Köln Z Soziol (2014) 66:549–573
DOI 10.1007/s11577-014-0284-3

ABHANDLUNGEN

Kölner Zeitschrift für Soziologie und Sozialpsychologie

Ausmaß und Risikofaktoren des Publication Bias in der deutschen Soziologie

Katrin Auspurg · Thomas Hinz · Andreas Schneck

Zusammenfassung Die statistische Signifikanz von Forschungsergebnissen wird oft fälschlicherweise als ein Indikator für deren Relevanz und Aussagekraft gehalten. Signifikante Ergebnisse werden eher veröffentlicht, obwohl nicht-signifikante Ergebnisse gleichermaßen für den Erkenntnisfortschritt bedeutsam sind. Die Folgen sind eine Überschätzung von Effektstärken und eine zu optimistische Beurteilung von Theorien. Im vorliegenden Beitrag wird dem Problem des Publication Bias (PB) in der deutschen Soziologie anhand von elf Jahrgängen der zwei wichtigsten deutschsprachigen Soziologie-Zeitschriften (*Kölner Zeitschrift für Soziologie und Sozialpsychologie, Zeitschrift für Soziologie*) mithilfe des Caliper-Tests nachgegangen. Lassen sich ebenso wie in US-amerikanischen Soziologie-Zeitschriften Hinweise auf einen PB finden, und wenn ja, unter welchen Bedingungen ist dieser besonders stark ausgeprägt? Im Mittelpunkt der Ursachenanalyse stehen Möglichkeiten der Datenmanipulation sowie der sozialen Kontrolle durch Forschende. Im Ergebnis finden sich auch für die deutsche Soziologie Hinweise auf einen PB, wenngleich in schwächerem Umfang als in US-amerikanischen Zeitschriften. Einfache Maßnahmen wie Herausgebervorgaben, wonach Daten für Replikationen zur Verfügung zu stellen sind, zeigen keine durchschlagende Wirkung. Es lässt sich

K. Auspurg (✉) · A. Schneck
Institut für Soziologie, Goethe-Universität Frankfurt a. M.,
Grüneburgplatz 1, 60323 Frankfurt a. M., Deutschland
E-Mail: auspurg@soz.uni-frankfurt.de

A. Schneck
E-Mail: schneck@soz.uni-frankfurt.de

T. Hinz
Fachbereich Geschichte und Soziologie, Universität Konstanz,
Universitätsstr. 10, 78457 Konstanz, Deutschland
E-Mail: thomas.hinz@uni-konstanz.de

lediglich eine leichte Tendenz feststellen, dass komplexe Arbeiten mit mehreren parallel zu testenden Hypothesen das PB-Risiko abmildern.

Schlüsselwörter Publication Bias · Wissenschaftssoziologie · Signifikanztest · Caliper-Test · Rational-Choice

Prevalence and Risk-Factors of Publication Bias in German Sociology

Abstract Statistical significance of research results is often misleadingly regarded as an indicator of relevance and explanatory power. Significant results have better chances of getting published than non-significant results, although both are equally important for scientific progress. Such a selection of significant results is accompanied by overestimated effect sizes and too optimistically (biased) evaluations of theories. In this article, the problem of publication bias (PB) is examined using the caliper test based on data from eleven volumes of the two leading German sociology journals (*Kölner Zeitschrift für Soziologie und Sozialpsychologie* and *Zeitschrift für Soziologie*). Is there any evidence for PB in these journals as it was detected for sociology journals in the US? Which conditions trigger the occurrence of PB? The analyses focus on the possibilities of data manipulation and social control by researchers. The results indicate that German sociology is indeed affected by PB though to a lesser extent than US journals. Editorial policies (e.g. policies that data have to be provided for replications) have not been effective so far. Only a slight tendency of a reduced PB-risk is found in case of more complex analyses, i.e. when multiple hypotheses are tested.

Keywords Publication bias · Sociology of science · Significance testing · Caliper test · Rational-choice

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).[1] 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, S. 13). Ergebnisse, welche die vom Forscher gewählte Signifikanzschwelle unterschreiten, demnach „statistisch signifikant“ sind, werden unabhängig von ihrer Qualität und Aussagekraft mit größerer Wahrscheinlichkeit

[1] 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).

niedergeschrieben, eingereicht, veröffentlicht und schlussendlich auch zitiert als nicht-signifikante Ergebnisse (Egger und 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.[2] 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 Autorenteams 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. 2014), zugleich aber einem fehlenden Wissen zu seinen Risikofaktoren relevant (Auspurg und Hinz 2011a). Für die Diagnose eines PB wird im vorliegenden Beitrag das Verfahren des Caliper-Tests (Gerber und Malhotra 2008a, b) eingesetzt.

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, S. 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, S. 1062). Es wird also fälschlicherweise die Nullhypothese, dass kein Effekt vorliegt, verworfen. Grund-

[2] 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. (2013, S. 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. € allein auf Bundesebene; Deutscher Bundestag 2013).

sätzlich 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 solchen Effekte bestehen. Allerdings werden selbst geringe Effekte bei entsprechend großer Stichprobe leichter als „überzufällig" und somit als „siginifikant" 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 und Pigott 2006, S. 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 verkannt (Rosenthal 1979, S. 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, S. 13). Hierbei lassen sich grundsätzlich zwei Vorgehensweisen unterscheiden. Erstens kann versucht werden, durch die wiederholte Erhebung von Daten (vgl. *objective publication bias*; Begg 1994, S. 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, S. 400). Im Gegensatz zum objektiven PB werden hier nicht die Daten selbst, sondern vielmehr die Ergebnisse der Auswertungen unterschlagen.[3]

Die Untersuchung von Forschungsergebnissen auf einen PB hat in den letzten Jahren in Meta-Analysen weite Verbreitung gefunden (s. z. B. Ferguson und

[3] 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.

Brannick 2012).[4] 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 und 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, S. 1127).

3 Theoretischer Rahmen

Erste Erkenntnisse zur Auftrittswahrscheinlichkeit des 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 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, S. 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 und Levy 1993, S. 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, S. 222) vergeben werden. Aufsehen erregende Publikationen, Aufsätze in (hoch renommierten) Zeitschriften und häufige Zitationen sind dabei wichtige Erfolgskriterien, um sich gegenüber der Konkurrenz durchzusetzen (Feigenbaum und Levy 1996, S. 263; Stephan 2010, S. 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

[4] 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.

2005–2010 nur zwischen 6% und 10% der eingereichten Manuskripte publiziert.[5] 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, S. 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 und Hinz 2011a; Dickersin 2005; Ferguson und 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, S. 177). In Bezug auf die Sanktionsschwere gilt allerdings, dass ein PB im Gegensatz zum offenen Betrug (wie dem Manipulieren von Daten) weit weniger stark sanktioniert wird.[6] Feigenbaum und 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, S. 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 sich jedenfalls angelehnt an das Konzept statistischer Diskriminierung (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

[5] http://www.asanet.org/journals/previous_editors_reports.cfm (Zugegriffen: 21.03.2014).

[6] 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.

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 und Hinz 2011a; Kerr 1998). Für viele Akteure dürfte unter den skizzierten Wettbewerbsstrukturen die Alternative, einen PB zu begehen, daher vermutlich lohnender erscheinen:

H1: In der untersuchten Literatur sind Anzeichen für einen PB zu finden.

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.[7] Die Konkurrenz um die limitierten Veröffentlichungsgelegenheiten ist also in der deutschen Soziologie erheblich geringer, womit zu erwarten ist:

H2: Ein PB tritt in der deutschsprachigen Soziologie in geringerem Ausmaß auf als in den höher gerankten US-amerikanischen Zeitschriften.

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:

H3: Je mehr Hypothesen getestet werden, desto geringer ist das PB-Risiko.[8]

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 Signifikanz-

[7] 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 (Zugegriffen: 21.03.2014).

[8] Eine alternative Interpretation wäre, dass bereits wenige signifikante Ergebnisse die Publikationschancen hinreichend erhöhen, sodass 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 und Hinz 2011a).

niveaus (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 und Levy 1993, S. 119). Generell ist daher anzunehmen, dass mit einer starken Dokumentationspflicht und zugänglichen Daten das Risiko für einen PB abnimmt:

H4a: Je umfangreicher die Berichtspflichten zu Datenmaterial und Analysen, desto geringer ist das PB-Risiko.

H4b: Wenn Daten allgemein zugänglich sind, dann kommt es seltener zu einem PB.

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. Ko-Autoren erhöhen zudem als „Mitwisser" die Gefahr eines sozialen Tadels (Auspurg und 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 *Whistleblower* 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.[9] Insgesamt ist der „Nettoeffekt" der Teamgröße daher schwer vorherzusagen. Vermutlich birgt aber gerade die Möglichkeit von *whistleblowern* das abschreckendste Risiko dafür, 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, S. 673 f.):

H5: Je mehr Autoren an einer Publikation beteiligt sind, desto geringer ist das PB-Risiko.

4 Forschungsstand und Methoden

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, S. 32). Diesen Befund konnten Sterling et al. in ihrer Replikation aus dem Jahr 1995 mit über 92 % solcher Artikel bestätigen.

[9] 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.

In der Medizin hingegen sind die Anteile überwiegend signifikanter Ergebnisse mit 43–82 % deutlich geringer (Sterling et al. 1995, S. 109). Eine starke Häufung von Artikeln mit überwiegend signifikanten Ergebnissen wurde überdies für US-amerikanische Soziologie-Zeitschriften (80 %) beobachtet (Wilson et al. 1973, S. 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, S. 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, S. 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 und Simes 1997), dieser Unterschied kann aber neben einem PB auch in der besseren methodischen und theoretischen Qualität der Studien begründet sein.

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)[10] 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 und Malhotra 2008a, S. 11 f.). 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 und Lalande (2012) sowie Leggett et al. (2013) finden dabei, dass *p*-Werte knapp unterhalb des 5 %-Signifikanzniveaus deutlich häu-

[10] Die Begriffe „Teststatistiken“ und „Testwerte“ werden im Folgenden synonym verwendet, gemeint sind beispielsweise *p*-, *t*-, oder *z*-Werte.

figer auftreten, als man erwarten würde (Leggett et al. 2013, S. 2305; Masicampo und Lalande 2012, S. 2274). Beide Studien zeigen damit Anzeichen für eine Überrepräsentanz knapp signifikanter Ergebnisse und damit für einen PB.[11] Allerdings haben diese Studien den Nachteil, die lediglich zwischen 0 und 1 definierte Verteilung der p-Werte, 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.[12]

Das von Gerber und Malhotra (2006) vorgeschlagene Verfahren des Caliper-Tests (CT)[13] 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 Abb. 1 schematisch dargestellt: Bei Vorliegen eines PB (rechte Grafik) ist die Verteilung unterhalb der Signifikanzschwelle (gestrichelte Linie bei einem *t*-Wert von 1,96. Dieser Wert markiert das 5 %-Signifikanzniveau) 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.

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 und Malhotra 2006, S. 12; für nähere Erläuterungen s. Online-Anhang).[14] 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, S. 652 f.) verbleiben von ursprünglich 587 Testwerten im 3 %-Caliper nur 22 Testwerte im OC und 11 im UC. Der Caliper 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.

[11] Auch Brodeur et al. (2013) finden in ihrer grafischen Analyse deutlich mehr knapp signifikante als nichtsignifikante *z*-Werte.

[12] Diese ist nur empirisch abgeleitet, jedoch nicht theoretisch begründet.

[13] 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 und Malhotra 2008b, S. 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?).

[14] Siehe http://www.uni-koeln.de/kzfss/materialien/KS-66-4-Auspurg.pdf

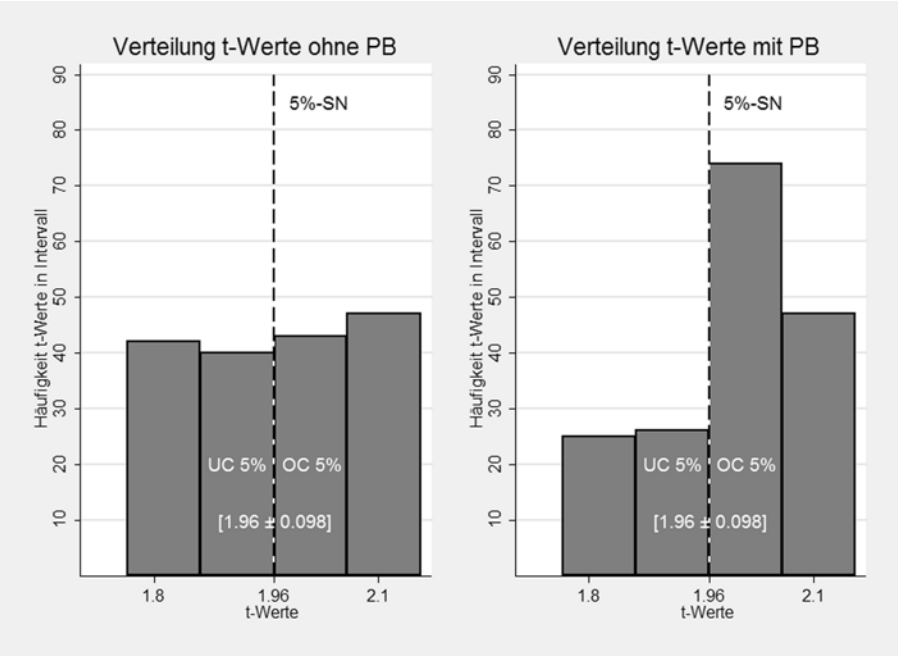

Abb. 1 Schematische Darstellung des CT. Dargestellt ist das Testprinzip für den 5%-Caliper um das 5%-Signifikanzniveau, welches einem *t*-Wert von 1,96 entspricht (gestrichelte Linie). Der 5%-Caliper deckt alle *t*-Werte ab, die in den Bereich von 1,862 bis 2,058 fallen. Die Verteilung dieser Werte wird 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

In ihren drei, den CT einsetzenden Studien stellen Gerber und Malhotra in der US-amerikanischen Soziologie (Gerber und Malhotra 2008a, S. 17) und ebenso in Zeitschriften der Politikwissenschaft (Gerber und Malhotra 2008b, S. 318) mehr knapp zum 5%-Niveau signifikante als knapp nicht-signifikante Koeffizienten 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 und Malhotra 2008a, S. 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 und 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 bis 2010 erschienenen sind, ebenfalls Hinweise auf einen PB fest, jedoch erscheint dort die Überrepräsentanz knapp signifikanter Testwerte im 3%-Caliper (mit 67%) und im 5%-Caliper (mit 60%) für das 5%-Signifikanzniveau deutlich weniger stark ausgeprägt als in den US-amerikanischen Zeitschriften (Auspurg und

Hinz 2011a, S. 653).[15] Die Ergebnisse von Auspurg und Hinz konnten auch von Weiß und Berning (2013) repliziert werden.[16]

Neben einer reinen Deskription untersuchen Auspurg und Hinz (2011a) sowie Brodeur et al. (2013) zudem den Einfluss von Text- und Autorenmerkmalen auf die Testwerteverteilung. Auspurg und Hinz (2011a) finden in bivariaten Analysen, dass explizit formulierte Hypothesen sowie Alleinautorschaften das Risiko für einen PB tendenziell erhöhen (Auspurg und Hinz 2011a, S. 654). Allerdings handelt es sich lediglich um eine Pilotstudie, die aufgrund der geringen Fallzahl keine aussagekräftigeren multivariaten Analysenemethoden verwendet und auch nur begrenzte Aussagen zur Prävalenz des PB in der deutschen Soziologie zulässt.[17]

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 nichtsignifikante Ergebnisse (Brodeur et al. 2013, S. 9). Die Verfügbarkeit der Daten scheint hingegen keinen Effekt zu haben. Allerdings wird hier lediglich explorative Forschung mit ad-hoc Annahmen und grafischen Inspektionen betrieben.

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.[18] 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.

[15] Im Gegensatz zu den anderen berichteten Studien verwenden Auspurg und Hinz (2011a) die genaueren *t*- statt *z*-Werte, was allerdings nur bei wenigen Artikeln mit sehr kleinen Fallzahlen zu unterschiedlichen Ergebnissen führt.

[16] Die Autoren kommen jedoch in ihrer Schlussfolgerung trotz signifikant überrepräsentierter Testwerte im OC des 3 %-Caliper zu dem Ergebnis, dass kein PB vorliegt.

[17] Auch Weiß und 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.

[18] Vgl. ein Impact Factor der ZfS von 0,604 und 0,481 bei der KZfSS (Journal Citation Report 2012).

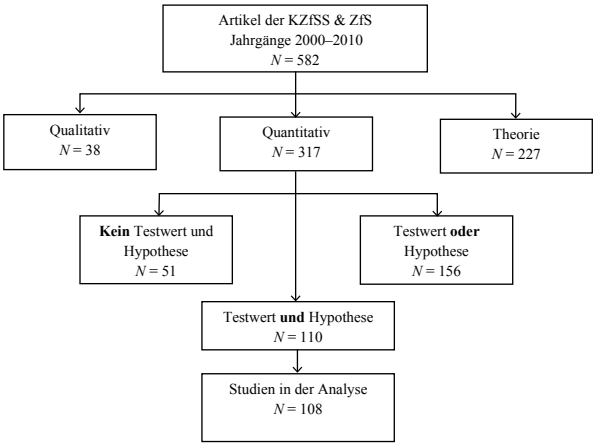

Abb. 2 Flussdiagramm zum Kodierprozess

Der in Abb. 2 dargestellte Kodierprozess[19] 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 Testwerte 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 oder rein deskriptive Auswertungen. Von den quantitativen Studien lassen sich mit diesen Auswahlkriterien insgesamt nur knapp 35 % für die nachfolgenden Auswertungen berücksichtigen ($N=108$ Studien).[20] 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.

Werden in Studien zum Test der Hypothesen mehrere Modelle berichtet, so werden (abweichend zu den Studien von Gerber und 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

[19] Die Darstellung des Kodierprozesses orientiert sich an den Empfehlungen der PRISMA Guidelines (Moher et al. 2009, S. 3).

[20] 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.

Tab. 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

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 und Hinz (2011a) hinaus.

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. Tab. 1 für deskriptive Statistiken). Die *Anzahl der in den Artikeln getesteten Hypothesen* oder *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.

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.[21] 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

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

deshalb der *Data Policy*.[22] Auf eine Unterscheidung zwischen der ZfS und der KZfSS wird in den Analysen auf Grund der großen Überschneidung zwischen Zeitschrift und Berichtspflicht bzw. *Data Policy* verzichtet. Ferner wird als Kontrollvariable aufgrund des Befundes von Auspurg und 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. Online-Anhang[23]). Im vorliegenden Datensatz sind in gut 33 % der Artikel die Hypothesen lediglich implizit formuliert.

5.1 Testverfahren und Schätzmodelle

Für die nachfolgenden Auswertungen wird der CT herangezogen (siehe Abschn. 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. Online-Anhang).[24] Zudem werden für die Testwerte zufällige Nachkommastellen imputiert, um robuste Ergebnisse im Hinblick auf potenzielle 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, S. 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).

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

[23] Siehe http://www.uni-koeln.de/kzfss/materialien/KS-66-4-Auspurg.pdf

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

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 und 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 und Clark 1999). Aus diesem Grund werden über die logistischen Regressionsmodelle 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, S. 687). Um der Mehrebenenstruktur (L1: Testwerte; L2: Artikel) Rechnung zu tragen, werden die Modelle mit geclusterten Standardfehlern geschätzt (Rogers 1994).

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.[25] Im Folgenden wird zuerst die Verteilung der Testwerte (Abb. 3) näher betrachtet, um darauf aufbauend den CT auf die in der Soziologie am weitesten verbreiteten Signifikanzniveaus (5 % und 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 (Abb. 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

[25] 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.

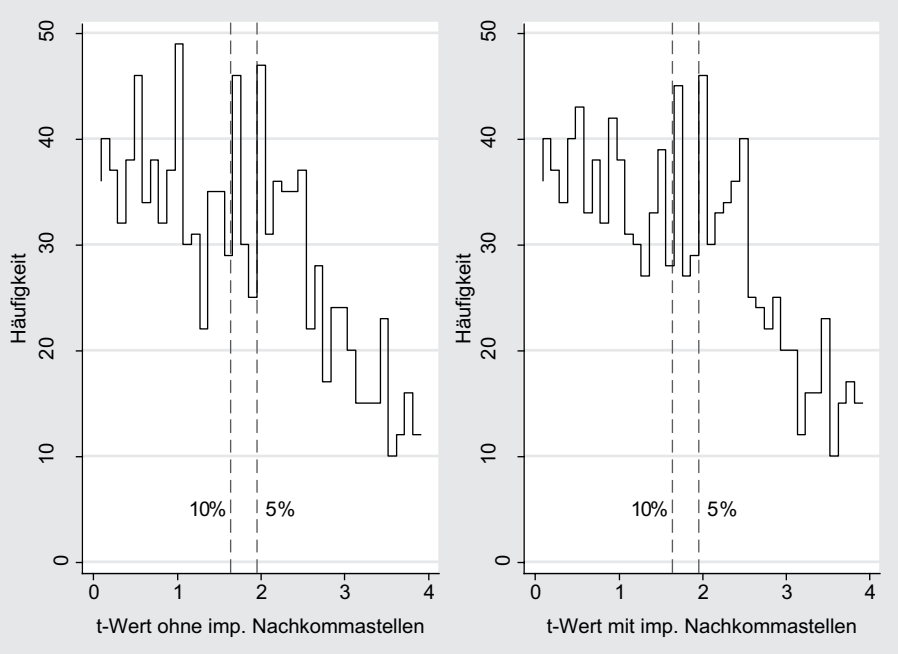

Abb. 3 Histogramm der *t*-Werteverteilung zum 5 %-Caliper

Tab. 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

um 0,5 bzw. 1 auftretenden Schwankungen sind hingegen schwer zu erklären. Die Imputation von zufälligen Nachkommastellen zur Korrektur von möglichen Rundungsungenauigkeiten (s. Abb. 3, rechts) löst dieses Problem nur in Teilen auf.

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 (Tab. 2). Im 3 %- und 5 %-Caliper sind statistisch signifikante Ergebnisse 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.[26] 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 kleinsten Caliper nur

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

Tab. 3 Caliper-Test 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

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 oder selbst nach der konservativen Holm-Korrektur 1:10 im 3- und 5 %-Caliper). Das Vorliegen eines PB ist somit sehr wahrscheinlich (*H1*).

Betrachten wir nun die Ergebnisse des CT um das 10 %-Signifikanzniveau (Tab. 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, 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 und Malhotra (2008a).[27] Diese deutlichen Unterschiede bleiben auch nach der Holm-Korrektur auf dem 10 %-Niveau signifikant (vgl. Onlineanhang, Tab. A4). Somit scheint die US-amerikanische Forschung zumindest im Bereich der Soziologie erwartungsgemäß (*H2*) deutlich stärker von einem PB betroffen zu sein als die deutschsprachige.

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

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.[28] Es werden im Folgenden AMEs aus

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

[28] Multikollinearitätsprobleme sind jedoch nicht zu befürchten ($r < 0{,}45$).

Tab. 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

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

logistischen Regressionen mit allen Kovariaten berichtet (s. Tab. 5).[29] 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 engsten berechneten Ausschnitt um das 5 %-Signifikanzniveau, dem 3 %-Caliper, lässt sich mit steigender Koeffizientenanzahl (*H3*) sowie bei Mehrfachautorenschaft (*H5*) 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 (*H4a*), die Datenzugänglichkeit (gemessen durch die potenzielle öffentliche Verfügbarkeit des Datensatzes; *H4b*) 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 (*H4b*) nur ein Nulleffekt (vgl. Onlineanhang Tab. A5).

Im 5 %-Caliper (s. ebenfalls Tab. 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.

[29] 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.

Tab. 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

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

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 asymmetrische Verteilung der herangezogenen Testwerte feststellen. Die Ergebnisse des CT sprechen insgesamt deutlich 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 und 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 der Methodik bedingt ist. Der größere Wettbewerbsdruck in der internationalen Forschung scheint also das Auftreten eines PB zu unterstützen.[30]

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 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.[31] So konnten von den quantitativen Arbeiten nur gut

[30] 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.

[31] 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)

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 und 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) 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, S. 44; Diekmann 2005, S. 26).[32] 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

die genauen Signifikanzwerte (oder damit assoziierte Testwerte, wie *t*-Statistiken oder Standardfehler) zu berichten.

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

Selbstheilungskräfte der Wissenschaft scheinen derzeit nicht auszureichen, um einen PB zu verhindern (Stroebe et al. 2012).

Literatur

Arrow, Kenneth. 1973. The theory of discrimination. In *Discrimination in labor markets,* Hrsg. Orley Ashenfelter und Albert Rees, 193–216. Princeton: Princeton UP.

Auspurg, Katrin, und Thomas Hinz. 2011a. What fuels publication bias? Theoretical and empirical analyses of risk factors using the Caliper Test. *Jahrbücher für Nationalökonomie und Statistik* 231:636–660.

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

Baerlocher, Mark O., Jeremy O'Brien, Marshall Newton, Tina Gautam und Jason Noble. 2010. Data integrity, reliability and fraud in medical research. *European Journal of Internal Medicine* 21:40–45.

Becker, Gary S. 1968. Crime and punishment: An economic approach. *Journal of Political Economy* 76:169–217.

Begg, Colon B. 1994. Publication Bias. In *The handbook of research synthesis,* Hrsg. Harris Cooper und Larry V. Hedges, 399–409. New York: Russell Sage Foundations.

Brodeur, Abel, Mathias Lé, Marc Sangnier und Yanos Zylberberg. 2013. *Star wars: The empirics strike back*. IZA Discussion Paper Series Nr. 7268. Institute for the Study of Labor. Bonn.

Brüderl, Josef. 2004. Meta-Analyse in der Soziologie: Bilanz der deutschen Scheidungsforschung oder „statistischer Fruchtsalat"? *Zeitschrift für Soziologie* 33:84–86.

Brüderl, Josef. 2013. *Sind die Sozialwissenschaften wissenschaftlich? Ergebnisse eines Replikationsexperiments*. Vortrag im Rahmen der Tagung Rational-Choice Sociology. VIU Venedig. Venedig.

Cohen, Jacob. 1994. The earth is round (p<0,05). *American Psychologist* 49:997–1003.

Dawes, Robyn M. 1980. Social dilemmas. *Annual Review of Psychology* 31:169–193.

Deutscher Bundestag. 2013. *Bundesregierung hält an Grippemittel Tamiflu fest*. Heute im Bundestag – Gesundheit/Antwort 08.05.2013. Berlin. Deutscher Bundestag.

Dickersin, Kay. 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. Hannah R. Rothstein, Alexander J. Sutton und Michael Borenstein, 11–33. Oxford: Blackwell Science.

Dickersin, Kay, Yan-I Min und Curtis L. Meinert. 1992. Factors influencing publication of research results. Follow-up of applications submitted to two institutional review boards. *JAMA* 267:374–378.

Diekmann, Andreas. 2005. Betrug und Täuschung in der Wissenschaft. Datenfälschung, Diagnoseverfahren, Konsequenzen. *Schweizerische Zeitschrift für Soziologie* 31:7–30.

Diekmann, Andreas. 2011. Are most published research findings false? *Jahrbücher für Nationalökonomie und Statistik* 231:628–636.

Diekmann, Andreas, Bettina Heintz, Richard Münch, Ilona Ostner und Hartmann Tyrell. 2002. Editorial. *Zeitschrift für Soziologie* 31:1–3.

Easterbrook, Philippa J., Jesse A. Berlin, Ramana Gopalan und David R. Matthews. 1991. Publication Bias in Clinical Research. *Lancet* 337:867–872.

Egger, Matthias, und George D. Smith. 1998. Meta-analysis bias in location and selection of studies. *British Medical Journal* 316:61–66.

Fanelli, Daniele. 2009. How many scientists fabricate and falsify research? A systematic review and meta-analysis of survey data. *PLoS ONE* 4:e5738.

Fanelli, Daniele. 2012. Negative results are disappearing from most disciplines and countries. *Scientometrics* 90:891–904.

Feigenbaum, Susan, und David M. Levy. 1993. The market for (ir)reproducible econometrics. *Accountability in Research* 3:25–43.

Feigenbaum, Susan, und David M. Levy. 1996. The technological obsolescence of scientific fraud. *Rationality and Society* 8:261–276.

Ferguson, Christopher J., und Michael 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.

Fisher, Ronald A. 1973. *Statistical methods for research workers*. New York: Hafner.
Gerber, Alan S., und Neil 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, Alan S., und Neil Malhotra. 2008a. Publication Bias in empirical sociological research. *Sociological Methods & Research* 37:3–30.
Gerber, Alan S., und Neil 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, Alan S., Neil Malhotra, Conor M. Dowling und David Doherty. 2010. Publication Bias in two political behavior literatures. *American Politics Research* 38:591–613.
Greene, William H. 2012. *Econometric analysis*. Boston: Prentice Hall.
Hart, Robert A., und David H. Clark. 1999. *Does size matter? Exploring the small sample properties of maximum likelihood estimation*. Annual Meeting of the Midwest Political Science Association.
Hirschi, Travis. 1969. *Causes of delinquency*. Berkeley: University of California Press.
Holm, Sture. 1979. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* 6:65–70.
Jefferson, Tom, Mark A. Jones, Peter Doshi, Chris B. Del Mar, Carl J. Heneghan, Rokuro Hama und Matthew J. Thompson. 2013. Neuraminidase inhibitors for preventing and treating influenza in healthy adults. *Cochrane Database of Systematic Reviews*.
Kerr, Norbert L. 1998. HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review* 2:196–217.
Kollock, Peter. 1998. Social dilemmas: The anatomy of cooperation. *Annual Review of Sociology* 24:183–214.
Labovitz, Sanford. 1968. Criteria for selecting a significance level: A note on the sacredness of.05. *The American Sociologist* 3:220–222.
Leggett, Nathan C., Nicole A. Thomas, Tobias Loetscher und Michael E. R. Nicholls. 2013. The life of p: „Just significant" results are on the rise. *Quarterly Journal of Experimental Psychology* 66:2303–2309.
Mahoney, Michael J. 1977. Publication prejudices: An experimental study of confirmatory bias in the peer review system. *Cognitive Therapy and Research* 1:161–175.
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.
Merton, Robert K. 1957. Priorities in scientific discovery: A chapter in the sociology of science. *American Sociological Review* 22:635–659.
Merton, Robert K. 1961. Singletons and multiples in scientific discovery: A chapter in the sociology of science. *Proceedings of the American Philosophical Society* 105:470–486.
Moher, David, Alessandro Liberati, Jennifer Tetzlaff, Douglas G. Altman und The Prisma Group. 2009. Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLoS Med* 6:e1000097.
Necker, Sarah. 2012. Wissenschaftliches Fehlverhalten – ein Problem in der deutschen Volkswirtschaftslehre? *Perspektiven der Wirtschaftspolitik* 13:267–285.
Nuzzo, Regina. 2014. Scientific method: Statistical errors. *Nature* 506:150–152.
Phelps, Edmund S. 1972. The statistical theory of racism and sexism. *The American Economic Review* 62:659–661.
Rogers, William. 1994. Regression standard errors in clustered samples. *Stata Technical Bulletin* 3:19–23.
Rosenthal, Robert. 1979. The file drawer problem and tolerance for null results. *Psychological Bulletin* 86:638–641.
Sahner, Heinz. 1979. Veröffentlichte empirische Sozialforschung – Eine Kumulation von Artefakten eine Analyse von Periodika. *Zeitschrift für Soziologie* 8:267–278.
Simonsohn, Uri, Leif D. Nelson und Joseph P. Simmons. 2014. P-curve: A key to the file drawer. *Journal of Experimental Psychology: General* 143: 534–547.
Skipper, James K., Jr., Anthony L. Guenther, und Gilbert 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, Michael. 1985. Utilitarianism, moral dilemmas, and moral cost. *American Philosophical Quarterly* 22:161–168.
Stephan, Paula E. 2010. The economics of science. In *Handbook of the economics of innovation,* Hrsg. Bronwyn H. Hall und Nathan Rosenberg, 1, 217–274. Amsterdam: North Holland.

Sterling, Theodore 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, Theodore. D., Wilfred L. Rosenbaum und James 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, Jerome M., und R. John Simes. 1997. Publication Bias: Evidence of delayed publication in a cohort study of clinical research projects. *British Medical Journal* 315:640–645.
Sterne, Jonathan A. C., David Gavaghan und Matthias 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.
Stroebe, Wolfgang, Tom Postmes und Russell Spears. 2012. Scientific misconduct and the myth of self-correction in science. *Perspectives on Psychological Science* 7:670–688.
Sutton, Alexander J., und Therese D. Pigott. 2006. Bias in meta-analysis induced by incompletely reported studies. In *Publication Bias in meta-analysis,* Hrsg. Hannah Rothstein, Alexander J. Sutton und Michael Borenstein, 221–239. Chichester: John Wiley & Sons.
Weiß, Bernd, und Michael Wagner. 2008. Potentiale und Probleme von Meta-Analysen in der Soziologie. *Sozialer Fortschritt* 10/11:250–255.
Weiß, Bernd, und Carl 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.
Wilson, Franklin D., Gale L. Smoke und J. David Martin. 1973. The replication problem in sociology: A report and a suggestion. *Sociological Inquiry* 43:141–149.
Zeitschrift für Soziologie. 2012. Editorial. *Zeitschrift für Soziologie* 41:2–4.

Katrin Auspurg, 1974, Dr. rer. soz., Professorin für Soziologie mit Schwerpunkt quantitative Methoden der empirischen Sozialforschung, Fachbereich Gesellschaftswissenschaften, Goethe-Universität Frankfurt. Forschungsgebiete: Methoden der empirischen Sozialforschung, Survey Methodology, Arbeitsmarktsoziologie, Soziale Ungleichheit. Veröffentlichungen: Berufliche Umzugsentscheidungen in Partnerschaften. Eine experimentelle Prüfung von Verhandlungstheorie, Frame-Selektion und Low-Cost-These. Kölner Zeitschrift für Soziologie und Sozialpsychologie 66, 2014 (mit C. Frodermann und T. Hinz). Gruppenvergleiche bei Regressionen mit binären abhängigen Variablen – Probleme und Fehleinschätzungen am Beispiel von Bildungschancen im Kohortenverlauf. Zeitschrift für Soziologie 40, 2011 (mit T. Hinz).

Thomas Hinz, 1962, Dr. rer. pol., Professor für Soziologie, Fachbereich Geschichte und Soziologie, Universität Konstanz. Forschungsgebiete: Methoden der empirischen Sozialforschung, Survey Methodology, Arbeitsmarktsoziologie. Veröffentlichungen: Arbeitsmarktsoziologie. Probleme, Theorien, empirische Befunde, Wiesbaden 2009 (hrsg. mit M. Abraham); Organisationssoziologie, Sonderheft der Kölner Zeitschrift für Soziologie und Sozialpsychologie 41. Wiesbaden 2002 (hrsg. mit J. Allmendinger).

Andreas Schneck, 1987, Soziologie (MA), wissenschaftlicher Mitarbeiter an der Professur für Soziologie mit Schwerpunkt quantitative Methoden der empirischen Sozialforschung, Fachbereich Gesellschaftswissenschaften, Goethe-Universität Frankfurt. Forschungsgebiete: Methoden der empirischen Sozialforschung, Meta-Analysen, Wissenschaftssoziologie (insb. Publication Bias).