidenz für unterschiedliche Leistungsstandards und erwies sich die Mehrheit der Wissenschaftlerinnen hinsichtlich ihrer Karrierechancen als benachteiligt (Long et al. 1993: 718f.).

Aktuellere internationale Studien legen – wie von Long et al. (1993) aufgrund des wachsenden Gleichstellungsdrucks schon Anfang der 1990er Jahre prognostiziert – abnehmende Geschlechterunterschiede nahe. So finden etwa Sanz-Menéndez et al. (2013) auf der Basis von Ereignisdatenanalysen von Befragungsdaten keine signifikanten Geschlechterunterschiede bei den Berufungschancen in Spanien. De Paola & Scoppa (2011) berichten für Italien, und damit für ein Land, das wie Spanien ein stark standardisiertes, zentrales Auswahlverfahren anwendet, lediglich dann geringe Chancennachteile für Bewerberinnen, wenn sie mit einem rein männlichen Gutachterteam konfrontiert sind.

Aus Deutschland liegen einige Studien vor, die entweder keine Diskriminierung oder sogar eine Positivdiskriminierung von Frauen finden. Der Großteil der Studien konzentriert sich dabei auf die Ökonomik. Schulze et al. (2008) befragten Kohorten an Ökonom/innen mit Habilitation in den Jahren 1985-2006 im deutschsprachigen Raum (realisiertes $N = 934$, bei einem Rücklauf von 54 Prozent); die Grundgesamtheit der Habilitierten wurde dabei durch Anfragen bei Dekanaten und Bekanntmachungen in Zeitschriften recherchiert. Mit Probit- und Ereignisdatenanalysen wurden die Chance und Dauer bis zu einem Erstruf analysiert; bei Kontrolle für Leistungsmerkmale fanden sich keine Geschlechterunterschiede. Die Erfolgsfaktoren variierten aber deutlich für Professuren in der BWL und VWL, was die These unterstreicht, dass es innerhalb von Fächern deutliche Unterschiede in den Auswahlregeln geben kann. Mit sehr ähnlichen Methoden untersuchten Jungbauer-Gans & Gross (2013) die Berufungschancen von Habilitierten in Deutschland in den Fächern Rechtswissenschaft, Mathematik und Soziologie (Kohorten 1985-2005; $N = 716$; bereinigte Rücklaufquote von 45 Prozent). Dabei finden sich in allen

Fächern Hinweise auf partikularistische Kriterien, so zeigen in der Rechtswissenschaft die Anzahl der elterlichen Bildungsjahre und in der Mathematik das Berufsprestige der Eltern einen signifikanten Einfluss auf die Berufungschance. Das Geschlecht hat lediglich in der Soziologie einen Einfluss: Wissenschaftlerinnen haben hier bessere Berufungschancen, die Übergangsrate in eine Professur ist bei ihnen um etwa den Faktor 2 und damit deutlich erhöht (vgl. die Tabelle in Jungbauer-Gans & Gross 2013: 86). Kontrolliert werden in dem Modell zusätzlich die Kohorte, das Alter bei der Habilitation, das Prestige der Herkunftsuniversität sowie der Anteil an Hausarbeit. Das Ergebnis kann man daher auch so interpretieren, dass Soziologinnen für die gleichen Berufungschancen wie männliche Soziologen vergleichsweise weniger SSCI-Publikationen benötigen. Aus unserer Sicht nicht nachvollziehbar ist, warum in dem Modell für den privaten Hausarbeitsteil kontrolliert wird: Zunächst, weil dieser den Kommissionen keinesfalls bekannt sein dürfte; vor allem aber auch, weil damit ein Endogenitätsproblem generiert wird (der Hausarbeitsteil in Partnerschaften dürfte allen Standardmodellen der Familiensoziologie zufolge ganz wesentlich vom beruflichen Erfolg, also der abhängigen Variable beeinflusst sein, womit alle Schätzergebnisse verzerrt wären; vgl. Wooldridge 2013: 86f.).

Weitere Studien versuchen, mittels Internetrecherchen das gesamte wissenschaftliche Personal an Hochschulen zusammen mit im Netz bereitgestellten Lebensverlaufsinformationen zu erfassen, um dann Unterschiede in der Chancenstruktur auf Erstrufe (Plümper & Schimmelfennig 2007) bzw. die Dauer bis zu einem Erstruf zu schätzen. Die Analysen von Plümper & Schimmelfennig (2007) für die deutsche Politikwissenschaft zeigen, dass Frauen seltener in frühen Karrierestadien berufen werden, jedoch lässt sich insgesamt, bei Kontrolle der Publikationsleistung, keine Benachteiligung feststellen (Plümper & Schimmelfennig 2007). Lutter & Schröder (2016: 1000) zufolge liegt die Chance von Soziologinnen, eine Professur zu erhalten, um 40 Prozent über der Chance vergleichbar ausgewiesener Soziologen. Allerdings erzielen sie diese Ergebnisse unter Einbezug der weniger prestigeträchtigen W2-Professuren, die deutlich häufiger von Wissenschaftlerinnen besetzt werden. Den Autoren zufolge zeigen die Resultate, "that women get tenure with fewer publications, and therefore each of their publications is more strongly rewarded in terms of tenure, as compared with men" (Lutter & Schröder 2016: 1009). Dies sei wohl auch ein Ergebnis der *affirmative action* Politik der Universitäten (ebd.). Die Studien von Plümper & Schimmelfennig (2007) sowie Lutter & Schröder (2016) sind jedoch datenbedingt durchaus problematisch. Der klare Vorteil, den sie aus der Verwendung von quasi-prozessproduzierten Daten ziehen, ist die Vermeidung eines selektiven Rücklaufs, wie er bei Befragungsstudien auftritt. Allerdings hat ihr Design den Nachteil, dass die Analysen womöglich von einem besonders starken *survivor bias* geprägt sind: Es werden *per definitionem* nur Personen in die Auswahlgesamtheit einbezogen, welche an Hochschulen verblieben sind. Damit werden die Erfolgschancen einer Professur lediglich mit einem anderweitigen Verbleib an Hochschulen verglichen. Haben etwa Männer eine bessere Chance (mit geringem Publikationsoutput) auch ohne Professorentitel langfristig an Universitäten zu verbleiben (an der hier untersuchten Universität gab es dafür tatsächlich Evidenz), während Frauen (mit wenigen Publikatio-

nen) eher das Wissenschaftssystem verlassen müssen, könnte dies fälschlicherweise zum Eindruck führen, dass Männer den Sprung von der Mitarbeiterstelle auf die Professur nur mit vergleichsweise mehr Publikationen schaffen. Mit anderen Worten, es fehlt ein entscheidender *outcome* – das Verlassen(müssen) des Wissenschaftssystem bei fehlendem Berufungserfolg –, womit Erfolgschancen (selektiv) falsch eingeschätzt werden (für eine ähnliche Kritik: Schulze et al. 2008).

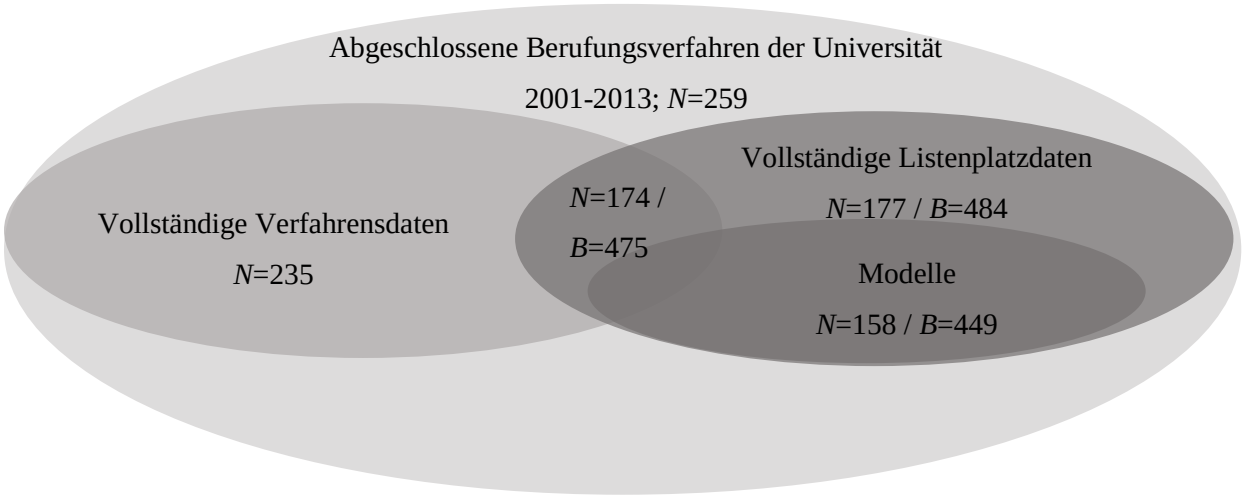
Die Idee, die Veränderung der Frauenanteile über die einzelnen Verfahrensstufen hinweg zu betrachten, wurde u. W. bislang nur von lediglich einer Studie aus den Niederlanden aufgegriffen. Bei einem solchen Vorgehen vergleicht man Frauenanteile auf der ersten Stufe der Bewerbung mit den Frauenanteilen im potenziellen Pool, womit man Mechanismen der Selbstselektion und die nachfolgende Entwicklung der Chancenstruktur in Berufungsverfahren trennen kann. In der niederländischen Studie finden sich Hinweise auf eine Selbstselektion in Form eines geringeren Anteils an Frauen bei der Bewerbung. Frauen hatten in dieser Studie außerdem eine um elf Prozentpunkte geringere Berufungswahrscheinlichkeit als ihre männlichen Mitbewerber (van den Brink et al. 2006: 531). In der Aggregatanalyse konnten allerdings keinerlei Informationen zu den Leistungen der Bewerber/innen einbezogen werden. Generell wurde das Konzept der *labor queues* unseres Wissens in der Wissenschaftssoziologie bislang nicht untersucht, obwohl es sich u. E. sehr gut auf die Situation von Berufungen übertragen lässt.

## 5.4. Daten und Analyseverfahren

Die Grundgesamtheit des vorliegenden Beitrags umfasst alle Berufungsverfahren einer deutschen Universität von 2001 bis 2013. Die Universität ist mit knapp 12.000 Studierenden mittelgroß und deckt nicht das gesamte Fächerspektrum ab. Es fehlen Medizin, Ingenieurwissenschaften (außer Informatik) und Theologie. Für die hier beabsichtigten Kausalanalysen ist diese Fächerbeschränkung unerheblich, da nichtsdestotrotz in den Geistes-, Sozial- sowie Naturwissenschaften Fächer mit sehr unterschiedlichen Frauenanteilen vertreten sind.

Für die nachfolgenden Analysen werden zwei anonymisierte Datensätze zu den Berufungsverfahren verwendet: ein Datensatz mit Informationen über die Verläufe der Berufungsverfahren (*Verfahrensdaten*) sowie ein Datensatz mit den auf dem abschließenden Berufungsvorschlag befindlichen Bewerber/innen, inkl. Publikations- und Forschungsleistungen sowie weiteren Informationen aus dem Lebenslauf (*Listenplatzdaten*). Die Grundlage beider Datensätze bilden Senatsprotokolle, Senatsvorlagen und Erfassungsbögen des Gleichstellungsreferats der Universität zum Verfahrensverlauf. Die vorliegenden prozessgenerierten Daten haben den Vorteil, dass sie nicht unter *non-response* Problemen leiden, wie sie in Befragungsstudien üblicherweise vorkommen, wenn vornehmlich erfolgreiche Personen teilnehmen. Trotzdem waren einige Informationen zu wenigen Verfahren bzw. Personen nicht mehr vollständig zu recherchieren.

**Abbildung 5-1 Überblick zu Verfahrens- und Listenplatzdaten**




*Anmerkung*: *N* ist die Fallzahl für Verfahren; *B* die Fallzahl für Bewerber/innen.

Insgesamt liegen für den Analysezeitraum Informationen aus 259 abgeschlossenen Verfahren vor. Über den gesamten Verlauf des Bewerbungsverfahrens bilden die *Verfahrensdaten* den Frauenanteil auf folgenden fünf Stufen ab: (1) Bewerbung, (2) Erstauswahl, (3) Berufungsvortrag, dem (4) Berufungsvorschlag (Liste) und (5) dem ersten Listenplatz. Im Vergleich zu der erwähnten Vorläuferstudie von van den Brink et al. (2006) sind zusätzlich zwei weitere Verfahrensstufen, die Erstauswahl und der Vorstellungsvortrag, erfasst. Durch diese zwei zusätzlichen Verfahrensstufen ist die deutliche Selektion der Bewerbungen (von teils über 100 eingegangenen auf gewöhnlich drei Listenplatzierte) genauer nachzuvollziehen. Diese *Verfahrensdaten* wurden mit Indikatoren für den Frauenanteil unter dem in der jeweiligen Disziplin und im Bewerbungsjahr bestehenden Bewerberpool angereichert (Schätzung mittels Personen, die im Bewerbungsjahr oder Vorjahr ihre Promotion bzw. Habilitation abgeschlossen haben; eigene Sonderauswertung von Daten des Statistischen Bundesamtes).[89] Insgesamt stehen für 235 Verfahren komplette Daten zur Verfügung. Um den Verfahrensablauf abzubilden, wird jede Verfahrensstufe als eigener Fall behandelt. Insgesamt besteht der Datensatz mit 235 Berufungsverfahren bei fünf Verfahrensstufen aus 1.175 Fällen.

Die *Listenplatzdaten* enthalten anonymisierte Informationen zu denjenigen Personen, die letztlich auf den Berufungslisten platziert wurden. Bei 177 Verfahren (mit $B = 484$ listenplatzierten Bewerber/innen) konnte hierfür auf die vollständige Information aus den bei der Bewerbung eingereichten Lebensläufen

[89] Speziell W1-Professuren setzen keine Habilitation voraus, aber auch in den anderen Verfahren bildet die Habilitation oftmals (und zunehmend) kein Anforderungskriterium mehr, weshalb alternativ zu den Habilitationen die Promotionen als Indikator für den Bewerberpool betrachtet werden. Die Schätzung der Frauenanteile erfolgte auf der Ebene von Fächern, die eine möglichst hohe Überschneidung mit der Fächerkategorie des Statistischen Bundesamtes aufweisen; der Zuordnungsschlüssel ist auf Anfrage erhältlich. Verfahren mit einer Bewerbungsfrist bis einschließlich Juni werden dem Bewerberpool des Vorjahrs, Verfahren mit einer Bewerbungsfrist ab Juli dem Bewerberpool des aktuellen Jahres zugerechnet.

zurückgegriffen werden.[90] Die *Verfahrensdaten* sowie die *Listenplatzdaten* überlappen dabei stark, sodass für 174 Verfahren sowohl Verfahrens- als auch Listenplatzinformationen komplett vorliegen. Nachfolgend wird auf die kompletten *Listenplatzdaten* zurückgegriffen, um nicht zu viele Fälle aus der Analyse auszuschließen.[91] Die vorliegenden Daten sollten für die beabsichtigten Analysen ein weitgehend verzerrungsfreies Bild liefern. Neben dem das „akademische Alter" definierenden Jahr der Promotion wurden zusätzlich der Familienstand, im Lebenslauf aufgeführte Kinder sowie die im Lebenslauf angegebenen Publikationen in den Datensatz aufgenommen.

Bei den Publikationen wurde zwischen Monographien, Sammelbandbeiträgen sowie Zeitschriftenartikeln mit *peer review* unterschieden. Dazu wurden aus den Lebensläufen der Kandidat/innen die dort genannten Monographien bzw. Sammelbände erfasst. Um eine möglichst einheitliche Definition von Zeitschriftenartikeln mit *peer review* zu erreichen, wurden für alle Bewerber/innen mit Lebenslaufdaten alle bis zum Bewerbungszeitraum veröffentlichten, in *Scopus* indizierten Zeitschriftenbeiträge exportiert.[92] Es stehen somit für 177 Verfahren auch Angaben zur Art und Anzahl der Publikationen sowie zu deren Rezeption zur Verfügung. Bedingt durch die im Folgenden näher erläuterte Modellierung fallen Verfahren mit nur einem Listenplatz aus der Analyse. Aus diesem Grund reduziert sich die Fallzahl in den Regressionsanalysen auf 158 Verfahren.

**Statistische Analyseverfahren**

Es wird in zwei Schritten vorgegangen: Zunächst werden die Veränderungen von Frauenanteilen über die Verfahrensstufen analysiert. Hier interessiert, ob es Anzeichen für eine Selbst- oder Fremdselektion gibt, und ob sich die Frauenanteile speziell auf einzelnen Verfahrensstufen verändern, wie das die theoretischen Annahmen erwarten lassen. Bei diesen Analysen bilden die einzelnen Verfahrensstufen die Analyseeinheit. Daten zu den Leistungen der einzelnen Bewerber/innen lassen sich noch nicht einbeziehen.

Diese spielen dann im zweiten Schritt, bei der Analyse der Kandidat/innen auf den Listenplätzen, eine Rolle. Hier bilden die einzelnen Bewerber/innen die Fälle; das Erklärungsziel sind ihre Erfolgschancen

[90] Die fehlenden Verfahren erklären sich durch teilweise unvollständige oder fehlende Archivdaten.

[91] Da der Ausfall zwischen den stattgefundenen Verfahren sowie den Verfahren mit verfügbaren Informationen zu den Listenplatzierten mit 82 Verfahren sehr groß ist, wurden die Daten auf mögliche selektive Ausfälle im Hinblick auf den Fachbereich und das Verfahrensjahr untersucht. Beide Datensätze zeigten dabei keine Anzeichen für einen selektiven Ausfall im Hinblick auf das Fach. Dies ändert sich leicht im Hinblick auf das Verfahrensjahr: Sowohl die *Verfahrensdaten* ($\chi^2 = 28{,}779$; df = 12; $p = 0{,}004$) als auch die *Listenplatzdaten* ($\chi^2 = 40{,}985$; df = 12; $p < 0{,}001$) zeigen hier leicht systematische Ausfälle. Vor allem im Jahr 2006 steht ein höherer Anteil an Verfahren nicht für die Analysen zur Verfügung.

[92] In Scopus werden Artikel aus Zeitschriften mit *peer review*, die mindestens einen englischen Titel sowie Abstract haben, aufgeführt (vgl. Scopus Content Coverage Guide). Scopus deckt die Naturwissenschaften ebenso wie die Geistes- und Sozialwissenschaften ab. Explorative Zitationsanalysen für einen Teil der Verfahren mit Google Scholar Daten, die den „Buchmarkt" stärker einbeziehen, haben zu sehr ähnlichen Ergebnissen geführt. Eingeschlossen wurden jeweils alle Publikationen bis zum Vorjahr der Bewerbung, sofern der Bewerbungsstichtag vor Juli war. Im Falle eines Bewerbungsschlusses nach Juli wurden auch die Publikationen des Bewerbungsjahres berücksichtigt.

im Hinblick auf den ersten Listenplatz. Es interessiert, welche (geschlechtsspezifische) Erklärungsrelevanz Leistungsmerkmale wie das Publikationsprofil haben; welchen Einfluss das Geschlecht unter Kontrolle dieser Merkmale (und weiterer Kontrollvariablen wie dem akademischen Alter) noch hat; und inwiefern sich Ergebnisse ändern, wenn für die spezielle Konkurrenzsituation in den einzelnen Verfahren kontrolliert wird.

Beide Analyseschritte erfordern aufgrund der Datenstruktur spezielle statistische Verfahren. Für die im ersten Schritt erfolgenden Auswertungen der *Verfahrensdaten* werden *fractional response* Logit-Modelle verwendet (Papke & Wooldridge 1996). Diese modellieren in der funktionalen Form eines herkömmlichen Logit-Modells Anteilswerte anstatt binärer Outcomes.[93]

Um die Frauenanteile in den einzelnen Stufen der Berufungsverfahren als abhängige Variable zu modellieren, stehen pro Berufungsverfahren die Frauenanteile auf den berichteten fünf Verfahrensstufen zur Verfügung. Die fünfte und letzte Stufe des Verfahrens (erster Listenplatz) ist dabei im Gegensatz zu den vorangehend verwendeten Frauenanteilen nur noch das dichotom kodierte Geschlecht der Person auf dem ersten Listenplatz. In diesem Fall reduziert sich das fraktionale Logit-Modell auf das gewöhnliche Logit-Modell.

Zur Untersuchung einer geringeren Bewerbungswahrscheinlichkeit von Wissenschaftlerinnen (Selbstselektion) werden der Bewerberpool und die erste Verfahrensstufe (Bewerbung) betrachtet. Kontrolliert werden das Fach sowie das Verfahrensjahr. Dabei werden separate Modelle für zwei unterschiedliche Indikatoren des Bewerberpools geschätzt: dem Frauenanteil an den Promotionen bzw. an den Habilitationen im jeweiligen Bewerbungszeitraum (Bewerbungsjahr und Vorjahr, s. Angaben zur Datengrundlage).

Um die weiteren Annahmen zur geschlechtsspezifischen Chancenstruktur gemäß Tabelle 5-1 testen zu können, werden analog zum vorangehenden Modell die vier Verfahrensstufen nach der Bewerbung herangezogen und mögliche Veränderungen des Frauenanteils mittels Stufen-Dummies geschätzt. Die theoretischen Überlegungen erfordern dabei oftmals Tests der Koeffizienten über einzelne Verfahrensstufen. So wird etwa im Rahmen der statistischen Diskriminierung angenommen, dass sich der Frauenanteil insbesondere auf den ersten Stufen der Verfahren reduziert, wenn zu den Bewerber/innen noch vergleichsweise wenig Information vorliegt. Getestet wird konkret, ob sich die Koeffizienten über die Verfahrensstufen unterscheiden (Nullhypothese wäre die Gleichheit). Hierzu werden sog. Wald-Tests verwendet (Wooldridge 2013: 588f.). Als Kontrollvariablen werden das Jahr des Berufungsverfahrens, die Wertigkeit der Professur (Juniorprofessur ja/nein) und der vorhandene Bewerberpool in das Modell auf-

[93] Im Gegensatz zu linearen Modellen mit Anteilen als abhängiger Variable erlauben es fraktionale Logit-Modelle den Zusammenhang zwischen Wahrscheinlichkeit und unabhängigen Variablen nichtlinear zu modellieren (ebenso wie binäre Logit-Modelle im Gegensatz zu linearen Wahrscheinlichkeitsmodellen), was insbesondere das Problem von vorhergesagten Werten außerhalb des möglichen Wertebereiches für Anteilswerte und Wahrscheinlichkeiten [0;1] vermeidet (Papke & Wooldridge 1996: 621).

genommen, um Unterschiede, welche bereits vor Verfahrensbeginn bestehen, nicht dem Verfahrensprozess zuzuschreiben. Ebenso wird die Anzahl der in den Verfahrensstufen befindlichen Personen kontrolliert. Um die Resultate abzusichern, schätzen wir zusätzlich ein Modell zur Berücksichtigung von Heteroskedastizität. Die Varianz der Residuen kann sich bekanntlich in Abhängigkeit von den unabhängigen Variablen unterscheiden (Papke & Wooldridge 1996; Williams 2009). Konkret kann sie im vorliegenden Fall z. B. von den Verfahrensschritten beeinflusst sein. Die von Papke & Wooldridge (1996) vorgeschlagene und hier angewandte Modellierung lässt es zu, den Effekt einer möglichen Heteroskedastizität direkt in der Modellierung zu berücksichtigen. Anders als bei der linearen Regression (OLS) können bei fehlender Modellierung von Heteroskedastizität in Logit-Modellen nicht nur ineffiziente, sondern sogar verzerrte Schätzer resultieren (Williams 2009: 554). Ein Problem des heteroskedastischen Modells ist allerdings dessen ausgeprägte Sensitivität gegenüber einer Fehlspezifikation und eine geringere Effizienz der Schätzungen bei Fallzahlen von weniger als 250 unabhängigen Beobachtungen (Keele & Park 2005: 11-13). Daher schätzen wir als weiteren Robustheitscheck ein lineares Wahrscheinlichkeitsmodell mit fixen Effekten (FE-LPM), welches alle verfahrenskonstanten Merkmale kontrolliert.[94] Ergebnisse werden in Form durchschnittlicher Marginaleffekte (*average marginal effects*, kurz AMEs) berichtet. Dieses Vorgehen erscheint insbesondere zum Vergleich der verschiedenen Modellspezifikationen zwingend (Auspurg & Hinz 2010; Mood 2010).

Die bisher gewählte Modellierung erlaubt es, Veränderungen des Frauenanteils über die Verfahrensstufen zu modellieren. Ebenso interessant sind Unterschiede zwischen Bewerber/innen, die es auf die Liste geschafft haben; für diese *Listenplatzdaten* liegen wie erwähnt umfangreiche Information aus den CVs vor. Um mögliche Differenzen in der Publikationsleistung von Listenkandidaten zu untersuchen, werden lineare Regressionen eingesetzt. Für die Analyse der unterschiedlichen Erfolgschancen, den ersten Listenplatz zu erreichen, werden dagegen Logit-Modelle geschätzt. Als unabhängige Variable werden neben den Lebenslaufdaten die Zitationen der in der Literaturdatenbank *Scopus* erfassten Publikationen mit *peer review* verwendet. Um die Eigenheiten bestimmter Verfahren zu berücksichtigen, werden mittels konditionaler *fixed effects*-Mehrebenenmodelle (Logit) alle verfahrenskonstanten Variablen kontrolliert (vgl. Fernandez & Mors 2008: 1074). Zeigen sich auch nach Kontrolle für die einzelnen Verfahren noch Unterschiede im Vorankommen von Wissenschaftlerinnen und Wissenschaftlern, können diese nicht mehr durch Kompositionseffekte, also eine unterschiedliche Verteilung von Männern und Frauen auf Verfahren mit jeweils anderen Auswahlregeln, Kandidatenpools und sonstigen verfahrens-

[94] Aufgrund der Verzerrung von *fixed-effects*-Dummies in nichtlinearen Modellen (Arellano & Hahn 2005) wird dabei auf ein lineares Modell zurückgegriffen. Die Robustheitschecks beziehen sich dabei nur auf die Richtung und Signifikanz der Effekte. Die vollen Modelle sind auf Anfrage erhältlich. Beim linearen Wahrscheinlichkeitsmodell wird die Heteroskedastizität für die Dummyvariablen der letzten beiden Verfahrensstufen kontrolliert (vierte Stufe: Liste; fünfte Stufe: erster Listenplatz – für die anderen Stufen gab es keine Hinweise auf Heteroskedastizität). Für die Schätzung des *fractional-response* Logit Modells wird das *fracglm* stata-ado (Williams 2015) verwendet.

konstanten Merkmalen bedingt sein. Ergebnisse werden in *odds ratios* berichtet, da bei einer Modellierung mit *fixed effects* keine durchschnittlichen Marginaleffekte mehr schätzbar sind. In Modellen ohne *fixed effects* wird die Mehrebenenstruktur der Daten mit geclusterten Standardfehlern adressiert.[95]

## 5.5. Ergebnisse

### *5.5.1. Verfahrensdaten*

Zunächst interessiert, wie sich der Frauenanteil über die Verfahrensstufen hinweg entwickelt. Schon in einer ersten Deskription lässt sich feststellen, dass die Frauenanteile bei Bewerbung (25 Prozent) die der Qualifizierungsstufen Promotion (40 Prozent) und Habilitation (27 Prozent) nicht erreichen (vgl. Tabelle 5-2). Nach einem mittleren Frauenanteil von 25 Prozent bei Bewerbungen steigt dieser im Verfahrensverlauf auf zunächst 34 Prozent an (Erstauswahl), und nimmt dann auf den weiteren Stufen wieder leicht ab. Die Verfahren differieren untereinander stark, was sich an der hohen Streuung zeigt. Zugleich schwankt die Zahl der Verfahren über die Jahre hinweg (die Spannweite reicht von 4 Verfahren im Jahr 2001 bis 29 Verfahren im Jahr 2002; in der Tabelle nicht dargestellt). Insgesamt 26 Prozent der Verfahren sind für Juniorprofessuren ausgeschrieben. Auch die Anzahl der Bewerbungen variiert immens, so gibt es auf einzelne Verfahren nur eine/n einzige/n Bewerber/in, wohingegen in anderen Verfahren bis zu 149 Bewerber/innen um die offene Position konkurrieren. Im Mittel bewerben sich 41 Personen, was eine deutliche Bewerberselektion in den einzelnen Verfahrensstufen erfordert. Anhand der Fächerverteilung zeigt sich zudem, dass die Natur-, Geistes- und Sozialwissenschaften mit annähernd gleicher Häufigkeit vertreten sind.

**Tabelle 5-2 Deskriptive Ergebnisse vollständige Verfahrensdaten (FA=Frauenanteil)**

| Variablen | MW | SD | Min | Max |
|---|---|---|---|---|
| FA Pool (Promotionen) | 0,403 | 0,157 | 0,098 | 0,698 |
| FA Pool (Habilitationen) | 0,271 | 0,134 | 0,032 | 0,592 |
| FA Bewerbungseingang | 0,252 | 0,194 | 0 | 1 |
| FA Erstauswahl | 0,340 | 0,267 | 0 | 1 |
| FA Vortrag | 0,329 | 0,262 | 0 | 1 |
| FA Liste | 0,319 | 0,330 | 0 | 1 |
| FA Erstplatzierte | 0,306 | | | |
| Verfahrensjahr | 2007 | 3,622 | 2001 | 2013 |
| Juniorprofessur | 0,264 | | | |
| Anzahl Bewerber/innen | 40,770 | 29,748 | 1 | 149 |
| *Disziplin/Fachbereich* | | | | |
| Naturwissenschaften | 0,336 | | | |
| Geisteswissenschaften | 0,310 | | | |
| Sozialwissenschaften | 0,353 | | | |
| *N* | | 235 | | |

Nach dieser ersten Deskription sollen nun die vermuteten Zusammenhänge multivariat getestet werden. Zu Beginn wird dabei die Selbstselektion von möglichen Bewerber/innen untersucht. Der deskriptive

[95] Im Fall der *Verfahrensdaten* sind mehrere Verfahrensstufen in einzelnen Verfahren geclustert, im Fall der *Listenplatzdaten* mehrere Bewerber/innen in einzelnen Verfahren.

Befund bleibt stabil (vgl. Tabelle 5-3), so ist der Frauenanteil beim Bewerbungseingang unter Kontrolle des Verfahrensjahrs sowie des Fachbereichs im Mittel um 15 Prozentpunkte geringer als bei den Promotionen. Dieser Befund ist hoch signifikant. Bei Verwendung der alternativen Operationalisierung über die Habilitationen bleibt die Differenz zwischen dem Frauenanteil an den Bewerbungen und dem Bewerberpool auf dem 10%-Niveau signifikant, sie ist jedoch mit einem Umfang von nurmehr knapp 2 Prozentpunkten deutlich reduziert. Insgesamt lässt sich gleichwohl festhalten, dass Frauen sich mit geringerer Wahrscheinlichkeit auf die ausgeschriebenen Stellen bewerben, als es ihrem Anteil im Bewerberpool, hier angenähert über Promotionen und Habilitationen, entspricht (Einschränkungen der Operationalisierung werden abschließend diskutiert).

**Tabelle 5-3 Frauenanteil Bewerbungen und Frauenanteil im Pool (erste Verfahrensstufe, *fractional response* Logit, AME)**

| | (1) Promotionen | (2) Habilitationen |
|---|---|---|
| FA | -0,150*** | -0,019+ |
| | (-13,86) | (-1,81) |
| *N* | 470 (235) | |

*Anmerkungen:* Verfahrensdaten; z-Werte in Klammern, geclusterte Standardfehler.
Kontrollvariablen: Jahr des Verfahrens, Fachbereich.
+ $p < 0{,}10$, * $p < 0{,}05$, ** $p < 0.01$, *** $p < 0{,}001$

Für die folgende Betrachtung der Veränderung des Frauenanteils über die verschiedenen Verfahrensstufen (vgl. Tabelle 5-4) soll auf das heteroskedastische Logit-Modell sowie das *fixed-effects* LPM-Modell – wie erläutert – nur als Robustheitscheck Bezug genommen werden. Lediglich auf der vierten und fünften Stufe tragen die Varianzkomponenten des heteroskedastischen Modells (genauer des Listen- und des Erstplatzierten-Dummies) zur Verbesserung des Modells bei ($\chi^2$=9,15; df = 2; $p$=0,01).

Im zweiten Verfahrensschritt, der Erstauswahl, sind Frauen um 9,1 Prozentpunkte häufiger vertreten als bei den Bewerbungen. Im weiteren Verfahrensverlauf ist die Tendenz leicht rückläufig. So liegt der Frauenanteil beim Vortrag um gut 8, bei der Liste um 7, und schlussendlich bei den Erstplatzierten um knapp 6 Prozentpunkte über dem Frauenanteil bei den Bewerbungen. Insgesamt sind Bewerberinnen demnach im gesamten Verfahren im Verhältnis zu ihrem Anteil an den Bewerbungen überrepräsentiert. In diesem Zusammenhang sei allerdings an die soeben vorgestellten Ergebnisse zur Unterrepräsentation von Wissenschaftlerinnen bei den Bewerbungen hingewiesen. Ohne Kontrolle auf die Leistung ist nicht zu sagen, ob hier eine Positivdiskriminierung zu beobachten ist oder die Überrepräsentation von Frauen lediglich eine Selbstselektion besonders leistungsfähiger Bewerberinnen widerspiegelt. Hinweise dazu bieten die späteren Auswertungen der *Listenplatzdaten*.

Hier sollen zunächst noch die Annahmen der verschiedenen Diskriminierungstheorien geprüft werden. Statistische Diskriminierung nach Geschlecht wäre insbesondere auf den ersten Verfahrensstufen zu erwarten. Ein Wald-Test prüft die Gleichheit der beiden ersten Verfahrensstufen (Erstauswahl und Vortrag) mit den letzten beiden (Liste und erster Listenplatz), bei denen, insbesondere bedingt durch den Bewerbungsvortrag sowie das anschließende Gespräch, erheblich mehr Informationen über die

Bewerber/innen verfügbar sein sollten. Hierbei zeigt sich kein signifikanter Unterschied zwischen den zwei Koeffizientengruppen. Zudem ist der Zusammenhang entgegengesetzt zu der vermuteten Tendenz, da Frauen in den letzten beiden Verfahrensstufen zu einer leicht *geringeren* Chance vertreten sind als in den vorgelagerten Verfahrensstufen. Ähnlich sind die Überlegungen zu *reward expectation* bzw. *status construction* schon allein durch das konträre Ergebnis von im Vergleich zu den Bewerbungen über- statt unterrepräsentierten Bewerberinnen widerlegt – wiederum gilt allerdings abzuwarten, ob das auch nach Kontrolle von Leistungen noch gilt. Der Test auf *stereotype threat* vergleicht den Unterschied zwischen der Einladung zum Vortrag und der daraus resultierenden Liste. Zwischen diesen Stufen ist zwar in der Tat ein leichter Rückgang des Frauenanteils festzustellen, dieser ist jedoch nicht statistisch signifikant.

**Tabelle 5-4 Entwicklung der Frauenanteile über weitere Verfahrensstufen**

| | (1) Fractional resp. Logit (AMEs) | | (2) Heterosked. Logit (AMEs) | | (3) FE-LPM (AMEs) | |
|---|---|---|---|---|---|---|
| *(Referenz: Bewerbung)* | | | | | | |
| Erstauswahl | 0,091*** | (7,11) | 0,105*** | (4,73) | 0,088*** | (4,53) |
| Vortrag | 0,081*** | (7,07) | 0,093*** | (4,45) | 0,077*** | (3,95) |
| Liste | 0,071*** | (3,97) | 0,080*** | (3,66) | 0,067*** | (3,43) |
| 1. Listenplatz | 0,059* | (2,13) | 0,065* | (2,35) | 0,055** | (2,80) |
| *N* | | | 1175 (235) | | | |
| Wald-Tests (Chi²/*F*-Tests auf Heteroskedastizität bzw. Gleichheit von Koeffizienten) | | | | | | |
| Heteroskedastizität Residuen | | | $\chi^2$ (2) = 9,15 $p$ = 0,01* | | | |
| Statistische Diskriminierung[a] | $\chi^2(1)$ = 1,37 $p$ = 0,242 | | $\chi^2$ (1) = 1,95 $p$ = 0,163 | | $F(1,936)$ = 2,52 $p$ = 0,113 | |
| *reward expectation*[b] | $\chi^2(4)$ = 61,42 $p < 0{,}001^{***}$ | | $\chi^2$ (4) = 31.51 $p < 0{,}001^{***}$ | | $F(4,936)$ = 6,22 $p < 0{,}001^{***}$ | |
| *stereotype threat*[c] | $\chi^2(1)$ = 0,47 $p$ = 0,494 | | $\chi^2$ (1) = 0,27 $p$ = 0,600 | | $F(1,936)$ = 0,27 $p$ = 0,600 | |

*Anmerkungen*: Verfahrensdaten; *z*/*t*-Werte in Klammern, geclusterte Standardfehler.
Kontrollvariablen: Jahr des Verfahrens, Juniorprofessur, Frauenanteil Promotionen, Frauenanteil Habilitationen, Personen in Verfahrensstufe. [a] $H_0$: Effekte Erstauswahl und Vortrag = Effekte Liste und erster Listenplatz. [b] $H_0$: Effekte Erstauswahl = Vortrag = Liste = erster Listenplatz = 0. [c] $H_0$: Vortrag = Liste.
$p < 0{,}10$, $^{*}$ $p < 0{,}05$, $^{**}$ $p < 0{,}01$, $^{***}$ $p < 0{,}001$

Logit- und heteroskedastisches Logit-Modell sowie das auf alle verfahrenskonstanten Merkmale kontrollierende *fixed-effects* LPM sind hinsichtlich der Analyse- und Kontrollvariablen in Vorzeichen und Signifikanz robust.

### *5.5.2. Listenplätze*

Für die Analysen der *Listenplatzdaten* lassen sich Informationen aus 158 Verfahren zu insgesamt 449 Bewerber/innen verwenden (vgl. Abbildung 5-1). Im Mittel publizierten Personen auf der Liste knapp 1,5 Monographien, acht Sammelbandbeiträge und 15 Artikel in Fachzeitschriften mit *peer review* (für eine Aufschlüsselung nach Geschlecht s. Tabelle 5-5).[96] Alle Publikationsdaten streuen allerdings breit.

[96] Koautorenschaften werden wie Alleinautorenschaften gezählt (da die Arbeitsanteile bei Mehrfachautoren kaum aus den CVs ablesbar sind), ebenso wie aufgrund der vielfach möglichen Operationalisierungen auf eine Einteilung in A-, B-, und C-Journals oder auch Gewichtung nach der Länge von Beiträgen verzichtet wird.

Das akademische Alter ist definiert als die Jahre, die seit der Promotion verstrichen sind (globaler Mittelwert: acht Jahre; Median: sieben Jahre). Die Anzahl der in Bewerbungen genannten Kinder beläuft sich im Mittel auf 0,3 (SD: 0,7).

**Tabelle 5-5 Deskriptive Ergebnisse nach Geschlecht (vollständige Listenplatzdaten)**

| | Männer | | | | Frauen | | | |
|---|---|---|---|---|---|---|---|---|
| | MW | SD | Min | Max | MW | SD | Min | Max |
| Monographien | 1,497 | 1,991 | 0 | 14 | 1,482 | 1,705 | 0 | 8 |
| Sammelbandbeiträge | 7,620 | 10,413 | 0 | 73 | 8,340 | 11,269 | 0 | 72 |
| Artikel (Scopus) | 18,873 | 30,717 | 0 | 211 | 6,582 | 10,744 | 0 | 61 |
| Akad. Alter | 8,331 | 4,983 | 0 | 25 | 7,312 | 4,612 | 0 | 20 |
| Anzahl Kinder | 0,299 | 0,728 | 0 | 3 | 0,255 | 0,578 | 0 | 2 |
| *N* Bewerber/innen | | 308 | | | | 141 | | |
| *N* Verfahren | | | | 158 | | | | |

Zunächst wird der Frage nachgegangen, ob Bewerberinnen auf der Liste weniger publizieren als ihre männlichen Mitbewerber. In Tabelle 5-6 werden die unter Kontrolle auf die Kinderanzahl sowie das akademische Alter zu beobachtenden Unterschiede berichtet. Männer publizieren, wenn auch nicht signifikant, im Mittel einen Sammelbandbeitrag mehr (Modell 1). Bei den Monographien (Modell 3) haben die Bewerberinnen etwa gleich viel veröffentlicht wie männliche Bewerber. Der größte und auf dem 0,1%-Niveau signifikante Unterschied zeigt sich bei den Publikationen mit *peer review*: Männliche Listenkandidaten haben hier im Schnitt elf Publikationen mehr aufzuweisen (Modell 5).

Die bislang präsentierten Modelle berücksichtigen jedoch noch nicht die *labor queue* Struktur von Berufungsverfahren. Unterschiede können somit reine Kompositionseffekte, fachspezifische Publikationspraktiken oder Variationen in der Bewerbungslage widerspiegeln. Bleiben die Ergebnisse stabil, wenn man für solche Unterschiede zwischen den einzelnen „Bewerberturnieren" kontrolliert? Tatsächlich ändern sich die Ergebnisse unter Kontrolle der einzelnen Verfahren mittels *fixed effects* (FE) Mehrebenenmodellen. Bei den Monographien zeigt sich nun, dass Bewerberinnen rund 0,34 Monographien (und damit auf einem 10%-Niveau statistisch signifikant) weniger veröffentlichen als ihre männlichen Mitkonkurrenten im selben Verfahren. Am deutlichsten ändert sich bei Berücksichtigung fixer Verfahrenseffekte der bislang größte und zuvor höchst signifikante Effekt des Geschlechts auf die Anzahl der Publikationen mit *peer review*. Frauen haben zwar immer noch im Mittel 3,9 Zeitschriftenpublikationen weniger als ihre männlichen Konkurrenten, jedoch ist dieser Unterschied nur noch auf dem 10%-Niveau signifikant (Tabelle 5-6). Dabei ist auch der Rückgang der Effektstärke mit einem Umfang von knapp 2/3 beachtlich. Bei den Kontrollvariablen hat erwartungsgemäß das akademische Alter einen positiven Effekt. Die Anzahl der Kinder hat hingegen keinen Einfluss auf die wissenschaftliche Produktivität. Die Ergebnisse bleiben zudem auch ohne Kontrolle auf das akademische Alter sowie die Kinderanzahl stabil.

**Tabelle 5-6 Anzahl der Veröffentlichungen (OLS, Listenplatzdaten)**

| | (1) | (2) | (3) | (4) | (5) | (6) |
|---|---|---|---|---|---|---|
| | Sammelbandbeiträge | | Monographien | | Artikel mit *peer review* | |
| | OLS | FE | OLS | FE | OLS | FE |
| Bewerberin | -1,011 | 1,606 | 0,097 | -0,342+ | -11,062*** | -3,874+ |
| | (-0,96) | (1,60) | (0,52) | (-1,79) | (-4,24) | (-1,73) |
| Akad. Alter | 0,639*** | 0,833*** | 0,104*** | 0,0563* | 1,139*** | 1,382*** |
| | (4,62) | (8,49) | (5,71) | (2,26) | (4,47) | (4,72) |
| Anzahl Kinder | 1,048 | 0,843 | 0,114 | 0,180 | 1,600 | -1,118 |
| | (1,45) | (1,21) | (0,88) | (1,39) | (0,88) | (-0,73) |
| *N* Bewerber/innen | 449 | | | | | |
| *N* Verfahren | 158 | | | | | |

*Anmerkungen*: Listenplatzdaten; *t*-Werte in Klammern, geclusterte Standardfehler.
$^{+}$ $p < 0{,}10$, $^{*}$ $p < 0{,}05$, $^{**}$ $p < 0{,}01$, $^{***}$ $p < 0{,}001$

Insgesamt lässt sich feststellen, dass sich die Leistungen von listenplatzierten Bewerberinnen und Bewerbern *innerhalb* der Verfahren nur geringfügig unterscheiden, auch wenn es *zwischen* den Verfahren Unterschiede gibt, die bei der bislang üblichen, undifferenzierten Aggregatbetrachtung fälschlich als Ungleichbehandlungen gedeutet werden können.

Wie jedoch werden die Leistungen von Bewerber/innen im Berufungsverfahren bewertet und zeigen sich Ungleichbehandlungen bei der Auswahl der/des Erstplatzierten? In Tabelle 5-7 zeigt sich, dass Frauen unter Berücksichtigung der Publikationen, des akademischen Alters sowie der Kinderanzahl eine (statistisch nicht signifikante) *leicht geringere* Chance auf den ersten Listenplatz haben als Männer (odds ratio < 1). Die Hinzunahme der Publikationsvariablen verkleinert dabei die Differenz, lässt aber die Chancen von Bewerberinnen immer noch hinter denen von Bewerbern zurück (mit einem Faktor von knapp 0,9). Interessant an diesen Modellschätzungen ist, dass Publikationen mit Ausnahme der Monographien keinen Einfluss auf die Chance haben, auf den ersten Listenplatz gesetzt zu werden.[97] Die als Kontrollvariablen eingeschlossenen Variablen des akademischen Alters und der Kinderanzahl haben keinen Effekt auf die Platzierungschance mit Ausnahme des auf dem 10%-Niveau signifikanten negativen Effekts des akademischen Alters im Logit-Modell (Modell 3). Die Ergebnisse unterscheiden sich kaum zwischen gewöhnlichem Logit und FE-Logit. Unabhängig vom Verfahrenskontext sind also die Faktoren der Erstplatzierung kaum in den erfassten Leistungsparametern auszumachen. In Berufungsverfahren spielen auf der *letzten Auswahlstufe* andere Argumente als ein einfacher Vergleich des Outputs an publizierten Artikeln die entscheidende Rolle. Dies kann einerseits bedeuten, dass sich die Kommission gar nicht an diesen Kriterien orientiert; oder andererseits – was in dem Kontext weitaus plausibler sein dürfte – dass sich das Bewerberfeld der Listenplatzierten in diesen Aspekten zu wenig

[97] Das Ergebnis für Monographien ist durchaus beachtenswert: Unter den Listenplatzierten kann die Autorenschaft von Monographien die Reihenfolge der Liste beeinflussen. Es handelt sich hierbei nicht um einen bloßen Fächereffekt, da der Effekt auch im FE-Modell bestehen bleibt.

unterscheidet (statistisch: zu wenig Varianz), als dass sie hilfreiche Entscheidungskriterien bilden würden. Letzteres würde nochmals für leistungsgerechte Verfahren sprechen: Die Bewerber/innen, die es auf die Liste geschafft haben, unterscheiden sich (innerhalb der einzelnen Verfahren) kaum mehr in ihren Leistungen oder Publikationsrekord, oder anders formuliert, sie sind (vermutlich aufgrund der erfolgten Auslese) ähnlich leistungsstark.

Abschließend sei dazu auch noch berichtet, dass es keinerlei signifikante Interaktionen des Geschlechts mit Leistungsindikatoren gibt (Artikel, Zitationen) – was Evidenz gegen *double standards* ist (Analysen hier nicht dargestellt, aber auf Anfrage verfügbar). Auch führen Robustheitsanalysen mit alternativen Operationalisierungen von Forschungsleistungen keineswegs zu anderen substanziellen Schlussfolgerungen (getestet wurden etwa alternative Modelle mit der Anzahl Zitationen oder dem H-Index von Bewerber/innen statt der reinen Anzahl an *peer-reviewed* Artikeln; aufgrund von hoher Kollinearität mit der Publikationsanzahl wurden diese Variablen nicht zusätzlich in die Analysen in Tabelle 5-7 aufgenommen).

**Tabelle 5-7 Erster Listenplatz (Logit, *odds ratios*, Listenplatzdaten)**

| | (1) Logit | (2) FE-Logit | (3) Logit | (4) FE-Logit |
|---|---|---|---|---|
| Bewerberin | 0,823 | 0,793 | 0,860 | 0,848 |
| | (-0,90) | (-1,01) | (-0,67) | (-0,70) |
| Akad. Alter | 0,977 | 0,973 | 0,956+ | 0,961 |
| | (-1,11) | (-0,90) | (-1,80) | (-1,17) |
| Anzahl Kinder | 0,867 | 0,850 | 0,841 | 0,832 |
| | (-0,91) | (-0,99) | (-1,08) | (-1,09) |
| Artikel (Scopus) | | | 1,005 | 1,002 |
| | | | (1,23) | (0,39) |
| Monographien | | | 1,203** | 1,225** |
| | | | (2,93) | (2,59) |
| Sammelbandbeiträge | | | 0,994 | 0,991 |
| | | | (-0,57) | (-0,69) |
| *N* Bewerber/innen | | 449 | | |
| *N* Verfahren | | 158 | | |

*Anmerkungen*: Listenplatzdaten; *odds ratios*; *z*-Werte in Klammern.
$^{+}$ $p < 0{,}10$, $^{*}$ $p < 0{,}05$, $^{**}$ $p < 0{,}01$, $^{***}$ $p < 0{,}001$

## 5.6. Zusammenfassung

Der vorliegende Aufsatz hat einen entscheidenden Schritt im akademischen Karriereverlauf von Wissenschaftlerinnen und Wissenschaftlern thematisiert: das Berufungsverfahren auf Professuren. Anhand eines umfangreichen Datensatzes mit Informationen zu über 230 Verfahren an einer deutschen Universität wurden theoretisch motivierte Annahmen zum Geschlechterbias untersucht. Bevor wichtige Ergebnisse und Limitationen der Analysen diskutiert werden, seien die Resultate knapp zusammengefasst.

Die Vergleiche des Frauenanteils an den Bewerbungen mit den Frauenanteilen unter Promovierten und Habilitierten deuten *erstens* darauf hin, dass Wissenschaftlerinnen sich zu einem geringeren Ausmaß, als es ihrem Anteil im Bewerbungspool entspricht, auf Professuren bewerben. Dieses Resultat bestätigt Ergebnisse anderer Studien, die auf die sensitive Karrierephase von Wissenschaftlerinnen nach der Promotion hinweisen. Die Promotion qualifiziert zumindest für Juniorprofessuren, welche hier mehr als ein Viertel aller Verfahren ausmachen. Aber selbst gemessen am Anteil an den Habilitationen bewerben sich Frauen zurückhaltender als Männer.

*Zweitens* zeigt sich, dass auf den Verfahrensstufen nach der Bewerbung die Frauenanteile zunächst im Mittel ansteigen und zum Ende des Verfahrens wieder leicht abfallen. Sie liegen auch bei den letzten Verfahrensstufen, den Listenplätzen und dem ersten Listenplatz, deutlich über dem Anteil an den Bewerbungen. Dies zeigt, dass in den untersuchten Verfahren Wissenschaftlerinnen auf den ersten Blick keine schlechteren Chancen haben, sich im Verfahren durchzusetzen – bis hin zum Abschluss des Auswahlprozesses und der Platzierung auf dem ersten Listenplatz.

Damit verwundert es *drittens* nicht, dass die vorgestellten theoretischen Überlegungen zur *Benachteiligung* von Wissenschaftlerinnen durchweg in Frage gestellt werden. Dies gilt zunächst für Theorien *statistischer Diskriminierung*. Gerade auf derjenigen Stufe, auf der noch vergleichsweise wenige Informationen vorliegen und man daher noch am ehesten eine solche Diskriminierung erwarten könnte (Auswahl aus allen Bewerbungen zur näheren Ansicht), ist ein *Anstieg* des Frauenanteils zu beobachten. Ebenso kritisch fällt die Bilanz zu Thesen aus *reward expectation* und *stereotype threat* Theorien aus. Gegenüber dem Frauenanteil an Bewerbungen bleiben Wissenschaftlerinnen zu einem überproportionalen Anteil in den Berufungsverfahren – und dies auch nach dem Vorstellungsvortrag, der sich im Hinblick auf Stereotypenbedrohungen negativ auswirken könnte.

*Viertens* gibt es keinerlei Hinweise, dass Frauen und Männer auf den (ersten) Listenplätzen signifikant bevor- oder benachteilt wären. Berücksichtigt man die einzelnen Verfahren, ist der Publikationsrekord sehr ähnlich. Die generell starke Vergleichbarkeit der Leistungsparameter bei den Listenplatzierten dürfte wohl auch der plausibelste Grund sein, warum zumindest die hier beobachteten Leistungskriterien bei der Auswahl von „Platz 1“ keine Rolle mehr spielen – das Bewerberfeld auf der Liste ist insgesamt sehr ähnlich aufgestellt. Die Ergebnisse sprechen soweit für eine nach Geschlecht faire Auswahlchance und entkräften die oft vermutete Geschlechterdiskriminierung beim wichtigen Karriereübergang in eine Professur.

*Schließlich* gibt es auch keine Hinweise auf *double standards*. Für die Verfahren auf den ersten Stufen und insbesondere bei der Bewerbung sind diese nicht gänzlich auszuschließen. Was man aufgrund der Analysen für die Listenplatzierten aber sagen kann: Wissenschaftlerinnen und Wissenschaftler haben fachspezifisch unterschiedlich ausgeprägte Publikationsprofile, einfach, weil weniger Wissenschaftlerinnen in Gebieten tätig sind, in denen insbesondere Aufsätze mit *peer review* als Produktivitätssignal gewertet werden. Unsere Perspektive auf Berufungsverfahren als Turniere ergibt aber auch, dass bei

einer adäquaten Berücksichtigung des jeweiligen Verfahrens keine oder nur geringe Unterschiede in den Publikationsleistungen oder ihrer Wertung bestehen.

## 5.7. Diskussion

Der Zugewinn der vorliegenden Studie besteht in der Verbindung von aussagekräftigen Daten mit einer in der sozialwissenschaftlichen Forschung gut etablierten Kausalanalyse. Die eingesetzten *fixed effects* Schätzungen erlauben es erstmals, den jeweiligen Verfahrenskontext bei Berufungen, das „Turnier", konstant zu halten. Damit werden auch mögliche Fächerunterschiede berücksichtigt. Die explizite Analyse solcher Fächerunterschiede stand allerdings nicht im Zentrum des Beitrags, sondern der möglichst genaue Blick auf die geschlechts- und turnierspezifische Chancenstruktur, für die es darauf ankommt, wer sich wann auf welche Professur bewirbt. Die Analyseebene waren die einzelnen Verfahren mit ihrer jeweiligen Konkurrenzsituation (Komposition). Je nach Verfahren wird etwa eine aktive Rekrutierung eingesetzt oder nicht, es unterscheiden sich Anforderungen an die Bewerber/innen, die Anzahl und das Ausmaß, zu dem sich leistungsfähige Kandidat/innen bewerben, und die Zusammensetzung der Berufungskommission. Diese Merkmale können *verfahrensintern* den Maßstab für erfolgreiche Bewerbungen setzen (etwa wenn Art und Umfang des wünschenswerten Publikationsprofils diskutiert und festgelegt werden). Im Vergleich zu den referierten Studien zur Stellenbesetzung in der Soziologie und Ökonomik (Lutter & Schröder 2016; Plümper & Schimmelfennig 2007) können wir die problematische Situation beim ersten Schritt der Verfahren identifizieren. Wissenschaftlerinnen bewerben sich zu einem geringeren Anteil, als es ihrem Anteil im potenziellen Pool entspricht.

Auch die vorliegende Studie weist selbstverständlich Limitationen auf. So ist die Tatsache, dass lediglich eine Universität untersucht wurde, eine wichtige Beschränkung. Möglich ist etwa, dass sich diese Universität durch besondere Verfahrensweisen auszeichnet, die eine Verallgemeinerung erschweren. Aufgrund der hohen Zahl an untersuchten Verfahren und eines beachtlichen Fächerspektrums bei einem zugleich sehr langen Beobachtungsfenster (Jahre 2001 bis 2013) halten wir die zentralen Ergebnisse für tragfähig, auch wenn sie an einer größeren Datenbasis validiert werden sollten. Weiterhin sind die eingesetzten Leistungsmessungen über Publikationen umstritten. Zumindest für Sozialwissenschaften wie die Soziologie gibt es aber Hinweise, dass eine qualitative Forschungsbewertung durch informierte *peers* stark mit der Anzahl von Aufsätzen mit *peer review* korreliert, und diese somit einen besonders aussagefähigen Leistungsindikator darstellen (Auspurg et al. 2015). Mit den eingesetzten Daten lässt sich überdies nicht aufklären, was auf der letzten Auswahlstufe für die Erstplatzierung entscheidend ist. Es ist naheliegend, dass unter den letztlich für die Berufungsliste ausgewählten Wissenschaftler/innen auch die besondere Passfähigkeit oder auch strategische Überlegungen wie die Gewinnbarkeit eine Rolle spielen können.

Auch der Blick auf die späte, aber entscheidende Karrierephase des Übergangs in eine Professur hat seinen Preis: Entscheidende Unterschiede nach Geschlecht können in früheren Karrierestadien, wie etwa zwischen Promotion und weiterem Karriereverlauf bestehen. In diesen Stadien setzt die festgestellte

Selbstselektion ein. Insgesamt erscheinen somit Initiativen gerechtfertigt, die darauf abzielen, den Frauenanteil an den Bewerbungen auf Professuren zu erhöhen. Aber anders als in der privaten Wirtschaft, wo die Chancenstrukturen vor allem beim Erklimmen der höchsten Stufen immer schiefer zu sein scheinen (Kohaut & Möller 2010), kann auf der Grundlage der hier verwendeten Daten festgestellt werden, dass Leistungsgerechtigkeit im akademischen Wissenschaftsbetrieb eher erreichbar scheint. Die Analysen zeigen, dass eine Benachteiligung nach Geschlecht für Berufungsverfahren genauso ausgeschlossen werden kann wie eine Bevorzugung. Dies ist für das Wissenschaftssystem und für Wissenschaftler/innen eine gute Botschaft.

# Literaturverzeichnis


Aigner, D.J. und G.G. Cain. 1977. Statistical Theories of Discrimination in Labor Markets. *Industrial and Labor Relations Review* 30:175-187.

Akademie für Soziologie. 2019. Wissenschaftliche Daten sind kein Privateigentum einzelner Forschender, sondern ein kollektives Gut. Die Bereitstellung von Forschungsdaten zur Nachnutzung und Replikation muss auch in der Soziologie die Norm sein *Positionspapier der Akademie für Soziologie*.

Albert, H. 1978. Science and the Search for Truth. In *Progress and Rationality in Science,* Hrsg. G. Radnitzky und G. Andersson, 203-220. Dordrecht: Springer Netherlands.

Alinaghi, N. und W.R. Reed. 2016. Meta-Analysis and Publication Bias: How Well Does the FAT-PET-PEESE Procedure Work? Workingpaper 26/2016. Christchurch. University of Canterbury.

Allgemeines Gleichbehandlungsgesetz. *In der Fassung vom 14. August 2006 (BGBl. I S. 1897), zuletzt geändert durch Artikel 8 des Gesetzes vom 3. April 2013 (BGBl. I S. 610)*.

Alper, J. 1993. The Pipeline Is Leaking Women All the Way Along. *Science* 260:409-411.

American Psychological Association. 1974. *Publication Manual of the American Psychological Association*. Washington, DC: American Psychological Association.

American Psychological Association. 1983. *Publication Manual of the American Psychological Association*. Washington, DC: American Psychological Association.

American Psychological Association. 1994. *Publication Manual of the American Psychological Association*. Washington, DC: American Psychological Association.

American Psychological Association. 2001. *Publication Manual of the American Psychological Association*. Washington, DC: American Psychological Association.

American Psychological Association. 2010. *Publication Manual of the American Psychological Association*. Washington, DC: American Psychological Association.

American Psychological Association. 2018. List of APA-Journals. http://www.apa.org/pubs/journals/browse.aspx?query=Title:*&type=journal (Zugriff: 01.10.2018).

Arellano, M. und J. Hahn. 2005. Understanding Bias in Nonlinear Panel Models: Some Recent Developments. In *Advances in Economics and Econometrics,* Hrsg. R. Blundell, W.K. Newey und T. Persson, Volume III, 381-409. Cambridge: Cambridge University Press.

Arrow, K.J. 1971. The Theory of Discrimination. In *Discrimination in Labor Markets,* Hrsg. O. Ashenfelter und A. Rees, 3-33. Princeton: Princeton University Press.

Arrow, K.J. 1973. The Theory of Discrimination. In *Discrimination in Labor Markets,* Hrsg. O. Ashenfelter und A. Rees, 193-216. Princeton: Princeton University Press.

Auspurg, K. und T. Hinz. 2010. *Antragsaktivitäten und Förderchancen von Wissenschaftlerinnen bei Einzelanträgen auf DFG-Einzelförderung im Zeitraum 2005-2008*. Bonn: Deutsche Forschungsgemeinschaft.

Auspurg, K. und T. Hinz. 2011a. What Fuels Publication Bias? Theoretical and Empirical Analyses of Risk Factors Using the Caliper Test. *Journal of Economics and Statistics* 231:636-660.

Auspurg, K. und T. Hinz. 2011b. Gruppenvergleiche bei Regressionen mit binären abhängigen Variablen – Probleme und Fehleinschätzungen am Beispiel von Bildungschancen im Kohortenverlauf. *Zeitschrift für Soziologie* 40:62-73.

Auspurg, K. und T. Hinz. 2017. Social Dilemmas in Science: Detecting Misconduct and Finding Institutional Solutions. In *Social Dilemmas, Institutions, and the Evolution of Cooperation,* Hrsg. B. Jann und W. Przepiorka, 189–214. Berlin, Boston: De Gruyter.

Auspurg, K., T. Hinz und A. Schneck. 2014. Ausmaß und Risikofaktoren des Publication Bias in der deutschen Soziologie. *Kölner Zeitschrift für Soziologie und Sozialpsychologie* 66:549-573.

Auspurg, K., T. Hinz und A. Schneck. 2017. Berufungsverfahren als Turniere: Berufungschancen von Wissenschaftlerinnen und Wissenschaftlern. *Zeitschrift für Soziologie* 46:283-302.

Auspurg, K., A. Diekmann, T. Hinz und M. Näf. 2015. Das Forschungsrating des Wissenschaftsrats für die Soziologie in Deutschland revisited. *Soziale Welt* 66:177-192

Axelrod, R. 1980a. Effective Choice in the Prisoner's Dilemma. *Journal of Conflict Resolution* 24:3-25.

Axelrod, R. 1980b. More Effective Choice in the Prisoner's Dilemma. *Journal of Conflict Resolution* 24:379-403.

Axelrod, R. und W.D. Hamilton. 1981. The Evolution of Cooperation. *Science* 211:1390-1396.
Baerlocher, M.O., J. O'Brien, M. Newton, T. Gautam und J. Noble. 2010. Data Integrity, Reliability and Fraud in Medical Research. *European Journal of Internal Medicine* 21:40-45.
Bakker, M., A. van Dijk und J.M. Wicherts. 2012. The Rules of the Game Called Psychological Science. *Perspectives on Psychological Science* 7:543-554.
Banks, G.C., S. Kepes und M.A. McDaniel. 2012. Publication Bias: A Call for Improved Meta-Analytic Practice in the Organizational Sciences. *International Journal of Selection and Assessment* 20:182-196.
Bassler, D., K.F. Mueller, M. Briel, J. Kleijnen, A. Marusic, E. Wager, G. Antes, E. von Elm, D.G. Altman und J.J. Meerpohl. 2016. Bias in Dissemination of Clinical Research Findings: Structured OPEN Framework of What, Who and Why, Based on Literature Review and Expert Consensus. *BMJ Open* 6:e010024.
Becker, G.S. 1968. Crime and Punishment: An Economic Approach. *Journal of Political Economy* 76:169-217.
Becker, G.S. 1971. *Economics of Dicrimination*. Chicago: Chicago University Press.
Begg, C.B. 1994. Publication Bias. In *The Handbook of Research Synthesis,* Hrsg. H. Cooper und L.V. Hedges, 399-409. New York: Russell Sage Foundations.
Begg, C.B. und M. Mazumdar. 1994. Operating Characteristics of a Bank Correlation Test for Publication Bias. *Biometrics* 50:1088-1101.
Berger, J. und W.J. Murray. 2006. Expectations, Status, and Behavior. In *Contemporary Social Psychological Theories,* Hrsg. P.J. Burke, 268-300. Standford: Standford University Press.
Berger, J., H.M. Fisek, R.Z. Norman und D.G. Wagner. 1985. Formation of Reward Expectations in Status Situations. In *Status, Rewards, and Influence,* Hrsg. J. Beger und M.J. Zelditch, 215-261. San Francisco: Jossey-Bass.
Berning, C.C. und B. Weiß. 2015. Publication Bias in the German Social Sciences: An Application of the Caliper Test to Three Top-Tier German Social Science Journals. *Quality & Quantity* 50:901-917.
Biemer, P.P. 2017. Errors and Inference. In *Big Data and Social Science – A Practical Guide to Methods and Tools,* Hrsg. I. Foster, R. Chani, R.S. Jarmin, F. Kreuter und J. Lane. Boca Raton: CRC Press.
Blázquez, D., J. Botella und M. Suero. 2017. The Debate on the Ego-Depletion Effect: Evidence from Meta-Analysis with the p-Uniform Method. *Frontiers in Psychology* 8:197.
Borenstein, M. 2011. *Introduction to Meta-Analysis*. Chichester: Wiley.
Borjas, G.J. und M.S. Goldberg. 1978. Biased Screening and Discrimination in the Labor Market. *The American Economic Review* 68:918-922.
Bornmann, L. 2011. Scientific Peer Review. *Annual Review of Information Science and Technology* 45:199-245.
Bornmann, L., R. Mutz, C. Neuhaus und H.-D. Daniel. 2008. Citation Counts for Research Evaluation: Standards of Good Practice for Analyzing Bibliometric Data and Presenting and Interpreting Results. *Ethics in Science and Environmental Politics* 8:93-102.
Bosco, F.A., H. Aguinis, K. Singh, J.G. Field und C.A. Pierce. 2015. Correlational Effect Size Benchmarks. *Journal of Applied Psychology* 100:431-449.
Bravetti, A. und P. Padilla. 2018. An optimal strategy to solve the Prisoner's Dilemma. *Scientific Reports* 8:1948.
Breen, R. 2018. Some Methodological Problems in the Study of Multigenerational Mobility. *European Sociological Review* 34:603-611.
Brodeur, A., M. Lé, M. Sangnier und Y. Zylberberg. 2013. Star Wars: The Empirics Strike Back. *IZA Discussion Paper Series* Nr. 7268.
Brodeur, A., M. Lé, M. Sangnier und Y. Zylberberg. 2016. Star Wars: The Empirics Strike Back. *American Economic Journal-Applied Economics* 8:1-32.
Brouns, M. 2000. The Gendered Nature of Assessment Procedures in Scientific Research Funding: The Dutch Case. *Higher Education in Europe* 25:193-199.
Brüderl, J. 2004. Meta-Analyse in der Soziologie: Bilanz der deutschen Scheidungsforschung oder "statistischer Fruchtsalat"? *Zeitschrift für Soziologie* 33:84-86.

Brüderl, J. 2013: Sind die Sozialwissenschaften wissenschaftlich? Ergebnisse eines Replikationsexperiments. Vortrag anlässlich des Seminars Rational Choice Sociology. VIU Venedig.

Bruns, S.B. und J.P.A. Ioannidis. 2016. p-Curve and p-Hacking in Observational Research. *PLOS ONE* 11:e0149144.

Bürkner, P.-C. und P. Doebler. 2014. Testing for Publication Bias in Diagnostic Meta-Analysis: A Simulation Study. *Statistics in Medicine* 33:3061-3077.

Callaham, M.L., R.L. Wears, E.J. Weber, C. Barton und G. Young. 1998. Positive-Outcome Bias and Other Limitations in the Outcome of Research Abstracts Submitted to a Scientific Meeting. *JAMA* 280:254-257.

Camerer, C.F., A. Dreber, E. Forsell, T.-H. Ho, J. Huber, M. Johannesson, M. Kirchler, J. Almenberg, A. Altmejd, T. Chan, E. Heikensten, F. Holzmeister, T. Imai, S. Isaksson, G. Nave, T. Pfeiffer, M. Razen und H. Wu. 2016. Evaluating Replicability of Laboratory Experiments in Economics. *Science* 351:1433-1436.

Carsey, T.M. und J.J. Harden. 2013. *Monte Carlo Simulation and Resampling Methods for Social Science*. Los Angeles: Sage Publications.

Ceci, S.J. und W.M. Williams. 2011. Understanding Current Causes of Women's Underrepresentation in Science. *Proceedings of the National Academy of Sciences* 108:3157-3162.

Chalmers, I. 1990. Underreporting Research is Scientific Misconduct. *JAMA* 263:1405-1408.

Chang, A. und P. Li. 2015. Is Economics Research Replicable? Sixty Published Papers from Thirteen Journals Say 'Usually Not'. *FEDS Working Paper No. 2015-083*.

Chatterjee, M. und A.K. Chakraborty. 2016. A Simple Algorithm for Calculating Values for Folded Normal Distribution. *Journal of Statistical Computation and Simulation* 86:293-305.

Cleveland, W.S. 1979. Robust Locally Weighted Regression and Smoothing Scatterplots. *Journal of the American Statistical Association* 74:829-836.

Cohen, J. 1962. The Statistical Power of Abnormal-Social Psychological Research: A Review. *The Journal of Abnormal and Social Psychology* 65:145-153.

Cohen, J. 1988. *Statistical Power Analysis for the Behavioral Sciences*. Hillsdale, N.J.: L. Erlbaum Associates.

Cohen, J. 1992. A Power Primer. *Psychological Bulletin* 112:155-159.

Cohen, J. 1994. The Earth Is Round (p<.05). *American Psychologist* 49:997-1003.

Cole, S.R., R.W. Platt, E.F. Schisterman, H. Chu, D. Westreich, D. Richardson und C. Poole. 2009. Illustrating Bias Due to Conditioning on a Collider. *International Journal of Epidemiology* 39:417-420.

Coleman, J.S. 1990. *Foundations of Social Theory*. Cambridge, Mass.: Belknap Press of Harvard University Press.

Correl, S.J. und S. Benard. 2006. Biased Estimators? Comparing Status and Statistical Theories of Gender Discrimination. *Advances in Group Processes* 23:89-116.

Coursol, A. und E.E. Wagner. 1986. Effect of Positive Findings on Submission and Acceptance: A Note of Meta-Analysis Bias. *Professional Psychology: Research and Practice* 17:136-137.

Dawes, R.M. 1980. Social Dilemmas. *Annual Review of Psychology* 31:169-193.

Dawes, R.M. und D.M. Messick. 2000. Social Dilemmas. *International Journal of Psychology* 35:111-116.

De Paola, M. und V. Scoppa. 2011. Gender Discrimination and Evaluators' Gender: Evidence From The Italien Academy. *IZA Discussion Paper Series* Nr. 9658.

De Paola, M. und V. Scoppa. 2015. Gender Discrimination and Evaluators' Gender: Evidence from Italian Academia. *Department of Economics and Statistics University of Calabria* 06-2011.

De Paola, M., M. Ponzo und V. Scoppa. 2018. Are Men Given Priority for Top Jobs? Investigating the Glass Ceiling in Italian Academia. *Journal of Human Capital* 12:475-503.

Deeks, J.J., P. Macaskill und L. Irwig. 2005. The Performance of Tests of Publication Bias and Other Sample Size Effects in Systematic Reviews of Diagnostic Test Accuracy Was Assessed. *Journal of Clinical Epidemiology* 58:882-893.

Deeks, J.J., J.P.T. Higgins und D.G. Altman. 2008. Analysing Data and Undertaking Meta-Analyses. In *Cochrane Handbook for Systematic Reviews of Interventions,* Hrsg. J.J. Deeks, J.P.T. Higgins und D.G. Altman, 243-296. Southern Gate, West Sussex: John Wiley & Sons, Ltd.

Descartes, R. 2006. *A Discourse on the Method of Correctly Conducting One's Reason and Seeking Truth in the Sciences*. Oxford, New York: Oxford University Press.
Deutsche Forschungsgemeinschaft. 2009. *Nachwuchswissenschaftlerinnen und Nachwuchswissenschaftler in DFG-geförderten Projekten. Rekrutierung, Erfahrungen und Perspektiven*. Bonn: Deutsche Forschungsgemeinschaft.
Deutscher Bundestag. 2013. *Bundesregierung hält an Grippemittel Tamiflu fest*. Berlin: Deutscher Bundestag.
Diamond, A.M. 1996. The Economics of Science. *Knowledge and Policy* 9:6-49.
Dickersin, K. 1990. The Existence of Publication Bias and Risk Factors for Its Occurrence. *JAMA* 263:1385-1389.
Dickersin, K. 2005. Publication Bias: Recognizing the Problem, Understanding Its Origins and Scope, and Preventing Harm. In *Publication Bias in Meta-Analysis: Prevention, Assessment and Adjustments,* Hrsg. H.R. Rothstein, A.J. Sutton und M. Borenstein, 11-33. Oxford: Blackwell Science.
Dickersin, K. und Y.-I. Min. 1993. Publication Bias – The Problem that Won't Go Away. *Annals of the New York Academy of Sciences* 703:135-148.
Dickersin, K., Y.-I. Min und C.L. Meinert. 1992. Factors Influencing Publication of Research Results. Follow-up of Applications Submitted to Two Institutional Review Boards. *JAMA* 267:374-378.
Diekmann, A. 2005. Betrug und Täuschung in der Wissenschaft. Datenfälschung, Diagnoseverfahren, Konsequenzen. *Schweizerische Zeitschrift für Soziologie* 31:7-30.
Diekmann, A. 2011. Are Most Published Research Findings False? *Jahrbücher für Nationalökonomie und Statistik* 231:628-636.
Diekmann, A. 2013. *Spieltheorie – Einführung, Beispiele, Experimente*. Reinbek bei Hamburg: Rowohlt.
Diekmann, A., B. Heintz, R. Münch, I. Ostner und H. Tyrell. 2002. Editorial. *Zeitschrift für Soziologie* 31:1-3.
Dijk, D., O. Manor und L.B. Carey. 2014. Publication Metrics and Success on the Academic Job Market. *Current Biology* 24:516-517
Doucouliagos, H. und T.D. Stanley. 2009. Publication Selection Bias in Minimum-Wage Research? A Meta-Regression Analysis. *British Journal of Industrial Relations* 47:406-428.
Duch, J., X.H.T. Zeng, M. Sales-Pardo, F. Radicchi, S. Otis, T.K. Woodruff und L.A.N. Amaral. 2012. The Possible Role of Resource Requirements and Academic Career-Choice Risk on Gender Differences in Publication Rate and Impact. *PLOS ONE* 7:e51332.
Duval, S. und R. Tweedie. 2000. Trim and Fill: A Simple Funnel-Plot-Based Method of Testing and Adjusting for Publication Bias in Meta-Analysis. *Biometrics* 56:455-463.
Easterbrook, P.J., J.A. Berlin, R. Gopalan und D.R. Matthews. 1991. Publication Bias in Clinical Research. *Lancet* 337:867-872.
Egger, M. und G.D. Smith. 1998. Meta-Analysis – Bias in Location and Selection of Studies. *British Medical Journal* 316:61-66.
Egger, M., G.D. Smith, M. Schneider und C. Minder. 1997. Bias in Meta-Analysis Detected by a Simple, Graphical Test. *British Medical Journal* 315:629-634.
Elia, N., E. von Elm, A. Chatagner, D.M. Pöpping und M.R. Tramèr. 2016. How Do Authors of Systematic Reviews Deal with Research Malpractice and Misconduct in Original Studies? A Cross-Sectional Analysis of Systematic Reviews and Survey of Their Authors. *BMJ Open* 6:1-10.
Elster, J. 1989. *Nuts and Bolts for the Social Sciences*. Cambridge: Cambridge University Press.
Engels, A., T. Ruschenburg und S. Zuber. 2012. Chancengleichheit in der Spitzenforschung: Institutionelle Erneuerung der Forschung in der Exzellenzinitiative des Bundes und der Länder. In *Institutionelle Erneuerungsfähigkeit der Forschung,* Hrsg. T. Heinze und G. Krücken, 187-217. Wiesbaden: VS Verlag für Sozialwissenschaften.
Epskamp, S. und M.B. Nuijten. 2018. statcheck: Extract Statistics from Articles and Recompute p Values (Version 1.3.0). https://cran.r-project.org/web/packages/statcheck/index.html.
Epstein, W.M. 1990. Confirmational Response Bias Among Social-Work Journals. *Science Technology & Human Values* 15:9-38.
Epstein, W.M. 2004. Confirmational Response Bias and the Quality of the Editorial Processes among American Social Work Journals. *Research on Social Work Practice* 14:450-458.

Europäische Kommission. 2016. Guidelines on FAIR Data Management in Horizon 2020.
Fanelli, D. 2009. How Many Scientists Fabricate and Falsify Research? A Systematic Review and Meta-Analysis of Survey Data. *PLOS ONE* 4:e5738.
Fanelli, D. 2010. "Positive" Results Increase Down the Hierarchy of the Sciences. *PLOS ONE* 5:e10068.
Fanelli, D. 2012. Negative Results Are Disappearing From Most Disciplines and Countries. *Scientometrics* 90:891-904.
Fanelli, D. 2013. Positive Results Receive More Citations, But Only in Some Disciplines. *Scientometrics* 94:701-709.
Fang, H. und A. Moro. 2011. Theories of Statistical Discrimination and Affirmative Action: A Survey. In *Handbooks of Social Economics,* Hrsg. J. Benhabib, M.O. Jackson und A. Bisin. Amsterdam: North-Holland, Elsevier.
Färber, C. und U. Spangenberg. 2008. *Wie werden Professuren besetzt? Chancengleichheit in Berufungsverfahren*. Frankfurt/Main: Campus.
Feigenbaum, S. und D.M. Levy. 1993. The Market for (Ir)reproducible Econometrics. *Accountability in Research* 3:25-43.
Feigenbaum, S. und D.M. Levy. 1996. The Technological Obsolescence of Scientific Fraud. *Rationality and Society* 8:261-276.
Ferguson, C.J. und M. Heene. 2012. A Vast Graveyard of Undead Theories: Publication Bias and Psychological Science's Aversion to the Null. *Perspectives on Psychological Science* 7:555-561.
Ferguson, C.J. und M.T. Brannick. 2012. Publication Bias in Psychological Science: Prevalence, Methods for Identifying and Controlling, and Implications for the Use of Meta-Analyses. *Psychological Methods* 17:120-128.
Fernandez-Mateo, I. und R.M. Fernandez. 2013. Women on Top: Executive Search and the Gender Gap in Allocation to Top Corporate Jobs. Annual Meeting American Sociological Associaton. New York.
Fernandez, Roberto M. und M.L. Sosa. 2005. Gendering the Job: Networks and Recruitment at a Call Center. *American Journal of Sociology* 111:859-904.
Fernandez, R.M. und M.L. Mors. 2008. Competing for Jobs: Labor Queues and Gender Sorting in the Hiring Process. *Social Science Research* 37:1061-1080.
Findeisen, I. 2011. *Hürdenlauf zur Exzellenz. Karrierestufen junger Wissenschaftlerinnen und Wissenschaftler*. Wiesbaden: VS Verlag für Sozialwissenschaften.
Findeisen, I., K. Auspurg und T. Hinz. 2010. *Konkurrenz oder Sichtbarkeit? Geschlechtsspezifische Förderchancen in der Deutschen Forschungsgemeinschaft*. Zürich: Rüegger.
Fisher, R.A. 1973. *Statistical Methods for Research Workers*. New York: Hafner.
Foschi, M. 1996. Double Standards in the Evaluation of Men and Women. *Social Psychology Quarterly* 59:237-354.
Foschi, M., L. Lai und K. Sigerson. 1994. Gender and Double Standards in the Assessment of Job Applicants. *Social Psychology Quarterly* 57:326-339.
Francis, G. 2012a. The Psychology of Replication and Replication in Psychology. *Perspectives on Psychological Science* 7:585-594.
Francis, G. 2012b. Too Good to Be True: Publication Bias in Two Prominent Studies from Experimental Psychology. *Psychonomic Bulletin & Review* 19:151-156.
Francis, G. 2012c. Evidence That Publication Bias Contaminated Studies Relating Social Class and Unethical Behavior. *Proceedings of the National Academy of Sciences of the United States of America* 109:e1587.
Francis, G. 2012d. Publication Bias and the Failure of Replication in Experimental Psychology. *Psychonomic Bulletin & Review* 19:975-991.
Francis, G. 2012e. The Same Old New Look: Publication Bias in a Study of Wishful Seeing. *i-Perception* 3:176-178.
Francis, G. 2013. Publication Bias in 'Red, Rank, and Romance in Women Viewing Men,' by Elliot et al. (2010). *Journal of Experimental Psychology: General* 142:292-296.
Franco, A., N. Malhotra und G. Simonovits. 2014. Publication Bias in the Social Sciences: Unlocking the File Drawer. *Science* 345:1-6.
Friedman, H. 1982. Simplified Determinations of Statistical Power, Magnitude of Effect and Research Sample Sizes. *Educational and Psychological Measurement* 42:521-526.

Gelman, A. 2013. Too Good to Be True. http://www.slate.com/articles/health_and_science/science/2013/07/statistics_and_psychology_multiple_comparisons_give_spurious_results.html (Zugriff: 6.1.2017).
Gemeinsame Wissenschaftskonferenz (GWK). 2018. *Chancengleichheit in Wissenschaft und Forschung 22. Fortschreibung des Datenmaterials (2016/2017) zu Frauen in Hochschulen und außerhochschulischen Forschungseinrichtungen*. Bonn GWK.
Gerber, A.S. und N. Malhotra. 2006. Can Political Science Literatures Be Believed? A Study of Publication Bias in the APSR and the AJPS. Vortrag im Rahmen des Annual Meeting of the Midwest Political Science Association. Chicago.
Gerber, A.S. und N. Malhotra. 2008a. Publication Bias in Empirical Sociological Research. *Sociological Methods & Research* 37:3-30.
Gerber, A.S. und N. Malhotra. 2008b. Do Statistical Reporting Standards Affect What Is Published? Publication Bias in Two Leading Political Science Journals. *Quarterly Journal of Political Science* 3:313-326.
Gerber, A.S., N. Malhotra, C.M. Dowling und D. Doherty. 2010. Publication Bias in Two Political Behavior Literatures. *American Politics Research* 38:591-613.
GESIS Leibniz-Institut für Sozialwissenschaften. 2017. Evaluation des Professorinnenprogramms des Bundes und der Länder: Zweite Programmphase und Gesamtevaluation.
Godlee, F. 2012. Research Misconduct is Widespread and Harms Patients. *BMJ* 344:e:14.
Gomolla, M. und F.-O. Radtke. 2009. *Institutionelle Diskriminierung*: VS Verlag für Sozialwissenschaften.
Greene, W.H. 2012. *Econometric Analysis*. Boston: Prentice Hall.
Greenland, S., J. Pearl und J.M. Robins. 1999. Causal Diagrams for Epidemiologic Research. *Epidemiology* 10:37-48.
Grigat, F. 2018. 45 Prozent neue Professuren an Frauen. https://www.forschung-und-lehre.de/politik/45-prozent-neue-professuren-an-frauen-1160/ (Zugriff: 10.01.2019).
Gross, C. und M. Jungbauer-Gans. 2007. Erfolg durch Leistung? Ein Forschungsüberblick zum Thema Wissenschaftskarrieren. *Soziale Welt* 58:453-471.
Groves, R.M. 2009. *Survey Methodology*. Hoboken, N.J.: Wiley.
Guetzkow, J., M. Lamont und G. Mallard. 2004. What is Originality in the Humanities and the Social Sciences? *American Sociological Review* 69:190-212.
Hamburger, H. 1973. N-Person Prisoner's Dilemma. *Journal of Mathematical Sociology* 3:27-48.
Hardin, G. 1968. The Tragedy of the Commons. *Science* 162:1243-1248.
Hart, R.A. und D.H. Clark. 1999. Does Size Matter? Exploring the Small Sample Properties of Maximum Likelihood Estimation. Annual Meeting of the Midwest Political Science Association. Chicago.
Hartgerink, C.H.J., R.C.M. van Aert, M.B. Nuijten, J.M. Wicherts und M.A.L.M. van Assen. 2016. Distributions of p-Values Smaller Than .05 in Psychology: What Is Going On? *PeerJ* 4:e1935.
Hartgerink, H.C. 2016. 688,112 Statistical Results: Content Mining Psychology Articles for Statistical Test Results. *Data* 1.
Hayashino, Y., Y. Noguchi und T. Fukui. 2005. Systematic Evaluation and Comparison of Statistical Tests for Publication Bias. *Journal of Epidemiology* 15:235-243.
Head, M.L., L. Holman, R. Lanfear, A.T. Kahn und M.D. Jennions. 2015. The Extent and Consequences of p-Hacking in Science. *PLOS Biol* 13:e1002106.
Hechtman, L.A., N.P. Moore, C.E. Schulkey, A.C. Miklos, A.M. Calcagno, R. Aragon und J.H. Greenberg. 2018. NIH Funding Longevity by Gender. *Proceedings of the National Academy of Sciences* 115:7943-7948.
Hedström, P. und P. Bearman. 2009. What is Analytical Sociology All About? An Introductory Essay. In *Analytical Sociology,* Hrsg. P. Hedström und P. Bearman, 3-24. Oxford, New York: De Gruyter.
Hessels, L.K. und H. van Lente. 2008. Re-thinking New Knowledge Production: A Literature Review and a Research Agenda. *Research Policy* 37:740-760.
Higgins, J.P.T. und S.G. Thompson. 2002. Quantifying Heterogeneity in a Meta-Analysis. *Statistics in Medicine* 21:1539-1558.
Higgins, J.P.T. und S. Green. 2008. *Cochrane Handbook for Systematic Reviews of Interventions*. Chichester: Wiley.

Hinz, T., I. Findeisen und K. Auspurg. 2008. *Wissenschaftlerinnen in der DFG. Förderprogramme, Förderchancen und Funktionen (1991 - 2004)*. Weinheim: Wiley-VCH.
Hirschauer, S. 2016. Der Diskriminierungsdiskurs und das Kavaliersmodell universitärer Frauenförderung. *Soziale Welt* 67:119-136.
Hirschi, T. 1969. *Causes of Delinquency*. Berkeley: University of California Press.
Hoenig, J.M. und D.M. Heisey. 2001. The Abuse of Power. *The American Statistician* 55:19-24.
Holm, S. 1979. A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics* 6:65-70.
Hubbard, R. 2015. *Corrupt Research: The Case for Reconceptualizing Empirical Management and Social Science*: Sage Publications.
Humphreys, M., R.S. de la Sierra und P. van der Windt. 2013. Fishing, Commitment, and Communication: A Proposal for Comprehensive Nonbinding Research Registration. *Political Analysis* 21:1-20.
Iafrate, F. 2015. *From Big Data to Smart Data*. Hoboken: John Wiley & Sons, Inc.
Ioannidis, J.P. 1998. Effect of the Statistical Significance of Results on the Time to Completion and Publication of Randomized Efficacy Trials. *JAMA* 279:281-286.
Ioannidis, J.P. 2005. Why Most Published Research Findings Are False. *PLOS Med* 2:e124.
Ioannidis, J.P.A. 2013. Clarifications on the Application and Interpretation of the Test for Excess Significance and Its Extensions. *Journal of Mathematical Psychology* 57:184-187.
Ioannidis, J.P.A. 2014. Discussion: Why "An Estimate of the Science-Wise False Discovery Rate and Application to the Top Medical Literature" Is False. *Biostatistics* 15:28-36.
Ioannidis, J.P.A. und T.A. Trikalinos. 2007. An Exploratory Test for an Excess of Significant Findings. *Clinical Trials* 4:245-253.
Ioannidis, J.P.A., T.D. Stanley und H. Doucouliagos. 2017. The Power of Bias in Economics Research. *The Economic Journal* 127:F236-F265.
Jager, L.R. und J.T. Leek. 2014. An Estimate of the Science-Wise False Discovery Rate and Application to the Top Medical Literature. *Biostatistics* 15:1-12.
Jefferson, T., M.A. Jones, P. Doshi, C.B. Del Mar, C.J. Heneghan, R. Hama und M.J. Thompson. 2012. Neuraminidase Inhibitors for Preventing and Treating Influenza in Healthy Adults. *Cochrane Database of Systematic Reviews* 2012:1-246.
John, L.K., G. Loewenstein und D. Prelec. 2012. Measuring the Prevalence of Questionable Research Practices With Incentives for Truth Telling. *Psychological Science* 23:524-532.
Johnson, V. und Y. Yuan. 2007. Comments on `An Exploratory Test for an Excess of Significant Findings' by JPA loannidis and TA Trikalinos. *Clinical Trials* 4:254-255.
Jungbauer-Gans, M. und C. Gross. 2013. Determinants of Success in University Careers. *Zeitschrift für Soziologie* 42:74-92.
Keele, L. und D.K. Park. 2005. Difficult Choices: An Evaluation of Heterogeneous Choice Models. Workingpaper. Chicago.
Kerr, N.L. 1998. HARKing: Hypothesizing After the Results are Known. *Personality and Social Psychology Review* 2:196-217.
Kicinski, M. 2014. How Does Under-Reporting of Negative and Inconclusive Results Affect the False-Positive Rate in Meta-Analysis? A Simulation Study. *BMJ Open* 4:1-8.
Knobloch-Westerwick, S., C.J. Glynn und M. Huge. 2013. The Matilda Effect in Science Communication: An Experiment on Gender Bias in Publication Quality Perceptions and Collaboration Interest. *Science Communication* 35:603-625.
Kohaut, S. und I. Möller. 2010. Frauen in Chefetagen. *Wirtschaftsdienst* 90:420-422.
Kollock, P. 1998. Social Dilemmas: The Anatomy of Cooperation. *Annual Review of Sociology* 24:183-214.
Konsortium Bundesbericht Wissenschaftlicher Nachwuchs. 2013. *Bundesbericht Wissenschaftlicher Nachwuchs 2013. Statistische Daten und Forschungsbefunde zu Promovierenden und Promovierten in Deutschland*. Bielefeld: Bertelsmann.
Kreckel, R. 2008. *Zwischen Promotion und Professur. Das wissenschaftliche Personal in Deutschland im Vergleich mit Frankreich, Großbritannien, USA, Schweden, den Niederlanden, Österreich und der Schweiz*. Leipzig: Akademische Verlagsanstalt.
Kreuter, F., S. Presser und R. Tourangeau. 2008. Social Desirability Bias in CATI, IVR, and Web Surveys: The Effects of Mode and Question Sensitivity. *Public Opinion Quarterly* 72:847-865.

Kühberger, A., A. Fritz und T. Scherndl. 2014. Publication Bias in Psychology: A Diagnosis Based on the Correlation between Effect Size and Sample Size. *PLOS ONE* 9:e105825.
Labovitz, S. 1968. Criteria for Selecting a Significance Level: A Note on the Sacredness of .05. *The American Sociologist* 3:220-222.
Labovitz, S. 1972. Statistical Usage In Sociology: Sacred Cows and Ritual. *Sociological Methods & Research* 1:13-37.
Lakens, D. 2015. What p-hacking Really Looks Like: A Comment on Masicampo and LaLande (2012). *The Quarterly Journal of Experimental Psychology* 68:829-832.
Landeshochschulgesetz Baden-Württemberg. *In der Fassung vom 1. Jannuar 2005 (GBl. 2005, 1), geändert durch Artikel 1 des Gesetzes vom 13. März 2018 (GBl. S. 85)*.
Laney, D. 2001. 3D Data Management: Controlling Data Volume, Velocity, and Variety. META Group. *http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf (Zugriff 25.11.2018)*.
Lau, J., J.P.A. Ioannidis, N. Terrin, C.H. Schmid und I. Olkin. 2006. The Case of the Misleading Funnel Plot. *BMJ* 333:597-600.
Leahey, E. 2007. Not by Productivity Alone: How Visibility and Specialization Contribute to Academic Earnings. *American Sociological Review* 72:533-561.
Leahey, E., J.L. Crockett und L.A. Hunter. 2008. Gendered Academic Careers: Specializing for Success? *Social Forces* 86:1273-1309.
Leahey, E., B. Keith und J. Crockett. 2010. Specialization and Promotion in an Academic Discipline. *Research in Social Stratification and Mobility* 28:135-155.
Lee, C.J. 2016. Revisiting Current Causes of Women's Underrepresentation in Science. In *Implicit Bias and Philosophy,* Hrsg. J. Saul und M. Brownstein, Volume 1: Metaphysics and Epistemology. Oxford: Oxford University Press.
Lee, C.J., C.R. Sugimoto, G. Zhang und B. Cronin. 2013. Bias in Peer Review. *Journal of the American Society for Information Science and Technology* 64:2-17.
Lee, D.S. und T. Lemieux. 2010. Regression Discontinuity Designs in Economics. *Journal of Economic Literature* 48:281-355.
Leek, J.T. und R.D. Peng. 2015. What Is the Question? *Science* 347:1314-1315.
Leemann, R.J. 2005. Geschlechterungleichheiten in wissenschaftlichen Laufbahnen. In *Institutionalisierte Ungleichheiten. Wie das Bildungswesen Chancen blockiert,* Hrsg. P.A. Berger und H. Kahlert, 179-214. Weinheim: Beltz Juventa.
Leggett, N.C., N.A. Thomas, T. Loetscher und M.E.R. Nicholls. 2013. The Life of p: "Just Significant" Results Are on the Rise. *Quarterly Journal of Experimental Psychology* 66:2303-2309.
Leone, F.C., L.S. Nelson und R.B. Nottingham. 1961. The Folded Normal Distribution. *Technometrics* 3:543-550.
Levelt Commitee, Noort Commitee und Drenth Commitee. 2012. Flawed Science: The Fraudulent Research Practices of Social Psychologist Diederik Stapel.
Liebeskind, U. 2004. Arbeitsmarktsegregation und Einkommen. *Kölner Zeitschrift für Soziologie und Sozialpsychologie* 56:630-652.
Light, R.J. und D.B. Pillemer. 1984. *Summing Up: The Science of Reviewing Research*. Cambridge, Mass.: Harvard University Press.
Lincoln, A.E., S. Pincus, J.B. Koster und P.S. Leboy. 2012. The Matilda Effect in science: Awards and prizes in the US, 1990s and 2000s. *Social Studies of Science* 42:307-320.
Lind, I. 2004. *Aufstieg oder Ausstieg? Karrierewege von Wissenschaftlerinnen*. Bielefeld: Kleine.
Long, J.S. 1992. Measures of Sex Differences in Scientific Productivity. *Social Forces* 71:159-178.
Long, J.S., P.D. Allison und R. McGinnis. 1993. Rank Advancement in Academic Careers: Sex Differences and the Effects of Productivity. *American Sociological Review* 58:703-722.
Löther, A. 2015. Hochschulranking nach Gleichstellungsaspekten 2015. cews.publik.no 19. GESIS Leibniz-Institut für Sozialwissenschaften.
Lutter, M. und M. Schröder. 2016. Who Becomes a Tenured Professor, and Why? Panel Data Evidence from German Sociology, 1980–2013. *Research Policy* 45:999-1013.
Macaskill, P., S.D. Walter und L. Irwig. 2001. A Comparison of Methods to Detect Publication Bias in Meta-Analysis. *Statistics in Medicine* 20:641-654.
MacCoun, R. und S. Perlmutter. 2015. Blind Analysis: Hide Results to Seek the Truth. *Nature* 526:187-189.

Macfarlane, B. und M. Cheng. 2008. Communism, Universalism and Disinterestedness: Re-examining Contemporary Support among Academics for Merton's Scientific Norms. *Journal of Academic Ethics* 6:67-78.
Mahoney, M.J. 1977. Publication Prejudices: An Experimental Study of Confirmatory Bias in the Peer Review System. *Cognitive Therapy and Research* 1:161-175.
Mallett, S. und M. Clarke. 2002. The Typical Cochrane Review. How Many Trials? How Many Participants? *International Journal of Technology Assessment in Health Care* 18:820-823.
Markowitz, D.M. und J.T. Hancock. 2014. Linguistic Traces of a Scientific Fraud: The Case of Diederik Stapel. *PLOS ONE* 9:e105937.
Marsh, H.W., L. Bornmann, R. Mutz, H.-D. Daniel und A. O'Mara. 2009. Gender Effects in the Peer Reviews of Grant Proposals: A Comprehensive Meta-Analysis Comparing Traditional and Multilevel Approaches. *Review of Educational Research* 79:1290-1326.
Masicampo, E.J. und D.R. Lalande. 2012. A Peculiar Prevalence of p Values just Below .05. *Quarterly Journal of Experimental Psychology* 65:2271-2279.
McCook, A. 2013. Barred From the Boardroom. *Nature* 495:25-27.
McShane, B.B., U. Bockenholt und K.T. Hansen. 2016. Adjusting for Publication Bias in Meta-Analysis: An Evaluation of Selection Methods and Some Cautionary Notes. *Perspectives on Psychological Science* 11:730-749.
Merton, R.K. 1957. Priorities in Scientific Discovery: A Chapter in the Sociology of Science. *American Sociological Review* 22:635-659.
Merton, R.K. 1961. Singletons and Multiples in Scientific Discovery: A Chapter in the Sociology of Science. *Proceedings of the American Philosophical Society* 105:470-486.
Merton, R.K. 1970. Behavior Patterns of Scientists. *Leonardo* 3:213-220.
Merton, R.K. 1973. The Normative Structure of Science. In *The Sociology of Science: Theoretical and Empirical Investigations,* Hrsg. N.W. Storer, 267-278. Chicago: University of Chicago Press.
Moher, D., A. Liberati, J. Tetzlaff, D.G. Altman und The Prisma Group. 2009. Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. *PLOS Med* 6:e1000097.
Mood, C. 2010. Logistic Regression: Why We Cannot Do What We Think We Can Do, and What We Can Do About It. *European Sociological Review* 26:67-82.
Mooney, C.Z. 1997. *Monte Carlo Simulation*. Thousand Oakes: Sage Publications.
Moreno, S.G., A.J. Sutton, A.E. Ades, T.D. Stanley, K.R. Abrams, J.L. Peters und N.J. Cooper. 2009. Assessment of Regression-based Methods to Adjust for Publication Bias through a Comprehensive Simulation Study. *BMC Medical Research Methodology* 2009:1-17.
Morey, R.D. 2013. The Consistency Test Does Not – and Cannot – Deliver What Is Advertised: a Comment on Francis (2013). *Journal of Mathematical Psychology* 57:180-183.
Munafò, M.R., B.A. Nosek, D.V.M. Bishop, K.S. Button, C.D. Chambers, N. Percie du Sert, U. Simonsohn, E.-J. Wagenmakers, J.J. Ware und J.P.A. Ioannidis. 2017. A Manifesto for Reproducible Science. *Nature Human Behaviour* 1:0021.
Nash, J.F. 1950. Equilibrium Points in N-Person Games. *Proceedings of the National Academy of Sciences* 36:48-49.
Nature. 2017. Getting Published in Nature: The Editorial Process. http://www.nature.com/nature/authors/get_published/ (Zugriff: 05.01.2017).
Necker, S. 2012. Wissenschaftliches Fehlverhalten – ein Problem in der deutschen Volkswirtschaftslehre? *Perspektiven der Wirtschaftspolitik* 13:267-285.
Nosek, B.A. und D.S.J.A.O. Lindsay. 2018. Preregistration Becoming the Norm in Psychological Science. *APS Observer* 31.
Nuijten, M.B., C.H.J. Hartgerink, M.A.L.M. van Assen, S. Epskamp und J.M. Wicherts. 2016. The Prevalence of Statistical Reporting Errors in Psychology (1985–2013). *Behavior Research Methods* 48:1205-1226.
Nuzzo, R. 2014. Scientific Method: Statistical Errors. *Nature* 506:150-152.
Olson, C.M., D. Rennie, D. Cook, K. Dickersin, A. Flanagin, J.W. Hogan, Q. Zhu, J. Reiling und B. Pace. 2002. Publication Bias in Editorial Decision Making. *JAMA* 287:2825-2828.
Open Science Collaboration. 2015. Estimating the Reproducibility of Psychological Science. *Science* 349.
Open Science Collaboration. 2018. Data Reproducibility Policies. https://osf.io/kgnva/wiki/home/ (Zugriff: 03.11.2018).

Paldam, M. 2013. Regression Costs Fall, Mining Ratios Rise, Publication Bias Looms, and Techniques Get Fancier: Reflections on Some Trends in Empirical Macroeconomics. *Econ Journal Watch* 10:136-156.
Paldam, M. 2015. Simulating an Empirical Paper by the Rational Economist. *Empirical Economics* 50:1-25.
Papke, L.E. und J.M. Wooldridge. 1996. Econometric Methods for Fractional Response Variables with an Application to 401(k) Plan Participation Rates. *Journal of Applied Econometrics* 11:619-632.
Patel, C., A. Patel und D. Patel. 2012. Optical Character Recognition by Open Source OCR Tool Tesseract: A Case Study. *International Journal of Computer Applications* 55:50-56.
Petticrew, M. 1998. Diagoras of Melos (500 BC): an Early Analyst of Publication Bias. *The Lancet* 352:1558.
Petticrew, M., M. Egan, H. Thomson, V. Hamilton, R. Kunkler und H. Roberts. 2008. Publication Bias in Qualitative Research: What Becomes of Qualitative Research Presented at Conferences? *Journal of Epidemiology and Community Health* 62:552-554.
Phelps, E.S. 1972. The Statistical Theory of Racism and Sexism. *The American Economic Review* 62:659-661.
Plümper, T. und F. Schimmelfennig. 2007. Wer wird Prof — und wann? Berufungsdeterminanten in der deutschen Politikwissenschaft. *Politische Vierteljahresschrift* 48:97-117.
Podlubny, I. 2005. Comparison of Scientific Impact Expressed by the Number of Citations in Different Fields of Science. *Scientometrics* 64:95-99.
Popper, K.R. 1968. *Conjectures and Refutations: The Growth of Scientific Knowledge*. New York: Harper & Row.
Raub, W. und T. Voss. 2016. Micro-Macro Models in Sociology: Antecedents of Coleman's Diagram. In *Social Dilemmas, Institutions, and the Evolution of Cooperation,* Hrsg. B. Jann und W. Przepiorka, 11-36. Berlin, Boston: De Gruyter.
Reed, W.R. 2015. A Monte Carlo Analysis of Alternative Meta-Analysis Estimators in the Presence of Publication Bias. *Economics-the Open Access Open-Assessment E-Journal* 9:1-40.
Reich, E.S. 2009. *Plastic Fantastic – How the Biggest Fraud in Physics Shook the Scientific World*. New York: Palgrave Macmillan.
Renkewitz, F. und M. Keiner. 2016. How to Detect Publication Biases from Published Data? A Monte Carlo Simulation of Different Methods. 50. Kongress der Deutschen Gesellschaft für Psychologie. Leipzig.
Reskin, B.F. 1991. Labor Markets as Queues: a Structural Approach to Changing Occupational Sex Composition. In *Micro-Macro Linkages in Sociology,* Hrsg. J. Huber, 170–192. Newbury Park, CA: Sage Publications.
Ridgeway, C.L. 1991. The Social Construction of Status Value – Gender and Other Nominal Characteristics. *Social Forces* 70:367-386.
Ridgeway, C.L. 2014. Why Status Matters for Inequality. *American Sociological Review* 79:1-16.
Ridley, J., N. Kolm, R.P. Freckelton und M.J.G. Gage. 2007. An Unexpected Influence of Widely Used Significance Thresholds on the Distribution of Reported p-Values. *Journal of Evolutionary Biology* 20:1082-1089.
Rogers, W. 1994. Regression Standard Errors in Clustered Samples. *Stata Technical Bulletin* 3:19-23.
Rosenthal, R. 1979. The File Drawer Problem and Tolerance for Null Results. *Psychological Bulletin* 86:638-641.
Rothstein, H., A.J. Sutton und M. Borenstein. 2005. *Publication Bias in Meta-Analysis: Prevention, Assessment and Adjustments*. Chichester: John Wiley.
Sackett, P.R., C.L.Z. DuBois und A.W. Noe. 1991. Tokenism in Performance Evaluation: The Effects of Work Group Representation on Male-Female and White-Black Differences in Performance Ratings. *Journal of Applied Psychology* 76:263-267.
Sahner, H. 1979. Veröffentlichte empirische Sozialforschung – Eine Kumulation von Artefakten eine Analyse von Periodika. *Zeitschrift für Soziologie* 8:267-278.
Salganik, M.J. 2018. *Bit by Bit: Social Research in the Digital Age*. Princeton: Princeton University Press.
Sandström, U. und M. Hällsten. 2008. Persistent Nepotism in Peer-Review. *Scientometrics* 74:175-189.

Sanz-Menéndez, L., L. Cruz-Castro und K. Alva. 2013. Time to Tenure in Spanish Universities: an Event History Analysis. *PLOS ONE* 8:e77028.
Schelling, T.C. 1973. Hockey Helmets, Concealed Weapons, and Daylight Saving: A Study of Binary Choices with Externalities. *The Journal of Conflict Resolution* 17:381-428.
Schimmack, U. 2012. The Ironic Effect of Significant Results on the Credibility of Multiple-Study Articles. *Psychological Methods* 17:551-566.
Schneck, A. 2017. Examining Publication Bias – A Simulation-Based Evaluation of Statistical Tests on Publication Bias. *PeerJ* 5:e4115.
Schöch, C. 2013. Big? Smart? Clean? Messy? Data in the Humanities. *Journal of Digital Humanities* 2:2-13.
Schulze, G., S. Warning und C. Wiermann. 2008. What and How Long Does it Take to Get Tenure? The Case of Economics and Business Administration in Austria, Germany and Switzerland. *German Economic Review* 9:473-505.
Schwarzer, G., G. Antes und M. Schumacher. 2002. Inflation of Type I Error Rate in Two Statistical Tests for the Detection of Publication Bias in Meta-Analyses with Binary Outcomes. *Statistics in Medicine* 21:2465-2477.
Science. 2017. The Science Contributors FAQ. http://www.sciencemag.org/site/feature/contribinfo/faq/#pct_faq (Zugriff: 05.01.2017).
Sharpe, D. 1997. Of Apples and Oranges, File Drawers and Garbage: Why Validity Issues in Meta-Analysis Will Not Go Away. *Clinical Psychology Review* 17:881-901.
Simmons, J.P. und U. Simonsohn. 2017. Power Posing: P-Curving the Evidence. *Psychological Science* 28:687-693.
Simmons, J.P., L.D. Nelson und U. Simonsohn. 2011. False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. *Psychological Science* 22:1359-1366.
Simon, H.A. 1983. *Reason in Human Affairs*. Stanford, CA: Stanford University Press.
Simonsohn, U. 2013. It Really Just Does Not Follow, Comments on Francis (2013). *Journal of Mathematical Psychology* 57:174-176.
Simonsohn, U., L.D. Nelson und J.P. Simmons. 2014a. P-Curve and Effect Size: Correcting for Publication Bias Using Only Significant Results. *Perspectives on Psychological Science* 9:666-681.
Simonsohn, U., L.D. Nelson und J.P. Simmons. 2014b. P-Curve: A Key to the File Drawer. *Journal of Experimental Psychology-General* 143:534-547.
Simonsohn, U., J.P. Simmons und L.D. Nelson. 2015. Better P-Curves: Making P-Curve Analysis More Robust To Errors, Fraud, and Ambitious P-Hacking, A Reply To Ulrich and Miller (2015). *Journal of Experimental Psychology-General* 144:1146-1152.
Skeels, J.W. und R.P. Fairbanks. 1968. Publish or Perish: An Analysis of the Mobility of Publishing and Nonpublishing Economists. *Southern Economic Journal* 35:17-25.
Skipper, J.K., Jr., A.L. Guenther und G. Nass. 1967. The Sacredness of .05: A Note Concerning the Uses of Statistical Levels of Significance in Social Science. *The American Sociologist* 2:16-18.
Slote, M. 1985. Utilitarianism, Moral Dilemmas, and Moral Cost. *American Philosophical Quarterly* 22:161-168.
Smaldino, P.E. und R. McElreath. 2016. The Natural Selection of Bad Science. *Royal Society Open Science* 3.
Spence, M. 1973. Job Market Signaling. *The Quarterly Journal of Economics* 87:355-374.
Stanley, T.D. 2017. Limitations of PET-PEESE and Other Meta-Analysis Methods. *Social Psychological and Personality Science* 8:581–591.
Stanley, T.D. und H. Doucouliagos. 2014. Meta-Regression Approximations to Reduce Publication Selection Bias. *Research Synthesis Methods* 5:60-78.
Steele, C.M. 1997. A Threat in the Air: How Stereotypes Shape Intellectual Identity and Performance. *American Psychologist* 52:613-629.
Steele, C.M. und J. Aronson. 1995. Stereotype Threat and the Intellectual Test Performance of African Americans. *Journal of Personality and Social Psychology* 69:797-811.
Steinpreis, R., K.A. Anders und D. Ritzke. 1999. The Impact of Gender on the Review of the Curricula Vitae of Job Applicants and Tenure Candidates: A National Empirical Study. *Sex Roles* 41:509-528.

Stephan, P.E. 2010. The Economics of Science. In *Handbook of the Economics of Innovation* Hrsg. B.H. Hall und N. Rosenberg, 217-274. Amsterdam: North Holland.
Sterling, T.D. 1959. Publication Decisions and Their Possible Effects on Inferences Drawn from Tests of Significance – Or Vice Versa. *Journal of the American Statistical Association* 54:30-34.
Sterling, T.D., W.L. Rosenbaum und J.J. Weinkam. 1995. Publication Decisions Revisited: The Effect of the Outcome of Statistical Tests on the Decision to Publish and Vice Versa. *The American Statistician* 49:108-112.
Stern, J.M. und R.J. Simes. 1997. Publication Bias: Evidence of Delayed Publication in a Cohort Study of Clinical Research Projects. *British Medical Journal* 315:640-645.
Sterne, J.A.C., D. Gavaghan und M. Egger. 2000. Publication and Related Bias in Meta-Analysis: Power of Statistical Tests and Prevalence in the Literature. *Journal of Clinical Epidemiology* 53:1119-1129.
Sterne, J.A.C., A.J. Sutton, J.P.A. Ioannidis, N. Terrin, D.R. Jones, J. Lau, J. Carpenter, G. Rücker, R.M. Harbord, C.H. Schmid, J. Tetzlaff, J.J. Deeks, J. Peters, P. Macaskill, G. Schwarzer, S. Duval, D.G. Altman, D. Moher und J.P.T. Higgins. 2011. Recommendations for Examining and Interpreting Funnel Plot Asymmetry in Meta-Analyses of Randomised Controlled Trials. *BMJ* 343:d4002.
Stroebe, W., T. Postmes und R. Spears. 2012. Scientific Misconduct and the Myth of Self-Correction in Science. *Perspectives on Psychological Science* 7:670-688.
Sutton, A.J. und T.D. Pigott. 2006. Bias in Meta-Analysis Induced by Incompletely Reported Studies. In *Publication Bias in Meta-Analysis,* Hrsg. H. Rothstein, A.J. Sutton und M. Borenstein, 221-239. Chichester: John Wiley & Sons.
Tang, J.L. und J.L.Y. Liu. 2000. Misleading Funnel Plot for Detection of Bias in Meta-Analysis. *Journal of Clinical Epidemiology* 53:477-484.
Terrin, N., C.H. Schmid, J. Lau und I. Olkin. 2003. Adjusting for Publication Bias in the Presence of Heterogeneity. *Statistics in Medicine* 22:2113-2126.
Thaler, R. 1980. Toward a Positive Theory of Consumer Choice. *Journal of Economic Behavior & Organization* 1:39-60.
Tibshirani, R. 1996. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society* 58:267-288.
Titus, S.L., J.A. Wells und L.J. Rhoades. 2008. Repairing Research Integrity. *Nature* 453:980-982.
Trinquart, L., A. Abbe und P. Ravaud. 2012. Impact of Reporting Bias in Network Meta-Analysis of Antidepressant Placebo-Controlled Trials. *PLOS ONE* 7:e35219.
Tsagris, M., C. Beneki und H. Hassani. 2014. On the Folded Normal Distribution. *Mathematics* 2:12-28.
Ulrich, R. und J. Miller. 2017. Some Properties of p-Curves, with an Application to Gradual Publication Bias. *Psychological Methods* Advance Online Publication:1-15.
Universität Konstanz. 2013. Entzug des Doktorgrades in letzter Instanz bestätigt. Presseinformation Nr. 98 vom 31. Juli 2013 (Zugriff: 01.03.2014).
van Aert, R.C.M., J.M. Wicherts und M.A.L.M. van Assen. 2016. Conducting Meta-Analyses Based on p Values. *Perspectives on Psychological Science* 11:713-729.
van Assen, M.A.L.M., R.C.M. van Aert und J.M. Wicherts. 2015. Meta-Analysis Using Effect Size Distributions of only Statistically Significant Studies. *Psychological Methods* 20:293-309.
van den Brink, M., M. Brouns und S. Waslander. 2006. Does Excellence Have a Gender? *Employee Relations* 28:523-539.
Vidgen, B. und T. Yasseri. 2016. P-Values: Misunderstood and Misused. *Frontiers in Physics* 4.
Wagner, M., I. Dunkake und B. Weiß. 2004. Schulverweigerung. Empirische Analysen zum abweichenden Verhalten von Schülern. *Kölner Zeitschrift für Soziologie und Sozialpsychologie* 56:793-793.
Wasserstein, R.L. und N.A. Lazar. 2016. The ASA's Statement on p-Values: Context, Process, and Purpose. *American Statistician* 70:129-131.
Weiß, B. und M. Wagner. 2008. Potentiale und Probleme von Meta-Analysen in der Soziologie. *Sozialer Fortschritt* 10/11:250-255.
Weiß, B. und C. Berning. 2013. Publication Bias in the German Social Sciences: An Application of the Caliper Test for Three High-Ranking German Social Science Journals. Poster präsentiert am Campell Colloquium. Chicago.

Weisshaar, K. 2017. Publish and Perish? An Assessment of Gender Gaps in Promotion to Tenure in Academia. *Social Forces* 96:529-560.

Wilkinson, M.D., M. Dumontier, I.J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L.B. da Silva Santos, P.E. Bourne, J. Bouwman, A.J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C.T. Evelo, R. Finkers, A. Gonzalez-Beltran, A.J.G. Gray, P. Groth, C. Goble, J.S. Grethe, J. Heringa, P.A.C. 't Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S.J. Lusher, M.E. Martone, A. Mons, A.L. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. van Schaik, S.-A. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M.A. Swertz, M. Thompson, J. van der Lei, E. van Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao und B. Mons. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* 3:160018.

Williams, R. 2009. Using Heterogeneous Choice Models to Compare Logit and Probit Coefficients Across Groups. *Sociological Methods & Research* 37:531-559.

Williams, R. 2015. fracglm. http://www3.nd.edu/~rwilliam/stata (Zugriff: 19.10.2015).

Wilson, F.D., G.L. Smoke und J.D. Martin. 1973. The Replication Problem in Sociology: A Report and a Suggestion. *Sociological Inquiry* 43:141-149.

Wold, A. und C. Wennerås. 1997. Nepotism and Sexism in Peer-Review. *Nature* 387:341-343.

Wooldridge, J.M. 2013. *Introductory Econometrics: A Modern Approach*. Mason, OH: South-Western Cengage Learning.

Wroblewski, A. und A. Leitner. 2013. Analyse von Gender-Indikatoren WB-Kennzahl 1.A.5 Gender Pay Gap und Datenbedarfskennzahl 1.3 Geschlechterrepräsentanz im Berufungsverfahren. Wien. Institut für Höhere Studien (IHS).

Yoder, S. und B.H. Bramlett. 2011. What Happens at the Journal Office Stays at the Journal Office: Assessing Journal Transparency and Record-Keeping Practices. *Ps-Political Science & Politics* 44:363-373.

Zeitschrift für Soziologie. 2012. Editorial. *Zeitschrift für Soziologie* 41:2-4.

Zou, H. und T. Hastie. 2005. Regularization and Variable Selection via the Elastic Net. *Journal of the Royal Statistical Society: Series B* 67:301-320.