Katholische Universität Eichstätt-Ingolstadt
Philosophisch-Pädagogische Fakultät
Allgemeine Psychologie II

Masterarbeit

Fragwürdige Forschungspraktiken in der Psychologie: Eine vergleichende *p*-Kurven-Analyse zum Stand der Replikationskrise

Questionable Research Practices in Psychology:
A Comparative *p*-Curve Analysis
of the State of the Replication Crisis

Alexander Rubenbauer
info@rubenbauer.info

Manuel Rausch, Ph.D.
Gutachter

Eingereicht am 18. September 2019

Inhaltsverzeichnis

Zusammenfassung....4
Abstract....5
Einleitung....6
Theoretischer Hintergrund....7
Theorie der *p*-Werte....7
Statistische Missverständnisse....8
Publikationsbias....9
Ursachen des Publikationsbias....9
Ausmaß des Publikationsbias....10
Auswirkungen des Publikationsbias....10
Bestätigungstendenz....10
Fragwürdige Forschungspraktiken....11
Betrug in der Sozialpsychologie....13
Replikationsdilemma....14
Ungenaue Methodenbeschreibungen....15
Ursachen für die Replikationskrise....16
Das „Reproducibility Project“....16
Problematischer Untersuchungsgegenstand....17
Kumulative Evidenz....18
Zu geringe Teststärken....19
Ausbleibende Zurücknahmen....19
Daten werden nicht geteilt....20
Hypothesen....21
Methode....22
Forschungsdesign....22
Voraussetzungen der *p*-Kurven-Analyse....23
Vorgehen bei der *p*-Kurven-Analyse....23
Methodik der *p*-Kurven-Analyse....24
Ablauf und Selektionsregel....25
Stichprobe....26

Auswertung und Ergebnisse....27
Ergebnisse der *p*-Kurven-Analyse....27
Hypothese A: Allgemeine Psychologie....27
Hypothese B: Sozialpsychologie....31
Hypothese C: Verschiedenheit der Verteilungen....33
Ergebnisse der *Z*-Kurven-Analyse....34
Diskussion....35
Diskussion der Allgemeinen Psychologie....35
Diskussion der Sozialpsychologie....36
Schwierigkeiten bei der Datenerhebung....37
Praktikabilität der Methode....38
Kritik an der Methode....38
Grenzen dieser Studie....40
Ausblick....40
Weitere Anwendungsgebiete....41
Maßnahmen gegen die Replikationskrise....42
Präregistrierung von Studien....42
Bayes'sches Testen....43
Open Access....43
Open Data....45
Höhere Teststärken....46
Feindliche Zusammenarbeit....47
Auszeichnungen....47
Replikationszähler....47
Offenlegungspflichten....48
Sci-Hub....49
Ein *Wiki* für Forschungsergebnisse....50
Positive Entwicklungen....51
Literaturverzeichnis....52

Zusammenfassung

Nachdem in der Sozialpsychologie mehrere Betrugsfälle aufgedeckt wurden und sich ein großer Teil der psychologischen Forschungsergebnisse als nicht reproduzierbar erwiesen hat, kam es zu einem Vertrauensverlust in die psychologische Forschung. Die vorliegende Studie beleuchtet die Ursachen dieser Replikationskrise. Der Einsatz fragwürdiger Forschungspraktiken, wie das so genannte *p-Hacking*, verfälscht Forschungsergebnisse und trägt dadurch entscheidend zur Replikationskrise bei. In dieser Studie wurde das gegenwärtige Ausmaß fragwürdiger Forschungspraktiken in der Psychologie empirisch untersucht, indem die Fachgebiete Allgemeine Psychologie und Sozialpsychologie einer vergleichenden *p*-Kurven-Analyse unterzogen wurden. Dabei wurde anhand von $N = 76$ *p*-Werten überprüft, ob deren Verteilung linkssteil oder rechtssteil ausgeprägt ist. Eine linkssteile Verteilung weist überwiegend hochsignifikante *p*-Werte auf und gilt als Indiz für die empirische Evidenz der zugrundeliegenden Forschungsergebnisse. Eine rechtssteile Verteilung hingegen weist mehrheitlich knapp signifikante *p*-Werte auf und gilt als Indiz für den Einsatz fragwürdiger Forschungspraktiken. In dieser Studie wurde kein signifikanter Unterschied festgestellt: In beiden Fachgebieten waren die *p*-Werte signifikant linkssteil verteilt, was für die empirische Evidenz der zugrundeliegenden Forschungsergebnisse spricht. Im Übrigen wurde die Praktikabilität der *p*-Kurven-Analyse diskutiert und mögliche Lösungen für die Replikationskrise aufgezeigt.

Stichworte: Allgemeine Psychologie, Experimentelle Psychologie, Forscher-Freiheitsgrade, Fragwürdige Forschungspraktiken, Kognitionspsychologie, *p*-Hacking, *p*-Kurven-Analyse, Publikationsbias, Replikationskrise, Sozialpsychologie

Abstract

After several cases of fraud have been discovered in social psychology, and a large part of psychological research has been proven to be non-reproducible, a crisis of confidence emerged. This study shed light on the causes of this replication crisis. The use of questionable research practices, such as *p*-hacking, falsifies research results and therefore contributes to the crisis. The current extent of the use of questionable research practices was empirically investigated by conducting a *p*-curve analysis comparing the fields of cognitive psychology and social psychology. It was examined whether the distribution of the fields' *p*-values is left-skewed or right-skewed. A right-skewed distribution would predominantly show highly significant *p*-values like $p \leq .01$, while a left-skewed distribution would mainly contain *p*-values just below the significance threshold of $p = .05$. The result of a *p*-curve analysis can be indicative of empirical evidence of the underlying research or the use of questionable research practices. Both fields' distributions of *p*-values were significantly right-skewed, suggesting empirical evidence of its research. Furthermore, *p*-curve's practicability has been discussed, as well as possible solutions for the replication crisis.

Keywords: cognitive psychology, experimental psychology, *p*-curve analysis, *p*-hacking, publication bias, questionable research practices (QRPs), replicability crisis, replication crisis, reproducibility crisis, researcher degrees of freedom, social psychology

Einleitung

In den vergangenen zehn Jahren kam es unter anderem in der Psychologie zu einer so genannten Replikationskrise. Diese hat in den Medien viel Aufmerksamkeit erfahren und in der Wissenschaft eine große Debatte über die Vertrauenswürdigkeit von Forschungsergebnissen ausgelöst (Pashler & Wagenmakers, 2012). Unter anderem wurde die Replikationskrise durch aufgedeckte Betrugsfälle in der Sozialpsychologie ausgelöst (Pashler & Wagenmakers, 2012). Zudem weisen psychologische Studien häufig zu geringe Teststärken auf, was *per se* zu geringen Replikationsraten führt (Bezeau & Graves, 2001; Cohen, 1962; Open Science Collaboration, 2015; Sedlmeier & Gigerenzer, 1989). Gleichzeitig sehen sich Wissenschaftler vonseiten der Universitäten, Geldgeber und Verlage einem zunehmenden Druck ausgesetzt, so oft wie möglich neuartige und positive Forschungsergebnisse zu veröffentlichen (Nosek et al., 2015). Nach dem Motto *„publish or perish“* werden Aufmerksamkeit und Forschungsgelder eher denjenigen Wissenschaftlern zuteil, die häufiger veröffentlichen. Dieser Publikationsbias ist einer der Hauptgründe dafür, warum Wissenschaftler zu fragwürdigen Forschungspraktiken greifen (Chambers, 2017; Fanelli, 2010, 2012, 2013; Nosek et al., 2015; Rosenthal, 1979; Sterling, 1959). Von 2.000 befragten Psychologen gab die Mehrheit an, mindestens einmal fragwürdige Forschungspraktiken genutzt zu haben (John, Loewenstein & Prelec, 2012). Ein weiterer Grund für scheiternde Replikationen liegt in der Motivation, theoretische Überlegungen zu Effekten empirisch fundieren zu wollen, von deren Existenz man überzeugt ist (*confirmation bias*) – auch wenn dies nur durch mehr oder weniger bewusste Manipulation erreicht werden kann (Kuhn, 1989, S. 677; Kunda, 1990; Nickerson, 1998, S. 177).

Eine Möglichkeit, die Reproduzierbarkeit von publizierten Ergebnissen in einem Fachgebiet oder in einem Forschungsfeld zu überprüfen, besteht in der Durchführung von Replikationsstudien. Dies ist jedoch mit einem hohen Aufwand verbunden. Eine weniger aufwendige Methode ist die *p*-Kurven-Analyse. Dabei wird die Verteilung der *p*-Werte in einer hinreichend großen Anzahl von Studien analysiert und überprüft, ob die Verteilung der *p*-Werte empirische Evidenz besitzt. Tut sie dies, dann ist von einem tatsächlichen Effekt auszugehen, da fragwürdige Forschungspraktiken als Hauptursache für die signifikanten Ergebnisse ausgeschlossen werden können. Mangelt es hingegen an

empirischer Evidenz, dann muss überprüft werden, ob die Verteilung auch bei Abwesenheit eines tatsächlichen Effekts entstanden sein könnte oder ob fragwürdige Forschungspraktiken eine Rolle gespielt haben könnten.

Vor dem Hintergrund der Replikationskrise, von der besonders die Sozialpsychologie betroffen war, soll in dieser Studie mittels einer *p*-Kurven-Analyse auf Basis aktueller Forschungsergebnisse überprüft werden, ob sich die Verteilung der *p*-Werte in der Allgemeinen Psychologie signifikant von der Verteilung der *p*-Werte in der Sozialpsychologie unterscheidet. Dem liegt die Annahme zugrunde, dass eine rechtssteile Verteilung der *p*-Werte ein Indiz für den Einsatz fragwürdiger Forschungspraktiken ist.

Theoretischer Hintergrund

In der Psychologie hat sich die Praxis durchgesetzt, nur Forschungsergebnisse zu veröffentlichen, die signifikante Ergebnisse aufweisen. Nichtsignifikante Ergebnisse verschwinden dagegen in der Schublade (*file drawer problem;* Rosenthal, 1979). Dies führt einerseits dazu, dass häufig Effekte untersucht werden, die es gar nicht gibt. Andererseits wird durch einseitige Publikationsweise die Forschungsrealität nicht repräsentativ abgebildet, weshalb es kaum möglich ist, den tatsächlichen Stand der Forschung anhand der veröffentlichten Literatur zu bestimmen (Fanelli, 2012; Ioannidis, 2005, 2008; Pashler & Harris, 2012). Eine mögliche Lösung für dieses Jahrzehnte alte Problem stellt die *p*-Kurven-Analyse dar (Simonsohn, Nelson & Simmons, 2014a).

Theorie der *p*-Werte

Ein *p*-Wert gibt die prozentuale Wahrscheinlichkeit an, mit der ein Effekt (oder ein noch größerer Effekt) in einem Datensatz beobachtet werden kann, wenn tatsächlich gar kein Effekt existiert, also die Nullhypothese gilt (Chambers, 2017, S. 24 ff.). Folglich gilt: Je kleiner der *p*-Wert ausfällt, desto besser bzw. desto „signifikanter“ ist das Ergebnis. Ein *p*-Wert sagt dabei weder etwas darüber aus, mit welcher Wahrscheinlichkeit ein Effekt existiert, noch mit welcher Wahrscheinlichkeit er *nicht* existiert, und auch nicht, wie groß der Effekt ist. Ein *p*-Wert besagt lediglich, wie überrascht man darüber sein müsste, dass ein beobachteter Effekt in Wahrheit gar nicht existiert: Je kleiner der erzielte *p*-Wert, desto größer wäre die Überraschung (Chambers, 2017, S. 24 ff.).

In der psychologischen Forschung spricht man in der Regel von signifikanten *p*-Werten, wenn diese ein Signifikanzniveau von α = .05 unterschritten haben. Diese kritische Schwelle wurde von Fisher (1926, S. 85) seinerzeit willkürlich gewählt: *„Personally, the writer prefers to set a low standard of significance at the 5 per cent point, and ignore entirely all results which fail to reach this level.“*. Fisher gehört neben Pearson, Gosset, Neyman sr. und Neyman jr. zu den Erfindern des *null hypothesis significance testing (NHST)*. Mit der Einführung einer Signifikanzschwelle von $p = .05$ wollte Fisher ein klares Entscheidungskriterium hinsichtlich der Frage einführen, ob ein Effekt als signifikant zu gelten hat oder nicht. Er verwies allerdings darauf, dass man auch ein strengeres Signifikanzniveau von 2 % oder 1 % nutzen könne (Fisher, 1926). Die von ihm bevorzugten 5 % wurden jedoch von den meisten Forschern übernommen, und haben sich in der Folge als Konvention durchgesetzt, die bis heute kaum hinterfragt wird (Chambers, 2017, S. 24).

Da man sich beim *NHST* mit jeder Testung irrt, ist es notwendig, die für ein Experiment benötigte Stichprobengröße zuvor (*a priori*) festzulegen, und die Daten erst im Anschluss – also nach Erreichen der benötigten Stichprobengröße – zu analysieren. Auf einem Signifikanzniveau von α = .05 irrt man sich mit einer Wahrscheinlichkeit von 5 %. Somit ist es theoretisch möglich, die Nullhypothese (die besagt, dass es keinen Effekt gibt) bei jedem 20. positiven Ergebnis fälschlicherweise abzulehnen (Chambers, 2017, S. 24 ff.). Durch den vorsätzlichen – oder zumindest unbewusst-fahrlässigen – Einsatz diverser Analyseverfahren steigt die Irrtumswahrscheinlichkeit noch weiter: Multiples Testen kann die Irrtumswahrscheinlichkeit schnell von 5 % auf 60,7 % und mehr erhöhen (Simmons, Nelson & Simonsohn, 2011). Zu einer solchen Alphafehlerkumulierung kann es schon dadurch kommen, dass Forscher aus praktisch-ökonomischen Erwägungen heraus den Umfang ihrer Stichprobe so lange erweitern, bis ein signifikantes Ergebnis erzielt wird (Chambers, 2017, S. 26 f.). In über 50 % der Analysen wird mindestens ein Haupteffekt oder eine Interaktion durch Zufall statistisch signifikant (Bishop, 2013).

Statistische Missverständnisse

Der Einsatz fragwürdiger Forschungspraktiken resultiert oftmals aus Missverständnissen bezüglich des *NHST*-Ansatzes (Chambers, 2017). So sagt ein *p*-Wert weder etwas über die Größe eines beobachteten Effekts aus, noch darüber, wie wahrscheinlich eine Hypothese tatsächlich zutrifft. Somit ist es durchaus möglich, dass ein großer Effekt

mit einem *p*-Wert von $p = .04$ für Theorie und Praxis bedeutender sein kann als ein kleiner Effekt mit einem *p*-Wert von $p = .001$, der aus einer deutlich größeren Stichprobe bezogen wurde.

Publikationsbias

Der Publikationsbias, *publication bias* (Sterling, 1959, Sterling, Rosenbaum & Weinkam, 1995) oder auch *file drawer problem* (Rosenthal, 1979) führt zu einer verzerrten Datenlage über den tatsächlichen Forschungsstand, weil „positive", d. h. signifikante Ergebnisse bevorzugt veröffentlicht werden. Auch Fachzeitschriften und insbesondere jene mit hohem Einflussfaktor (*impact factor*) veröffentlichen häufig nur „positive" Ergebnisse, die oftmals einen Neuigkeitswert aufweisen müssen (Chambers, 2017). Wissenschaftler sind darauf angewiesen, befristete Verträge verlängert zu bekommen, Stellenangebote zu erhalten und Forschungsgelder einzuwerben. Dadurch steigt der Druck, statistisch signifikante Ergebnisse zu erzielen, um die eigenen Ideen einem breiteren Fachpublikum präsentieren zu können. Solange dieser Erfolgsdruck mit einem Verhalten in Konflikt steht, welches den kollektiven Nutzen für die Wissenschaft in den Vordergrund stellt, ist der Einsatz fragwürdiger Forschungspraktiken rational nachvollziehbar (Chambers, 2017, S. 27).

Ursachen des Publikationsbias. Eine Ursache für den Publikationsbias liegt in der statistischen Methode des weit verbreiteten *null hypothesis significance testing (NHST),* da diese nur ermöglicht, eine Nullhypothese nicht abzulehnen – nicht aber sie anzunehmen (Chambers, 2017, S. 72). Bei einem nichtsignifikanten Ergebnis ist deshalb nicht klar, ob es lediglich noch an „Beweisen" für die Annahme der Alternativhypothese mangelt, oder ob die Nullhypothese tatsächlich die Realität abbildet (Hoekstra, Finch, Kiers & Johnson, 2006, S. 1033; Świątkowski & Dompnier, 2017, S. 115). Insbesondere kann keine Aussage darüber getroffen werden, mit welcher Wahrscheinlichkeit die Nullhypothese wahr ist. Dementsprechend ist beim *NHST*-Ansatz nie klar, ob tatsächlich kein Effekt vorliegt, ob nur der Stichprobenumfang zu gering war, oder ob das Messinstrument ungeeignet war oder falsch kalibriert wurde. Darum können nichtsignifikante Effekte als Inkompetenz aufseiten des durchführenden Wissenschaftlers missverstanden werden (Greenwald, 1975). Bei wissenschaftlichen Novizen kann es zwar geboten sein, genauer hinzusehen; ansonsten gibt es jedoch keine Veranlassung, zu glauben, dass ein nichtsignifikantes Ergebnis weniger aussagekräftig ist als ein signifikantes (Greenwald,

1975). Inkompetenz kann auch falsch-positive Ergebnisse verursachen (z. B. durch multiples Testen) (Greenwald, 1975). Zudem führt die regelmäßige Nichtakzeptanz von Nulleffekten jegliche Replikationsversuche *ad absurdum* und es stellt sich die Frage, was der Sinn von Replikationen sein soll, wenn gescheiterte Replikationen letztlich nicht anerkannt werden (Ferguson & Heene, 2012, S. 559).

Ausmaß des Publikationsbias. Es entbehrt nicht einer gewissen Ironie, dass die Psychologie, die den Publikationsbias entdeckt hat, mit 91,5 % den höchsten Anteil positiver Ergebnisse berichtet (Fanelli, 2010, 2012). Bei den Weltraumwissenschaften ist der Anteil mit 70,2 % zwar am geringsten (Fanelli, 2010), weist damit aber noch immer ein erstaunliches Maß an Publikationsbias auf. Der Anteil positiver Ergebnisse stieg zwischen 1990 und 2007 insgesamt um 22 % an (Fanelli, 2012).

Auswirkungen des Publikationsbias. Ein System, das negative Ergebnisse missbilligt, verzerrt nicht nur unmittelbar die wissenschaftliche Literatur (Fanelli, 2012), sondern verhindert auch die Durchführung riskanter Forschungsprojekte. Zudem nötigt es Wissenschaftler dazu, ihre Daten zu manipulieren. Der Publikationsbias ist ein Grund dafür, dass fragwürdige Forschungspraktiken bis hin zum vorsätzlichen Betrug überhaupt vorkommen (Chambers, 2017, S. 106). Wenn Wissenschaftler positive Ergebnisse erzielen müssen, werden sie sicherstellen, dass diese Ergebnisse entstehen. Zudem sorgt der Publikationsbias dafür, dass mehrere Wissenschaftler am gleichen Untersuchungsgegenstand forschen, voneinander aufgrund unveröffentlichter (weil nichtsignifikanter) Ergebnisse aber nichts erfahren. Dies führt zu einem unökonomischen Einsatz von Ressourcen wie Zeit, Steuergeldern, etc. (Chambers, 2017).

Bestätigungstendenz

Die Bestätigungstendenz (*confirmation bias*) beschreibt die Tendenz von Menschen, theoretische Überlegungen zu Effekten, von deren Existenz sie überzeugt sind, empirisch fundieren zu wollen, auch wenn dies nur durch mehr oder weniger bewusste Manipulation erreicht werden kann (Kunda, 1990; Nickerson, 1998). Anders ausgedrückt werden beim Testen einer Hypothese bevorzugt Informationen gesucht, die dazu geeignet sind, die Hypothese zu bestätigen (Dorsch, 2019). Somit liefert die Bestätigungstendenz einen weiteren Baustein zur Erklärung verzerrter Forschungsergebnisse und somit auch für scheiternde Replikationen (ein alltägliches Beispiel findet sich bei Johnson, 2013).

Die Studien von Wason (1960, 1968) zeugen von der Existenz der Bestätigungstendenz und diskutieren dessen Implikationen für die wissenschaftliche Praxis: Beispielsweise stellte sich heraus, dass Gutachter starke Vorurteile gegen Forschungsergebnisse hatten, die ihren eigenen Erwartungen zuwiderliefen (Mahoney, 1977).

Fragwürdige Forschungspraktiken

Wicherts et al. (2016) listen 34 Freiheitsgrade auf, die Forscher bei der Formulierung ihrer Hypothesen, während der Studienkonzeption und -durchführung sowie bei der Auswertung und Publikation der Ergebnisse nutzen können; darunter unter anderem:

- Durchführung explorativer Forschung ohne *a priori* formulierte Hypothesen
- Erheben von Zusatzvariablen, die auf unterschiedliche Weise in die spätere Analyse einbezogen werden können
- Messung der selben abhängigen Variable auf unterschiedliche Weisen
- Messung von Zusatzvariablen, die den späteren Ausschluss von Studienteilnehmern rechtfertigen
- Unterlassen einer Teststärkenanalyse
- Unterlassen von Randomisierung
- Unterlassen von Blindsetzung der Studienteilnehmer und/oder des Versuchsleiters über die jeweilige Versuchsbedingung
- Durchführung von Signifikanztests, während die Erhebung noch läuft
- Einzelfallbasierte Entscheidungen anstelle von *a priori* festgelegten Kriterien über den Umgang mit Ausreißern (z. B. Extremwerte ausschließen oder beibehalten, je nachdem, was sich günstiger auf den *p*-Wert auswirkt) sowie den Umgang mit fehlenden Daten und Verletzungen statistischer Voraussetzungen
- Ausprobieren verschiedener statistischer Methoden und Analyseverfahren
- Bestimmung des Signifikanzniveaus im Nachhinein
- Behinderung von Replikationsversuchen durch ungenaue Beschreibung der zugrundeliegenden Methodik
- Verschweigen gescheiterter (Pilot-)Studien (*file drawer effect*)
- Fehlerhafte Angabe von *p*-Werten oder Freiheitsgraden (z. B. durch abschneiden, abrunden, fälschen)
- Nachträgliches *Framing* explorativer Analysen als konfirmatorische Hypothesen (*HARKing: hypothesizing after results are known*; vgl. Kerr, 1998).

John et al. (2012) befragten 2.000 Psychologen, ob diese bereits einmal fragwürdige Forschungspraktiken angewendet hätten. 100 % der befragten Psychologen gaben an, je nach Auswirkung Ausreißer weggelassen oder beibehalten zu haben. 100 % gaben an, weitere Daten erhoben zu haben, weil noch kein signifikantes Ergebnis erreicht worden war. 75 % gaben an, nicht alle durchgeführten Bedingungen berichtet zu haben. 50 % haben mit der Datenerhebung aufgehört, sobald „ein erwünschtes Ergebnis“ erzielt wurde. Basierend auf diesen „überraschend hohen“ (John et al., 2012, S. 524) Selbsteingeständnissen schätzen die Forscher, dass alle Psychologen bei mindestens einer Gelegenheit fragwürdige Forschungspraktiken angewendet haben (eine Kritik findet sich bei Fiedler & Schwarz, 2015). Chambers (2017, S. 29) geht deshalb davon aus, dass *p-Hacking* in der Psychologie eher die Norm und weniger die Ausnahme darstellt. Wenn *p-Hacking* aber die Norm ist, dann müsste sich ein großer Teil der *p*-Werte knapp unter der Schwelle von $p = .05$ befinden. Dies dürfte umso mehr für die Sozialpsychologie gelten, da in diesem Fachgebiet bereits der Einsatz fragwürdiger Forschungspraktiken – bis hin zu Fällen von tiefgreifendem Betrug – nachgewiesen werden konnte.

Multiples Testen. Ein großes Missbrauchspotential hinsichtlich fragwürdiger Forschungspraktiken besteht beim multiplen Testen. Dabei werden die bereits erhobenen Daten immer wieder auf Erreichen der Signifikanzschwelle von $p \leq .05$ hin analysiert. Bei Nichterreichen werden so lange weitere Daten erhoben, bis die Signifikanzschwelle unterschritten ist. Da sich die Zusammensetzung der Daten (die Datenbasis) mit jedem weiteren Datensatz ändert, und man bei jeder Analyse eine Irrtumswahrscheinlichkeit von 5 % hat, liegt die Wahrscheinlichkeit, ein signifikantes Ergebnis zu erhalten, obwohl in Wahrheit die Nullhypothese gilt, aufgrund der so genannten Alphafehlerkumulierung bereits nach fünf Durchgängen bei 25 % (Chambers, 2017).

Anschaulich wurde die Problematik des multiplen Testens in einer Studie von Bennett, Baird, Miller und Wolford (2009), die einen neuen Magnetresonanztomographen (MRT) testeten, indem sie einen toten atlantischen Lachs scannten. Für den Scan wurde das digitale Abbild des Lachses in 130.000 Voxel (dreidimensionale Pixel) eingeteilt. Anschließend wurden dem Lachs Bilder von Menschen gezeigt, deren Emotionen er erkennen sollte. Dabei wurde scheinbar eine signifikante Aktivität im Gehirn des Lachses festgestellt. Die Forscher wollten damit demonstrieren, dass unkorrigiertes multiples Testen bei 130.000 Voxel in einer typischen fMRT-Studie fast zwangsweise zu

falsch-positiven Effekten führt und spannende, aber nicht reproduzierbare, Publikationen zur Folge haben kann. Vor der Veröffentlichung von Bennett et al. (2009) haben zwischen 25 und 40 % der fMRT-Studien keine entsprechenden Korrekturen durchgeführt (Hemmelder, 2016). Diese Hypothese wird von Masicampo und Lalande (2012) gestützt: Die beiden Forscher untersuchten 3.627 *p*-Werte aus drei renommierten psychologischen Fachzeitschriften. Kleinere *p*-Werte wurden zwar häufiger veröffentlicht; diejenigen knapp unter der Signifikanzschwelle waren jedoch fünf Mal häufiger als erwartet. Leggett, Thomas, Loetscher und Nicholls (2013) konnten zudem nachweisen, dass die Häufigkeiten zwischen den Jahren 1965 und 2005 weiter zugenommen haben. Leggett et al. (2013) vermuten, dass dies mit der Verbreitung schneller Analysemöglichkeiten durch Software wie *JASP* (JASP Team, 2019), *PSPP* (GNU Project, 2019), *R* (R Core Team, 2013) und *SPSS* (IBM, 2017) zusammenhängt, während die Berechnungen der *p*-Werte früher aufwendig und teilweise sogar von Hand geschehen mussten. Dieser Aufwand war gewissermaßen ein natürlicher Schutz gegen multiples Testen.

Betrug in der Sozialpsychologie

Neben fragwürdigen Forschungspraktiken, zu deren Anwendung sich die Mehrheit von 2.000 befragten Psychologen anonym bekannt hat (John et al., 2012), gab es auch Fälle von explizitem Betrug, beispielsweise durch den Sozialpsychologen Diederik Stapel (Chambers, 2017; Pashler & Wagenmakers, 2012). Stapel hat Daten für mindestens 55 Publikationen frei erfunden (Levelt, Noort & Drenth, 2012, S. 25) und bis dato 58 Publikationen zurückgezogen (Palus, 2015). Auch der Sozialpsychologe Dirk Smeesters gab zu, eine „Datenmassage“ durchgeführt zu haben, um statistische Signifikanz zu erzielen. Dazu habe er die Datensätze jener Teilnehmer aussortiert, welche die Anleitung zu seinem Experiment nicht gründlich gelesen hätten. Ein solches Vorgehen sei Smeesters zufolge „üblich“ und damit „Teil einer Kultur“ (van der Heijden, Groenen, Zeelenberg & te Lindert, 2014, zitiert nach Chambers, 2017, S. 107). Zudem sei er durch den Publikationsbias motiviert gewesen (Chambers, 2017, S. 108).

Gegen den deutschen Sozialpsychologen Jens Förster wurden ebenfalls Betrugsvorwürfe laut. Es lässt sich zwar nicht belegen, dass die Daten bewusst manipuliert wurden; Försters Forschungsergebnisse sind jedoch mit einer Wahrscheinlichkeit von „1 in 508 Trillionen“ (Schuyt & van Arkel, 2012, S. 1) „statistisch quasi unmöglich“ (Kara, 2015, o. S.). So fehlen beispielsweise von keinen der 2.284 Versuchspersonen

Daten und es gab keine Abbrüche, was für psychologische Experimente „atypisch“ ist (Schuyt & van Arkel, 2012, S. 1). Das *Dutch National Board of Research Integrity* befand, dass es Manipulationen gab, die nicht allein durch fragwürdige Forschungspraktiken zu erklären sind (Chambers, 2017, S. 108 ff.; Schuyt & van Arkel, 2014).

Der Anteil der Zurücknahmen aufgrund von Betrugs hat sich seit dem Jahr 1975 um das Zehnfache erhöht (Fang, Steen & Casadevall, 2012). Dabei entfallen rund 67 % auf wissenschaftliches Fehlverhalten (davon wiederum 43 % auf Betrug und vermuteten Betrug), 14 % auf redundante Publikationen und 10 % auf Plagiate. Nelson, Simmons und Simonsohn (2018) teilen eine entsprechend negative Sicht auf die Dinge: Während viele glaubten, Betrug komme in den Sozialwissenschaften so selten vor wie der Einschlag eines Asteroiden, zeichneten die bekannt gewordenen Fälle ein anderes Bild. Offensichtlicher Betrug würde meist bereits von aufmerksamen Mitautoren, Gutachtern oder Herausgebern bemerkt und gelange meist nicht ans Licht der Öffentlichkeit. Komme es trotz dieser Schutzmaßnahmen zu einer Veröffentlichung betrügerischer Ergebnisse, würden nur wenige versierte Leser ihn erkennen. Von diesen wiederum würden nur sehr wenige Leser ihre Bedenken äußern. Noch weniger würden ihre Bedenken öffentlich machen, da dies auch Risiken für die eigene Reputation berge und einen großen Zeitaufwand ohne konkreten Nutzen bedeute. Dementsprechend sei anzunehmen, dass die Dunkelziffer betrügerischer Arbeiten deutlich größer ist als die Anzahl der bekannt gewordenen Fälle (Nelson et al., 2018; Simonsohn, 2013; siehe auch Simonsohn, Simmons & Nelson, 2015, S. 1148).

Replikationsdilemma

Im Bestreben, positive Effekte und Neuheiten veröffentlichen zu können, werden oftmals konzeptuelle – das heißt an bisher erforschte Theorien und Konzepte lediglich angelehnte – Replikationen durchgeführt. Konzeptuelle Replikationen eröffnen jedoch oftmals Sackgassen, die durch direkte Replikationen der vorausgehenden Forschung bereits hätten geschlossen werden können (Chambers, 2017, S. 35). Direkte Replikationen haben es jedoch schwer, da sie als zeitintensiv angesehen werden, Energie und Ressourcen von eigenen Ideen abziehen und schwerer zu publizieren sind. Selbst wenn sie publiziert würden, würden sie als unbedeutend angesehen und in der Folge weniger Anerkennung und Forschungsgelder einbringen (Everett & Earp, 2015; Earp & Trafimow, 2015). Nur etwa 1 % aller veröffentlichten psychologischen Studien sind direkte

Replikationen, von denen noch nicht einmal alle durch unabhängige Forscher durchgeführt wurden (Chambers, 2017, S. 50; Makel, Plucker & Hegarty, 2012). Dabei hat die Psychologie keinen Grund, sich zurückzuhalten, da die Mehrheit der durchgeführten Replikationen erfolgreich war. Allerdings verlaufen Replikationen, die von unabhängigen Forschern durchgeführt werden, seltener erfolgreich (Makel et al., 2012).

Daniel Kahneman vertritt die Auffassung, dass Replikationen nur in Kooperation mit den Erstautoren erfolgen dürften, da u. a. die in den Publikationen beschriebenen Methoden regelmäßig zu vage formuliert seien (Chambers, 2017, S. 52 f.). Statt Replikationen auf diese Weise zu erschweren, wäre es sicherlich sinnvoller, die Methoden möglichst exakt zu beschreiben. Replikationen sollten zudem nicht als Zweifel an der Fähigkeit eines Wissenschaftlers verstanden werden, sondern als Wertschätzung, da dessen Forschung als interessant und relevant genug betrachtet wird, meint Jon Butterworth (Chambers, 2017, S. 53). Die Lösung sieht Chambers (2017, S. 49) in einer Kultur, in der direkte Replikationen wertgeschätzt und dementsprechend auch durchgeführt werden: Nach erfolgreicher Replikation sollten die den Ursprungsarbeiten zugrunde liegenden Konzepte inkrementell erweitert und in kleinen Schritten getestet werden, um sowohl wissenschaftlichen Fortschritt zu gewährleisten, als auch genauere Erkenntnisse darüber zu fördern, welche Variablen tatsächlich wirken und welche nicht.

Ungenaue Methodenbeschreibungen

In veröffentlichten Studien erfüllt die Methodensektion den Zweck, die Untersuchungsbedingungen so eindeutig und detailliert wie möglich zu beschreiben, damit andere Forscher die ursprünglichen Untersuchungsbedingungen reproduzieren und die durchgeführten Analysen wiederholen können (Chambers, 2017, S. 61). Basierend auf der Beschreibung der Methodik muss ein kritischer Leser außerdem die Frage beantworten können, ob das Forschungsdesign valide ist. Bei missglückten Replikationen kann es ansonsten passieren, dass nachträgliche Aufklärungsversuche als unaufrichtig oder als Versuch zur Gesichtswahrung verstanden werden (Chambers, 2017, S. 62).

In der Studie von John et al. (2012) gaben mehr als 70 % der befragten Psychologen an, nicht alle Untersuchungsbedingungen offen beschrieben zu haben, und fast 100 % haben nicht alle durchgeführten Messungen offengelegt. Chambers (2017, S. 62) zufolge ist ein Teufelskreis entstanden, nach dem vage Methodenteile damit begründet werden, dass kaum Replikationen durchgeführt werden und umfangreiche „Replikations-

rezepte“ darum unnötig seien. Gleichzeitig werden Replikationen durch vage Methodenteile erschwert bis unmöglich gemacht. Mag dies in gedruckten Zeitschriften noch durch Platzmangel begründet gewesen sein, könnte heutzutage problemlos auf umfangreichere Informationen im Internet verwiesen werden.

Ursachen für die Replikationskrise

Begley und Ioannidis (2015) sehen fünf Hauptgründe für die Replikationskrise: (1.) Daten und Publikationen werden in einer nie da gewesenen Geschwindigkeit generiert. (2.) Dabei ist die Mehrzahl der Entdeckungen nicht von Dauer. (3.) Dies liegt insbesondere daran, dass gute wissenschaftliche Praxis nicht mehr eingehalten wird, da man aus Verzweiflung lieber möglichst viel publiziert. (4.) Dafür ist niemand alleinverantwortlich, weshalb es keine einfache Lösung gibt. (5.) Es handelt sich um ein vielschichtiges Problem verschiedener Interessengruppen.

Das „Reproducibility Project“

Aufgrund der Replikationskrise unternahmen Nosek et al. (Open Science Collaboration, 2015) im Rahmen ihres *Reproducibility Projects* in einer Gruppe von 270 Wissenschaftlern den Versuch, 100 psychologische Studien zu replizieren, die im Jahr 2008 in den Fachzeitschriften *Psychological Science*, *Journal of Personality and Social Psychology*, sowie dem *Journal of Experimental Psychology: Learning, Memory, and Cognition* veröffentlicht wurden. Dafür zogen sie die Autoren der Originalstudien hinzu. 97 Studien berichteten signifikante Ergebnisse, aber lediglich 35 Studien konnten definitiv erfolgreich repliziert werden. Die Replikationen wiesen dabei meist geringere Effekte auf als die Originalstudien (Open Science Collaboration, 2015). Je kleiner die ursprünglich berichteten *p*-Werte ausfielen, desto wahrscheinlicher war eine erfolgreiche Replikation. Die Autoren führen dies zurück auf den Publikationsbias, auf Versuchsdesigns mit geringer Teststärke und auf Herausgeber von Fachzeitschriften, die nur positive, neuartige Erkenntnisse veröffentlichen möchten. Die Autoren weisen jedoch darauf hin, dass dieses Problem nicht nur die Psychologie betrifft, sondern die gesamte Wissenschaft umfasst: Innovative Forschung zeige mögliche Wege auf. Replizierende Forschung zeige auf, welche Wege wahrscheinlich „richtig“ sind. Fortschritt beruhe auf beidem (Open Science Collaboration, 2015; Weir, 2015).

Selbst sehr bekannte Befunde aus der Sozialpsychologie, die längst Eingang in die Lehrbücher gefunden hatten (Schimmack, 2018c), erwiesen sich als häufig nicht replizierbar. Dazu zählen zum Beispiel *ego-depletion* (Hagger et al., 2016; Schimmack, 2016; Erwiderung bei Dang, Liu, Liu & Mao, 2017; Überblick bei Friese, Loschelder, Gieseler, Frankenbach & Inzlicht, 2018), *social priming* (Doyen, Klein, Pichon & Cleeremans, 2012), *facial feedback* (Wagenmakers et al., 2016), *power posing* (Simmons & Simonsohn, 2017; Erwiderung bei Cuddy, Schultz & Fosse, 2018) und andere (vgl. Jarrett, 2016; Schimmack, 2018b, 2019a, 2019b, 2019c). Demgegenüber konnten neun von neun grundlegende Ergebnisse der Allgemeinen Psychologie (je drei aus den Bereichen Wahrnehmung, Gedächtnis und Sprache) von Zwaan et al. (2017) repliziert werden. Dementsprechend werden Ergebnisse aus der kognitiven Psychologie typischerweise als „vergleichsweise robust“ angesehen (Open Science Collaboration, 2015, zitiert nach Zwaan et al., 2017, S. 1969). Es gibt allerdings auch prominente Ergebnisse aus der kognitiven Psychologie, die nicht repliziert werden konnten – zum Beispiel *retrieval-induced forgetting* (Maslany & Campbell, 2013) und *eye movements on recall* (Matzke et al., 2015). Entscheidend ist: Während die Quote erfolgreicher Replikationen bei Studien aus der Allgemeinen Psychologie 48 bzw. 53 % betrug, lag die Replikationsquote sozialpsychologischer Studien nur bei 23 bzw. 29 % (Open Science Collaboration, 2015, S. 3). Aus diesem zentralen Befund speisen sich die Hypothesen dieser Studie.

Problematischer Untersuchungsgegenstand

Forschungsgegenstand der Sozialpsychologie sind komplexe soziale Variablen, die kulturell bestimmt sind und sich kontextabhängig und im Laufe der Zeit ändern können (Cesario, 2014, S. 45; Schramm, 2013). Dementsprechend muss die Sozialpsychologie mehr als andere Disziplinen auf das Erschließen setzen und kann weniger auf direkte Messungen bauen. Bereits Comte (1830) vermutete eine „Hierarchie der Wissenschaften“ – von den „harten“ Wissenschaften der Physik mit vermeintlich bestmöglicher Messbarkeit, größter Präzision der Messungen und umfassendster Generalisierbarkeit der Ergebnisse über die Biologie „irgendwo dazwischen“ hin zu den „weichen“ Sozialwissenschaften (zur Unterscheidung „hart“ und „weich“ siehe z. B. Fanelli, 2010; Russell, 1929; Storer, 1967; Whewell, 1840). Für Comte (1830) stellt die Soziologie dabei die komplexeste, aber auch die am wenigsten fortgeschrittene Wissenschaft dar (Fanelli, 2010). Als Teilgebiet der Psychologie und der Soziologie ist hier auch

die Sozialpsychologie einzuordnen. Da Menschen aufgrund unterschiedlicher Sozialisation verschiedene Dinge mit anderen Menschen und Situationen assoziieren, ist der Einfluss der Versuchspersonen auf das Experiment größer. Dementsprechend vergrößert sich auch die Varianz (Brookshire, 2018; Reuning & Meyer, 2019). Nichtsdestotrotz sollten tatsächlich vorhandene Effekte repliziert und Kontextabhängigkeiten nicht überbewertet werden (Schönbrodt, zitiert in Reuning & Meyer, 2019): Die *Many-Labs-2*-Studie (Klein et al., 2018) ergab, dass es zwischen sechzig Laboren, die alle dieselbe Studie durchgeführt hatten, kaum Variabilität in den Effekten gab. Das spricht Schönbrodt zufolge dafür, dass es die Effekte, die tatsächlich existieren, auch „in jedem Labor, in jedem Land, in jeder Kultur“ gibt und zwar „mit kaum Varianz“. „Die Effekte, die es nicht gab, die gab es auch nirgends“. Dementsprechend sei es für die Psychologie „durchaus beruhigend“, zu wissen, dass psychologische Effekte nicht „so labil sind, dass sie mal da sind und mal nicht“ und „bei kleinsten Kontextänderungen schon zusammenbrechen“ würden (Schönbrodt, zitiert in Reuning & Meyer, 2019, o. S.).

Kumulative Evidenz

Nosek et al. (Open Science Collaboration, 2015, S. 7) plädieren dafür, anzuerkennen, dass keine einzige Studie einen endgültigen Beweis für die Existenz oder Nichtexistenz eines Effektes liefern kann (vgl. Jussim, 2016). Man kann lediglich gemeinsam mit anderen Wissenschaftlern so viele Wiederholungen durchführen, dass die Existenz eines Effekts mit einer gewissen Wahrscheinlichkeit angenommen werden kann. Dies gilt umso mehr, wenn es keine direkten, sondern nur konzeptuelle Replikationen gibt. Die durchgeführten Replikationen können die getroffenen Annahmen entweder untermauern, oder aufzeigen, wo weitere Forschung nötig ist. Wissenschaft kann aber keine Sicherheit schaffen, sondern lediglich Unsicherheit reduzieren, indem möglichst viele Indizienbeweise kumuliert werden. Replikationen sind somit das Immunsystem der Wissenschaft (Chambers, 2017, S. 47). Aufgrund ihres komplexen Forschungsgegenstands dürfte die Psychologie besonders stark auf kumulative Evidenz angewiesen sein. Die Wissenschaft muss ihr größter Skeptiker bleiben und die Sicherheit ihrer Methoden kontinuierlich anzweifeln – im Bestreben, die eigenen Methoden stetig zu verbessern, um die Unsicherheit zu reduzieren, die der Wissenschaft immanent ist (Johnson, 2013; Open Science Collaboration, 2015).

Zu geringe Teststärken

Psychologische Studien weisen im Durchschnitt eine Teststärke von 1-β = .50 auf, um einen mittleren Effekt zu finden, womit die Teststärke nach wie vor viel zu gering ausgeprägt ist (Bezeau & Graves, 2001; Chambers, 2017, S. 56; Cohen, 1962; Sedlmeier & Gigerenzer, 1989). Eine Teststärke von 1-β = .50 bedeutet, dass mit einer Wahrscheinlichkeit von 50 % die Nullhypothese irrtümlicherweise nicht abgelehnt wird, obwohl in Wahrheit ein Effekt existiert – womit die Wahrscheinlichkeit, einen tatsächlich existierenden Effekt zu finden, ebenfalls nur 50 % beträgt. Cohen (1965, zitiert nach Chase & Chase, 1976) empfiehlt eine Teststärke von 1-β = .80. An der Forschungspraxis hat sich in den vergangenen vierzig Jahren wenig geändert. In der Folge werden Effekte, die tatsächlich existieren, nicht veröffentlicht – was wiederum der weiteren Generierung von Theorien schadet (Chambers, 2017, S. 56 ff.).

Geringe Teststärken stellen die andere Seite des Publikationsbias dar: Zum einen werden nicht-existente Effekte als signifikant und damit als gegeben betrachtet. Zum anderen werden tatsächlich existierende Effekte häufig nicht gefunden, weil die Teststärken zu gering sind. Der Glaubwürdigkeit der Psychologie ist damit doppelt geschadet (Chambers, 2017, S. 56 ff.). Häufig wird die Größe der benötigten Stichprobe nicht durch eine *a priori*-Analyse der Teststärke bestimmt, sondern angelehnt an Studien, die – mit möglichst geringer Stichprobe – signifikante Ergebnisse erzielten (Chambers, 2017, S. 57). Dabei sorgt eine hohe Teststärke dafür, sowohl falsch-negative als auch falsch-positive Ergebnisse zu reduzieren (Chambers, 2017, S. 59 f.). Eine Reihe von Simulationen ergaben, dass die Rate falsch-negativer Ergebnisse – bei typischen Effekt- und Stichprobengrößen – in der psychologischen Forschung bei etwa 84 % liegt. Forscher schenken der Teststärke demnach zu ihrem eigenen Nachteil zu wenig Beachtung (Smith, Levine, Lachlan & Fediuk, 2002).

Ausbleibende Zurücknahmen

In der Psychologie werden vielfach weniger Artikel zurückgezogen als in anderen Forschungsgebieten (27 %, berechnet vom Mittelwert aus über 200 Forschungsgebieten) (Chambers, 2017, S. 66, 231 f.; Grieneisen & Zhang, 2012; Zhang & Grieneisen, 2013): Während eine gescheiterte Replikation in der Physik ein Grund zur Sorge sei, zeige sich die Psychologie davon eher unberührt. Chambers sieht das als Indiz dafür, dass die

Psychologie ihren eigenen Methoden weniger vertraut und darum eine tiefere Messlatte zur Publikation anlegt. Wenn Replikationen das Immunsystem der Wissenschaft sind, dann sind Zurücknahmen der letzte Schutzwall, vergleichbar mit einem chirurgischen Eingriff (Chambers, 2017). Wenn die Messlatte für Publikationen – bedingt durch geringe statistische Standards und fragwürdige Forschungspraktiken – allerdings tief ist, dann wird automatisch die Messlatte für eine Zurückziehung von Artikeln erhöht (Chambers, 2017, S. 65 ff.).

Daten werden nicht geteilt

In der Psychologie ist eine bemerkenswerte Verschlossenheit hinsichtlich der Veröffentlichung oder wenigstens des Teilens von Daten mit anderen Wissenschaftlern nachweisbar (Chambers, 2017, S. 78 ff.): So haben Wicherts, Borsboom, Kats und Molenaar (2006) über 400 E-Mails mit der Bitte um Zusendung von Rohdaten versendet. 73 % der Autoren gaben keine Daten heraus, obwohl die Richtlinien der *American Psychological Association (APA)* die Herausgabe verlangen (APA, 2002). Von den 73 % haben 18 % nicht geantwortet, 35 % verweigerten die Herausgabe und 20 % versprachen die Herausgabe, hielten ihr Versprechen aber nicht ein. Gründe für diese mangelnde Herausgabebereitschaft könnten Chambers (2017, S. 79) zufolge sein, dass fragwürdige oder zumindest nachlässige Forschungspraktiken eine Rolle spielten, die keiner öffentlichen Debatte zugeführt werden sollen. Insbesondere „statistische Fehler“, die Studienergebnisse knapp unter die Signifikanzschwelle bringen, dürften in Arbeiten, deren Rohdaten offenbar geheim gehalten werden sollen, häufiger vorkommen. Entsprechend fanden Wicherts, Bakker und Molenaar (2011) heraus, dass *p*-Werte von Autoren, die ihre Rohdaten nicht geteilt haben, mehr als doppelt so häufig falsch berichtet wurden. Obwohl einige der Autoren gegen den Verhaltenskodex der *APA* verstießen, blieb eine Rüge der *APA* oder eine Zurücknahme der veröffentlichten Artikel aus (Chambers, 2017, S. 80). Rohdatenanalysen führten in den Fällen von Lawrence Sanna und Dirk Smeesters zu Rücktritten aufgrund fingierter Daten (Simonsohn, 2013).

Hypothesen

Aufgrund des ausführlich dargestellten theoretischen Hintergrundes der Replikationskrise (zusammengefasst bei Pashler & Wagenmakers, 2012) und dem mit ihr in Zusammenhang stehenden Einsatz fragwürdiger Forschungspraktiken werden in dieser Studie folgende Annahmen getroffen: Fachgebiete mit überwiegend nicht-manipulierten Forschungsergebnissen berichten häufiger wahre oder eindeutigere Effekte und erzielen kleinere *p*-Werte. Deren *p*-Kurve fällt somit linkssteil aus, da sie mehr kleinere (z. B. .01) als größere (z. B. .05) *p*-Werte enthält. Fachgebiete mit überwiegend manipulierten Forschungsergebnissen (in denen *p-Hacking* die Norm ist) berichten häufiger größere *p*-Werte, die sich knapp unter der Signifikanzschwelle von $p = .05$ befinden. Deren *p*-Kurve fällt somit rechtssteil aus, da sie mehr größere (z. B. .05) als kleinere (z. B. .01) *p*-Werte enthält. In dieser Studie werden dementsprechend folgende Hypothesen aufgestellt und empirisch überprüft:

- Hypothese A: Die Verteilung der *p*-Werte in der Allgemeinen Psychologie ist linkssteil ausgeprägt, da Studien aus diesem Fachgebiet häufiger repliziert werden konnten als jene aus der Sozialpsychologie.
- Hypothese B: Die Verteilung der *p*-Werte in der Sozialpsychologie ist rechtssteil ausgeprägt, da Studien aus diesem Fachgebiet seltener repliziert werden konnten als jene aus der Allgemeinen Psychologie.
- Hypothese C: Die Verteilungen der *p*-Werte unterscheiden sich signifikant zwischen den Fachgebieten, da die Sozialpsychologie von der Replikationskrise stärker betroffen war als die Allgemeine Psychologie. Dementsprechend ist die Verteilung der *p*-Werte in der Sozialpsychologie rechtssteil und die Verteilung der *p*-Werte in der Allgemeinen Psychologie linkssteil ausgeprägt.

Methode

Forschungsdesign

Es wird eine *p*-Kurven-Analyse nach Simonsohn et al. (2014a) durchgeführt, die in diesem Forschungsdesign adaptiert verwendet wird. Klassischerweise verwendet man *p*-Kurven-Analysen, um die Verteilung von *p*-Werten zu einem bestimmten Forschungsthema abzubilden. Beispielsweise könnte sich ein Leser fragen, ob die berichteten Effekte zu einem psychologischen Phänomen tatsächlich echt sind oder allesamt dem Publikationsbias unterliegen, oder gar durch gezieltes *p-Hacking* zustande kamen. In der vorliegenden Studie wird das Prinzip der *p*-Kurven-Analyse auf psychologische Fachgebiete übertragen. Dabei werden unterschiedliche Thematiken, Konstrukte und Effekte innerhalb der Fachgebiete stichprobenartig erhoben und dann verglichen, inwiefern die Fachgebiete in ihren berichteten Ergebnissen voneinander abweichen.

Die *p*-Kurven-Analyse unterscheidet linkssteile Verteilungen (vgl. *Abbildung 1*), rechtssteile Verteilungen (vgl. *Abbildung 2*) und Gleichverteilungen. Eine linkssteile Verteilung der *p*-Werte eines Fachgebiets ist ein Hinweis darauf, dass es sich um legitime Forschungsergebnisse handelt, die empirische Evidenz besitzen. Bei einer rechtssteilen Verteilung der *p*-Werte eines Fachgebiets wird dagegen angenommen, dass die Verteilung ein Hinweis auf mögliches *p-Hacking* ist, mit dessen Hilfe die *p*-Werte unter die Signifikanzschwelle von $p = .05$ gebracht wurden, weshalb diese sich überwiegend nah an dieser Schwelle befinden (mehr *p*-Werte um $p = .04$ als um $p = .01$). So genannte „flache *p*-Kurven“, die fast gleich viele *p*-Werte um $p = .01$ wie um $p = .04$ aufweisen, sind ein Hinweis darauf, dass ein untersuchter Effekt in der Realität nicht existiert. Die zugrundeliegenden Studien weisen demnach weder empirische Evidenz auf, noch ist anzunehmen, dass sie durch *p-Hacking* zustande kamen. So genannte „marginale“ *p*-Werte wie $p = .051$ bleiben bei der *p*-Kurven-Analyse unberücksichtigt, da diese sehr selten vorkämen, sollte ein Effekt tatsächlich existieren (Simonsohn et al., 2014a).

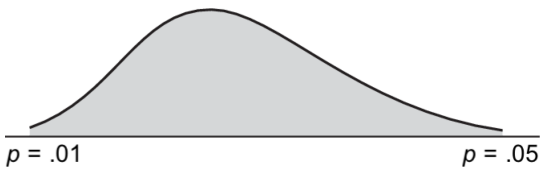

Abbildung 1. Darstellung einer linkssteilen Verteilung.

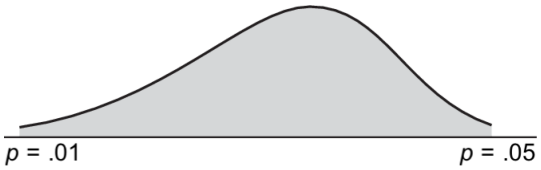

Abbildung 2. Darstellung einer rechtssteilen Verteilung.

Voraussetzungen der *p*-Kurven-Analyse

Nach Simonsohn et al. (2014a) müssen Studien und *p*-Werte in geeigneter Weise ausgewählt werden, um Schlussfolgerungen aus der *p*-Kurven-Analyse ziehen zu können. Dabei darf die *p*-Kurven-Analyse ausschließlich *p*-Werte heranziehen, die statistisch unabhängig voneinander sind und die die zu ihr gehörenden Haupthypothesen unmittelbar testen. Zudem müssen die *p*-Werte unter der Nullhypothese stetig gleichverteilt sein. Häufig gibt es mehr als einen signifikanten *p*-Wert pro Studie (Bishop & Thompson, 2016). Da die *p*-Kurven-Analyse es erfordert, dass die *p*-Werte unabhängig voneinander sind kann nur *ein* signifikanter *p*-Wert pro Studie „zufällig" ausgewählt werden (Bishop & Thompson, 2016, S. 5). In dieser Studie ist dies der erstgenannte signifikante *p*-Wert, der die Haupthypothese unmittelbar testet. Eine Bildung von Mittelwerten basierend auf Teststatistiken oder *p*-Werten ist nicht zulässig (Simonsohn, Nelson & Simmons, 2014b, S. 676). Studien ohne signifikanten *p*-Wert bleiben unberücksichtigt. Sofern mehr als ein *p*-Wert die angegebene Hypothese unmittelbar testet und die *p*-Werte abhängig voneinander sind, muss ein Test auf Robustheit durchgeführt werden. Dabei wird die *p*-Kurven-Analyse zuerst mit dem erstgenannten, anschließend mit dem letztgenannten *p*-Wert durchgeführt. Ergibt sich keine signifikante Veränderung, wird der erstgenannte *p*-Wert in die Analyse einbezogen. Verändert sich die *p*-Kurve jedoch signifikant, dann bleiben die entsprechenden *p*-Werte aufgrund ihres verfälschenden Potentials unberücksichtigt.

Vorgehen bei der *p*-Kurven-Analyse

Bei der Durchführung einer *p*-Kurven-Analyse wird zuerst eine sinnvolle Selektionsregel aufgestellt, welche Studien ausgewählt und welche *p*-Werte analysiert werden sollen. Anschließend werden die einzubeziehenden Studien sowie deren Autoren, Titel, Veröffentlichungsdaten und Internetadressen in eine so genannte *p*-Wert-Tabelle eingetragen. Darin werden auch die Haupthypothesen, Studiendesigns, Arten der statistischen Ergebnisse, die statistischen Ergebnisse in Form von Zitaten sowie deren *t-, F-, r-, Z-* und χ^2-Werte, Freiheitsgrade und dazugehörige *p*-Werte dokumentiert. Wenn sinnvoll und notwendig, werden zusätzlich alternative *p*-Werte angegeben, um einen Test auf Robustheit der *p*-Kurve durchführen zu können. Die Ergebnisse der *p*-Kurven-Analyse gelten dann als robust, wenn sich im Rahmen des Tests auf Robustheit keine signifikanten Unterschiede zu den Ergebnissen des zuerst durchgeführten Tests ergeben.

Grundsätzlich werden die *p*-Werte anhand der in den einbezogenen Studien berichteten *t-, F-, r-, Z-* und χ^2-Werte sowie der korrespondierenden Freiheitsgrade exakt nachberechnet. Die berichteten *p*-Werte werden lediglich aufgenommen, um bei Abweichungen von den nachberechneten Ergebnissen Hinweise darauf finden zu können, ob *p*-Werte abgeschnitten, unzulässig abgerundet oder gar vorsätzlich manipuliert wurden. Nach der Erstellung der *p*-Wert-Tabelle werden die Ergebnisse basierend auf dem *R*-Skript von Simonsohn (2017b) in Version 4.06 ausgewertet. Dabei wird eine graphische Ausgabe der *p*-Kurve sowie eine Tabelle mit statistischen Analysen erzeugt.

Methodik der *p*-Kurven-Analyse

Die *p*-Kurven-Analyse nach Simonsohn et al. (2014a) kombiniert sowohl die halbe als auch die volle *p*-Kurve, um Schlussfolgerungen über deren empirische Aussagekraft zu treffen. Wenn sowohl die halbe *p*-Kurve mit $p < .05$ linkssteil ist, oder sowohl die halbe als auch die volle *p*-Kurve mit $p < .10$ linkssteil ist, dann nimmt die *p*-Kurven-Analyse an, dass die *p*-Kurve empirische Evidenz besitzt. Dieser kombinierte Test ist robuster gegenüber tiefgreifendem *p-Hacking* als der einfache *p*-Kurven-Test und damit konservativer. Das bedeutet, dass er Linkssteilheit – und damit empirische Evidenz – weniger schnell als gegeben einschätzt und tiefgreifendem *p-Hacking* auch bei einer großen Anzahl von *p*-Werten widersteht (Simonsohn et al., 2015, S. 1149 ff.). Zwar werden bei der ausschließlichen Betrachtung von *p*-Werten < .025 nur die ohnehin aussagekräftigeren *p*-Werte beurteilt, bei denen *p-Hacking* weniger wahrscheinlich ist; durch die reduzierte Anzahl an *p*-Werten sinkt jedoch aufgrund der kleineren Stichprobe auch die Teststärke, was wiederum die Irrtumswahrscheinlichkeit in dieser Hinsicht erhöht.

Die *p*-Kurven-Analyse nutzt stetige Tests, basierend auf Stouffers Methode (Simonsohn et al., 2015, S. 1149 ff.). Hierbei wird für jeden signifikanten *p*-Wert ein so genannter *pp*-Wert berechnet, der auf einem angenommenen Signifikanzniveau von $\alpha = .05$ die Wahrscheinlichkeit angibt, einen mindestens so großen *p*-Wert zu erhalten, wenn in Wahrheit die Nullhypothese gilt (ein *p*-Wert des *p*-Werts). Die *pp*-Werte werden anschließend in mit $N(0,1)$ normalverteilte *Z*-Werte konvertiert (Mittelwert 0, Standardabweichung 1). Die Summen dieser *Z*-Werte werden durch die Quadratwurzel der Anzahl der eingegebenen Teststatistiken dividiert. Das Ergebnis sind die in den *Tabellen 1* und *2* berichteten *Z*-Werte mit ihren dazugehörigen *p*-Werten. Zudem führt die *p*-Kurven-Analyse binominale Tests durch. Das bedeutet, dass sämtliche einbezogenen *p*-Werte

zuerst in kleine ($p < .025$) und große ($p > .025$) p-Werte kategorisiert werden. Anschließend wird der Anteil kleiner p-Werte mit dem erwarteten Anteil (50 %) verglichen, der vorläge, wenn kein Effekt (keinerlei empirische Evidenz) existieren würde. Dieser Test ist einfach und resistent gegenüber Extremwerten. Darüber hinaus wird der Anteil kleiner p-Werte mit dem erwarteten Anteil (ca. 70 %, abhängig von den Freiheitsgraden der einbezogenen Teststatistiken) verglichen, der vorläge, wenn alle Studien eine Teststärke von 33 % aufwiesen. Resultiert eine flachere Verteilung als zu erwarten wäre, wenn die zugrundeliegenden Studien eine durchschnittliche Teststärke von 33 % hätten, dann ist die empirische Evidenz als mangelhaft einzustufen. Die Schwelle von 33 % wird von Simmons und Simonsohn (2017, S. 692) zwar als willkürlich, aber begründet, beschrieben, da eine mit einer Teststärke von 33 % durchgeführte Studie in zwei von drei Fällen kein signifikantes Ergebnis erzielen würde (Simonsohn et al., 2014a, S. 538). Sind die Ergebnisse der p-Kurve nicht eindeutig, müssen weitere p-Werte einbezogen werden. Die durchschnittliche Teststärke (statistische *Power*) der zugrundeliegenden Studien wird berechnet, indem eine erwartete p-Kurve für jede mögliche Teststärke zwischen 5 % und 99 % erstellt wird, und diese wiederum mit der tatsächlich beobachteten p-Kurve verglichen wird. Anschließend wird die Teststärke der erwarteten p-Kurve ausgewählt, die der beobachteten p-Kurve am ähnlichsten ist. Studien mit größerer Teststärke erzielen *linkssteilere* p-Kurven (Simonsohn et al., 2014a, S. 536 f., 2014b; Simmons & Simonsohn, 2017, S. 689).

Ablauf und Selektionsregel

Zuerst wurde eine Liste deutscher Universitäten erstellt (siehe elektronischer Anhang), basierend auf einer Übersicht der *Deutschen Gesellschaft für Psychologie* (DGPs, 2019). Anschließend wurde in alphabetischer Reihenfolge recherchiert, welche Lehrstühle es an den entsprechenden Universitäten gibt. Um veröffentlichte Studien zu finden, wurde die Publikationsliste des jeweiligen Lehrstuhls herangezogen. Bei der Auswahl wurde nicht nach Erstautoren differenziert. In seltenen Fällen kann es deshalb vorgekommen sein, dass der Erstautor einer anderen Universität angehörte. Wenn keine gemeinsame Publikationsliste des Lehrstuhls vorhanden war, wurde die Publikationsliste des Lehrstuhlinhabers (Professors) herangezogen. Wenn auch eine solche nicht vorhanden war, wurde diejenige des alphabetisch erstgenannten wissenschaftlichen Mitarbeiters herangezogen. In umgekehrt chronologischer Reihenfolge wurde dann das erste zugäng-

liche Ergebnis ausgewertet. Wurde kein eindeutiger *p*-Wert gefunden, wurde die nächste Arbeit herangezogen. Dies kam insbesondere dann vor, wenn kein Zugriff möglich war, es sich um keine empirische Arbeit handelte, kein signifikantes Ergebnis vorlag oder sich aufgrund einer unklaren Haupthypothese kein repräsentativer *p*-Wert ausmachen ließ. Wenn die Auswertung von mindestens fünf Studien eines Lehrstuhls erfolglos blieb, dann wurde dieser nicht weiter berücksichtigt (*stopping rule*). Mögliche Gründe dafür waren: Kein Zugriff auf die Studie; keine empirisch-experimentelle Arbeit; Angabe von Teststatistiken, die im Rahmen der *p*-Kurven-Analyse nicht auswertbar sind (z. B. β-Gewichte); eine Aufbereitung von Ergebnissen, deren Auswertung mit unverhältnismäßig hohem Aufwand verbunden gewesen wäre.

Die ausgewerteten Studien wurden in die *p*-Wert-Tabelle eingetragen (siehe elektronischen Anhang). Anhand der in den Studien berichteten Teststatistiken wurden die *p*-Werte in *Microsoft Excel for Mac, Version 16.28* u. a. anhand der *TDIST*-, *FDIST*- und *CHISQ.DIST*-Funktionen exakt nachberechnet (siehe *p*-Wert-Tabelle im elektronischen Anhang). Diese Neuberechnung ist erforderlich, weil *p*-Werte häufig nur als „kleiner als"-Wert angegeben und manchmal nicht korrekt berichtet werden (Bakker & Wicherts, 2011). Die berichteten Teststatistiken wurden in die webbasierte Applikation zur Durchführung der *p*-Kurven-Analyse von Simonsohn (2017a) übertragen, um eine entsprechende graphische und statistische Auswertung zu erhalten. Die nachberechneten *p*-Werte wurden genutzt, um zu überprüfen, ob sich die Verteilungen der *p*-Werte zwischen den beiden Fachgebieten signifikant voneinander unterscheiden. Hierzu wurde in *R* (R Core Team, 2013) mittels der *ks.test*-Funktion (R Core Team, 2013) ein Kolmogorow-Smirnow-Test sowie mittels der *cramer.test*-Funktion (Franz, 2019) ein Cramér-Test durchgeführt, da letzterer insbesondere im Zwei-Stichproben-Fall als trennschärfer gilt.

Stichprobe

Aufgrund knapper zeitlicher sowie personeller Ressourcen wurde der Mindestumfang der Stichprobe *a priori* auf 20 *p*-Werte pro Fachgebiet festgelegt, da dies laut Simonsohn et al. (2014a, S. 544) ausreicht, um empirische Evidenz zu finden – selbst dann, wenn die Teststärke der zugrundeliegenden Studien lediglich 50 % betragen würde. Insgesamt wurden mehrere hundert Studien recherchiert, von denen 218 detailliert ausgewertet und dokumentiert wurden. Aus dem Fachgebiet der Allgemeinen Psychologie wurden 114 Studien ausgewertet und aus dem Fachgebiet der Sozialpsychologie 104 Stu-

dien. Aus 88 empirischen Arbeiten konnten *p*-Werte extrahiert werden (Allgemeine Psychologie: 51 *p*-Werte; Sozialpsychologie: 37 *p*-Werte). Letztlich wurden $N = 76$ Studien in die Auswertung einbezogen, deren *p*-Werte einer Haupthypothese eindeutig zuordenbar waren oder deren Ergebnis als repräsentativ für die Haupthypothese betrachtet wurde. Von den $N = 76$ Studien, die in die *p*-Kurven-Analyse einbezogen wurden, stammen $n = 39$ aus der Allgemeinen Psychologie und $n = 37$ aus der Sozialpsychologie. Die Aggregation der Daten wurde bis zur genannten Anzahl durchgeführt, da somit fast alle Lehrstühle mit je einem *p*-Wert aus ihrem Fachgebiet repräsentiert werden konnten.

Auswertung und Ergebnisse

Die Auswertung erfolgte mittels *R* (R Core Team, 2013), *RStudio* (RStudio, 2018) sowie der webbasierten Applikation für die *p*-Kurven-Analyse von Simonsohn (2017a). Die Skripte befinden sich im elektronischen Anhang. Die Abbildungen der *p*-Kurven wurden durch die webbasierte Applikation „*p*-checker“ (Schönbrodt, 2018) generiert. Ergänzend wurde eine *Z*-Kurven-Analyse (Brunner & Schimmack, 2018) durchgeführt. Die entsprechende Ausgabe wurde von der webbasierten Applikation „*Z*-Curve Shiny-App“ (Schimmack & Brunner, 2019) erzeugt.

Ergebnisse der *p*-Kurven-Analyse

Hypothese A: Allgemeine Psychologie. Wie die in *Abbildung 3* dargestellten *p*-Kurven sowie die in *Tabelle 1* berichteten Ergebnisse zeigen, ist die *p*-Kurve für das Fachgebiet der Allgemeinen Psychologie signifikant linkssteil ausgeprägt. Hypothese A, wonach die Verteilung der *p*-Werte in der Allgemeinen Psychologie linkssteil ausgeprägt sei, konnte somit bestätigt werden.

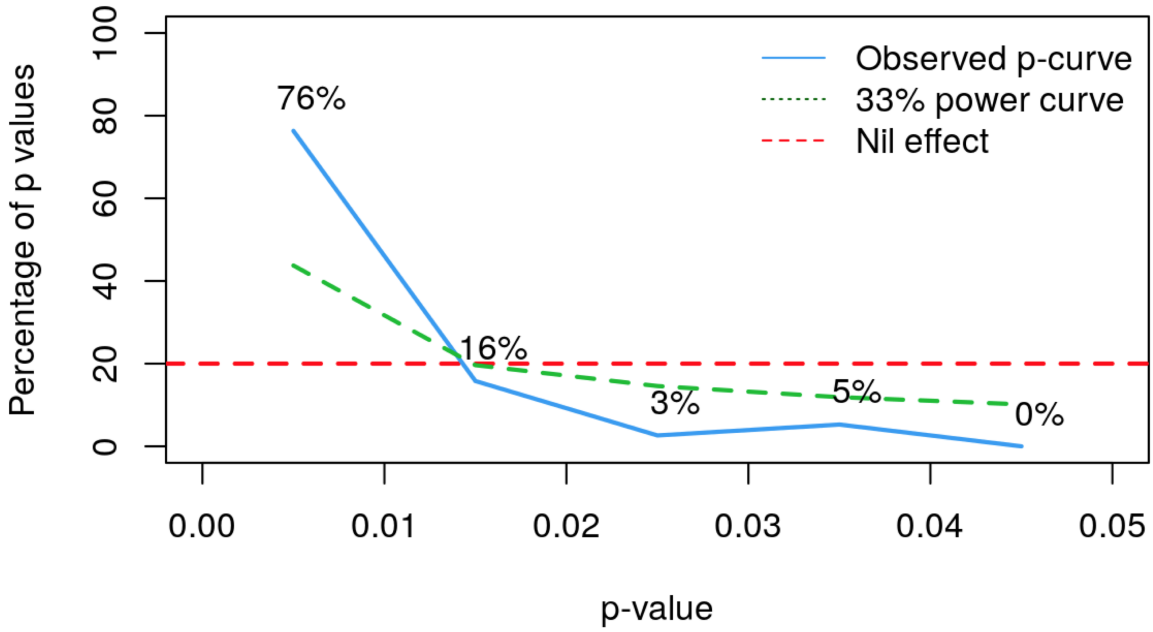

Abbildung 3. *p*-Kurve der Allgemeinen Psychologie (n = 39). Die vertikale Achse gibt die prozentuale Häufigkeit der *p*-Werte an, die auf der horizontale Achse abgebildet sind. 76 % der *p*-Werte sind $p < .01$. Die blaue Linie zeigt die linkssteile Verteilung der in die Analyse einbezogenen *p*-Werte. Die grüne Linie zeigt eine Verteilung, die zu erwarten wäre, wenn empirische Evidenz gegeben ist und die zugrundeliegenden Studien eine Teststärke von 33 % aufweisen. Die rote Linie zeigt eine Gleichverteilung der *p*-Werte, die angenommen wird, wenn es die in den Studien berichteten Effekte in Wahrheit nicht gäbe.

Nach Simonsohn et al. (2015, S. 1151) besitzt eine Reihe von Studien dann empirische Evidenz, wenn die halbe *p*-Kurve mit $p < .05$ linkssteil ist, oder sowohl die halbe als auch die volle *p*-Kurve mit $p < .10$ linkssteil ist. In diesem Fall sind beide Bedingungen erfüllt, was auf eine empirische Aussagekraft der *p*-Kurve bzw. empirische Evidenz der zugrundeliegenden Stichprobe hindeutet. Darüber hinaus testet die *p*-Kurven-Analyse, ob die *p*-Kurve signifikant flacher ausfällt als eine, deren zugrundeliegende Studien eine Teststärke von 33 % aufweisen. Erzielt dieser Test für die volle *p*-Kurve ein $p < .05$ oder für die halbe *p*-Kurve sowie den binominalen Powertest ein $p < .10$, dann wird ein Mangel an empirischer Evidenz angenommen. In diesem Fall ist die empirische Evidenz weder ausgeschlossen noch unzureichend.

Tabelle 1

Ergebnisse der p-Kurven-Analyse für das Fachgebiet der Allgemeinen Psychologie

	Binominaler Test (Anteil der Ergebnisse mit $p < .025$)	Stetiger Test (Aggregiert mittels Stouffers Methode)	
		Volle *p*-Kurve (*p*'s < .05)	Halbe *p*-Kurve (*p*'s < .025)
1) Studien enthalten empirische Evidenz (linkssteile Verteilung)	$p < .0001$	$Z = -21.98$, $p < .0001$	$Z = -20.60$, $p < .0001$
2) Studien sind von unzureichender empirischer Evidenz (flacher als 33 % Teststärke)	$p > .9999$	$Z = 15.82$, $p > .9999$	$Z = 18.76$, $p > .9999$
Teststärke der Tests, die in die *p*-Kurve mit einbezogen wurden (korrigiert für selektives Berichten)	Teststärke Schätzung: 99 % 90 % Konfidenzintervall: [99 %; 99 %]		

Anmerkung. Binominale und stetige Tests sind einseitig (H_1: „linkssteiler als flach"). Negative *Z*-Werte implizieren Linkssteilheit. (Test auf Abweichung in Richtung der Alternativhypothese – hier: Linkssteilheit – entgegen der Nullhypothese, dass die Verteilung der *p*-Kurve flach ist.)

Wie *Tabelle 1* zeigt, sind alle drei Kriterien für die Annahme empirischer Evidenz signifikant (alle $p < .001$) und deuten somit auf eine linkssteile Verteilung der zugrundeliegenden *p*-Werte hin. Darüber hinaus sind die drei Kriterien für die Abwesenheit empirischer Evidenz nicht signifikant (alle $p > .999$). Somit kann die empirische Evidenz der zugrundeliegenden Studien weder ausgeschlossen noch als unzureichend gewertet werden. Aufgrund des gewählten Konfidenzniveaus (*CI*) von 90 % kann angenommen werden, dass die Grenzen des Konfidenzintervalls in 90 % der Fälle den wahren Parameter umschließen. Dementsprechend wird die durchschnittliche Teststärke der zugrundeliegenden Studien auf 99 % geschätzt.

Extremwertanalyse. Eine Auswahl von Studien, deren empirische Evidenz noch immer eine gleichbleibend hohe Güte aufweist, nachdem man die kleinsten oder größten *p*-Werte aus der Analyse ausgeschlossen hat, ist robuster, die statistischen Ergebnisse vertrauenswürdiger und die empirische Aussagekraft somit größer. Für die Ergebnisse des Fachgebiets der Allgemeinen Psychologie ergaben sich keine die Aussagekraft beeinträchtigende Abweichungen gegenüber der vollständigen *p*-Kurve: Wie *Abbildung 4* zeigt, wäre die volle *p*-Kurve selbst dann signifikant linkssteil ausgeprägt, wenn bis zu 17 der kleinsten bzw. bis zu 17 der größten *p*-Werte aus der Analyse ausgeschlossen

worden wären (alle $p < .05$). Dies trifft ebenfalls für die halbe p-Kurve zu, die nur p-Werte kleiner als $p = .025$ berücksichtigt (alle $p < .05$). Darüber hinaus wäre die volle p-Kurve auch nach Ausschluss von bis zu 17 der kleinsten bzw. bis zu 17 der größten p-Werte nicht signifikant flacher ausgefallen, als eine, die auf Studien mit einer Teststärke von nur 33 % beruht (alle $p > .9$).

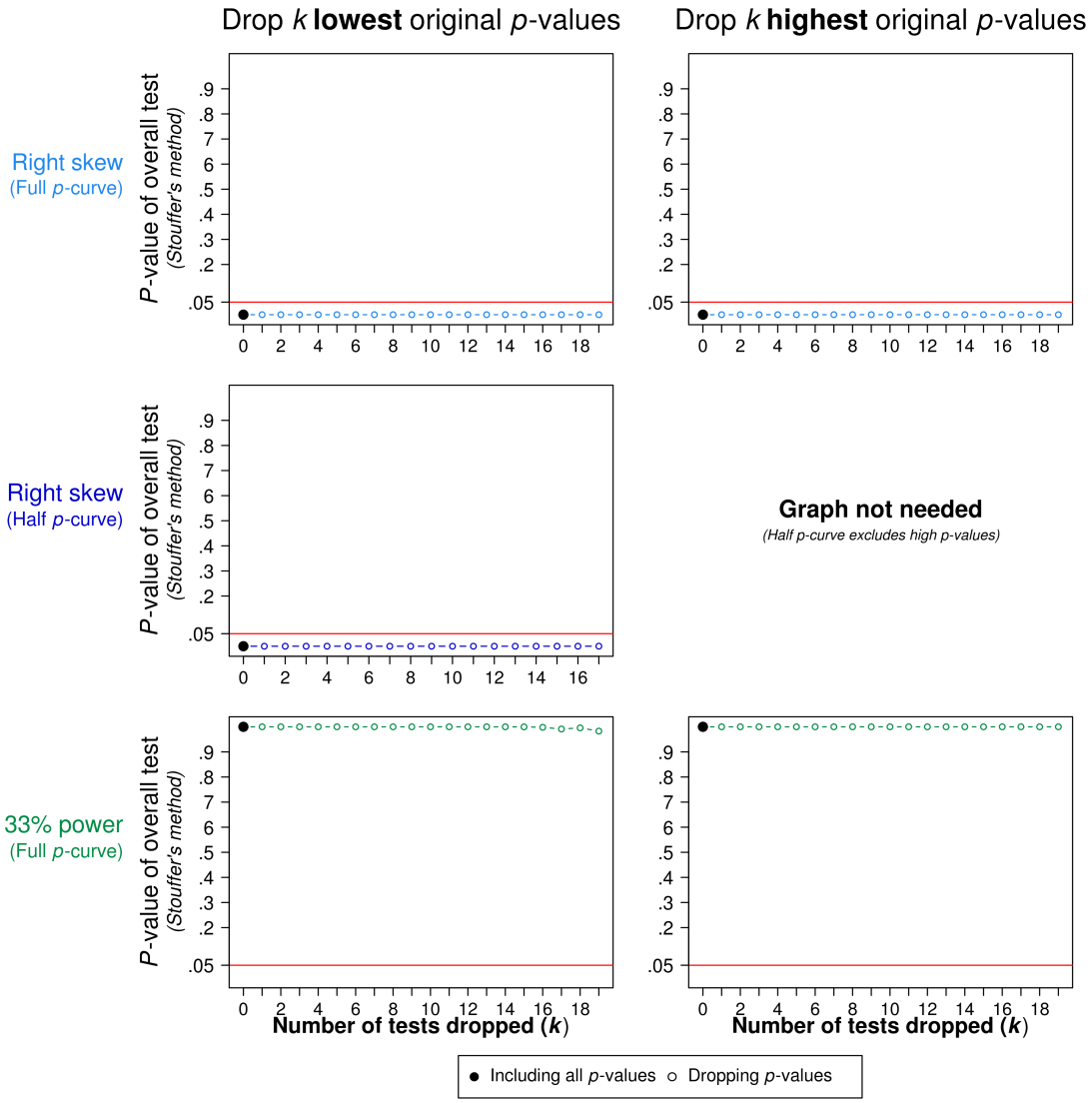

Abbildung 4. Bei den Tests auf Linkssteilheit hätten sich bei Ausschluss der kleinsten und größten p-Werte aus dem Fachgebiet der Allgemeinen Psychologie keine Auswirkungen auf das Ergebnis der p-Kurven-Analyse ergeben. Das obere Drittel der Abbildung zeigt die Auswirkungen auf die volle p-Kurve, das mittlere Drittel auf die halbe p-Kurve und das untere Drittel auf den Test auf Flachheit vs. 33 % Teststärke. Die horizontale Achse repräsentiert die Anzahl (k) ausgeschlossener p-Werte, die vertikale Achse den p-Wert der daraus resultierenden p-Kurve. Die rote Linie markiert die Signifikanzschwelle von $\alpha = .05$.

Hypothese B: Sozialpsychologie. Wie die in *Abbildung 5* dargestellten *p*-Kurven sowie die in *Tabelle 2* berichteten Ergebnisse zeigen, ist die *p*-Kurve für das Fachgebiet der Sozialpsychologie signifikant linkssteil ausgeprägt. Hypothese B, wonach die Verteilung der *p*-Werte in der Sozialpsychologie rechtssteil ausgeprägt gewesen wäre, konnte demnach nicht bestätigt werden.

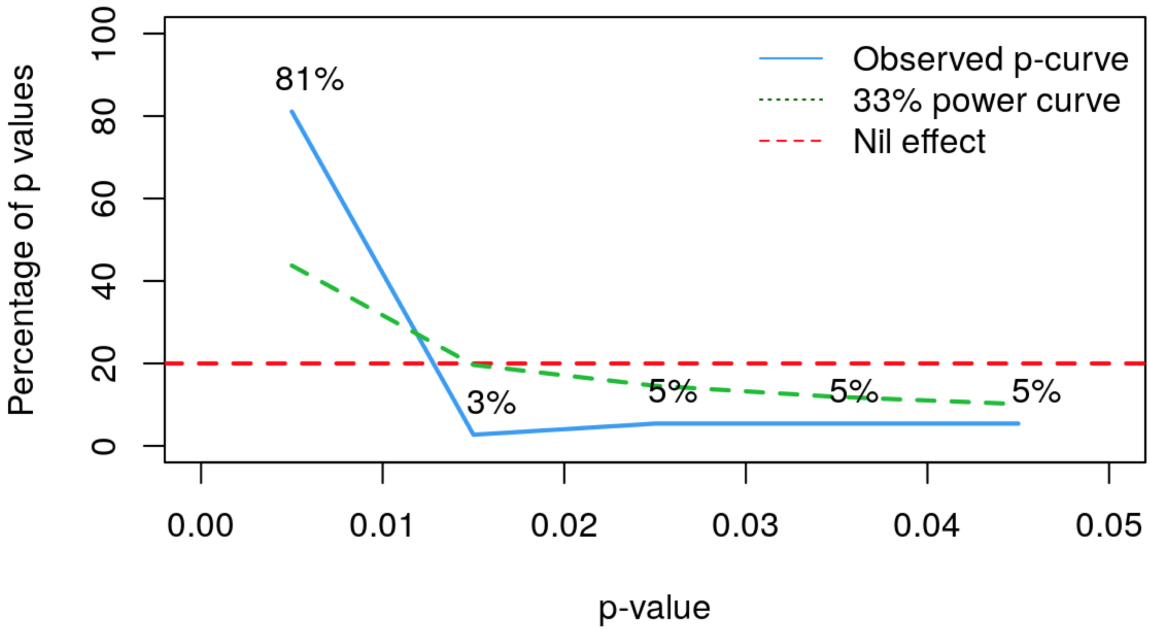

Abbildung 5. *p*-Kurve der Sozialpsychologie ($n = 37$). Die vertikale Achse gibt die prozentuale Häufigkeit der *p*-Werte an, die auf der horizontale Achse abgebildet sind. 81 % der *p*-Werte sind $p < .01$. Die blaue Linie zeigt die linkssteile Verteilung der in die Analyse einbezogenen *p*-Werte. Die grüne Linie zeigt eine Verteilung, die zu erwarten wäre, wenn empirische Evidenz gegeben ist und die zugrundeliegenden Studien eine Teststärke von 33 % aufweisen. Die rote Linie zeigt eine Gleichverteilung der *p*-Werte, die angenommen wird, wenn es die in den Studien berichteten Effekte in Wahrheit nicht gäbe.

Wie *Tabelle 2* zeigt, sind alle drei Kriterien für die Annahme empirischer Evidenz signifikant (alle $p < .001$) und deuten somit auf eine linkssteile Verteilung der zugrundeliegenden *p*-Werte hin. Darüber hinaus sind die drei Kriterien für die Abwesenheit empirischer Evidenz nicht signifikant (alle $p > .992$). Somit kann die empirische Evidenz der zugrundeliegenden Studien weder ausgeschlossen noch als unzureichend gewertet werden. Aufgrund des gewählten Konfidenzniveaus (*CI*) von 90 % kann angenommen werden, dass die Grenzen des Konfidenzintervalls in 90 % der Fälle den wahren Parameter umschließen. Dementsprechend wird die durchschnittliche Teststärke der zugrundeliegenden Studien auf 99 % geschätzt.

Tabelle 2

Ergebnisse der p-Kurven-Analyse für das Fachgebiet der Sozialpsychologie

	Binominaler Test (Anteil der Ergebnisse mit $p < .025$)	Stetiger Test (Aggregiert mittels Stouffers Methode)	
		Volle *p*-Kurve (*p*'s < .05)	Halbe *p*-Kurve (*p*'s < .025)
1) Studien enthalten empirische Evidenz (linkssteile Verteilung)	$p < .0001$	$Z = -19.98$, $p < .0001$	$Z = -20.83$, $p < .0001$
2) Studien sind von unzureichender empirischer Evidenz (flacher als 33 % Teststärke)	$p = .9924$	$Z = 13.94$, $p > .9999$	$Z = 18.17$, $p > .9999$
Teststärke der Tests, die in die *p*-Kurve mit einbezogen wurden (korrigiert für selektives Berichten)	Teststärke Schätzung: 99 % 90 % Konfidenzintervall: [99 %; 99 %]		

Anmerkung. Binominale und stetige Tests sind einseitig (H_1: „linkssteiler als flach"). Negative *Z*-Werte implizieren Linkssteilheit. (Test auf Abweichung in Richtung der Alternativhypothese – hier: Linkssteilheit – entgegen der Nullhypothese, dass die Verteilung der *p*-Kurve flach ist.)

Extremwertanalyse. Für die Ergebnisse des Fachgebiets der Sozialpsychologie ergaben sich keine die Aussagekraft beeinträchtigende Abweichungen gegenüber der vollständigen *p*-Kurve: Wie *Abbildung 6* zeigt, wäre die volle *p*-Kurve selbst dann signifikant linkssteil ausgeprägt, wenn bis zu 17 der kleinsten bzw. bis zu 17 der größten *p*-Werte aus der Analyse ausgeschlossen worden wären (alle $p < .05$). Dies trifft ebenfalls für die halbe *p*-Kurve zu, die nur *p*-Werte kleiner als $p = .025$ berücksichtigt (alle $p < .05$). Darüber hinaus wäre die volle *p*-Kurve auch nach Ausschluss von bis zu 17 der kleinsten bzw. bis zu 17 der größten *p*- Werte nicht signifikant flacher ausgefallen, als eine, die auf Studien mit einer Teststärke von nur 33 % beruht (alle $p > .9$).

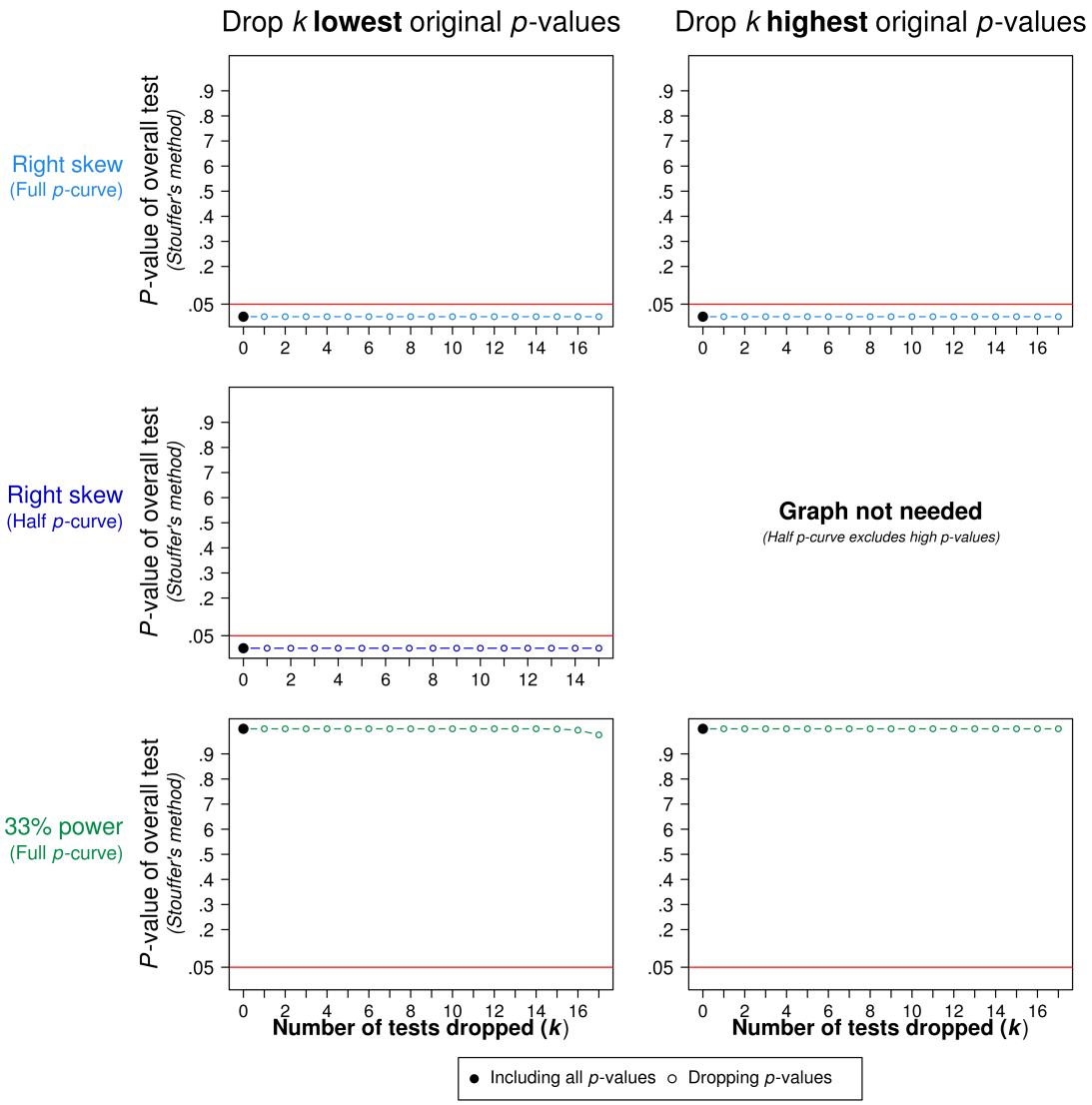

Abbildung 6. Bei den Tests auf Linkssteilheit hätten sich bei Ausschluss der kleinsten und größten p-Werte aus dem Fachgebiet der Sozialpsychologie keine Auswirkungen auf das Ergebnis der p-Kurven-Analyse ergeben. Das obere Drittel der Abbildung zeigt die Auswirkungen auf die volle p-Kurve, das mittlere Drittel auf die halbe p-Kurve und das untere Drittel auf den Test auf Flachheit vs. 33 % Teststärke. Die horizontale Achse repräsentiert die Anzahl (k) ausgeschlossener p-Werte, die vertikale Achse den p-Wert der daraus resultierenden p-Kurve. Die rote Linie markiert die Signifikanzschwelle von $\alpha = .05$.

Hypothese C: Verschiedenheit der Verteilungen. Um zu prüfen, ob sich die Verteilung der p-Werte zwischen den Fachgebieten signifikant unterscheidet, wurde ein Kolmogorow-Smirnow-Test sowie ein Cramér-Test durchgeführt. Die Voraussetzungen für den Einsatz der beiden nonparametrischen Testverfahren waren gegeben. Das Ergebnis des Kolmogorow-Smirnow-Tests deutet darauf hin, dass sich die beiden Verteilungen auf einem Signifikanzniveau von $\alpha = .05$ nicht signifikant voneinander unterscheiden, $D(39, 37) = 0.115$, *n. s.*, was der Cramér-Test bestätigt, $C(39, 37) = 0.004$, *n. s.*.

Ergebnisse der *Z*-Kurven-Analyse

Ergänzend zur *p*-Kurven-Analyse wurden folgende *Z*-Kurven erstellt (Brunner & Schimmack, 2018; Schimmack & Brunner, 2019), auf deren Grundlage die Replizierbarkeit der Studien aus den Fachgebieten Allgemeine Psychologie (*Abbildung 7*) und Sozialpsychologie (*Abbildung 8*) eingeschätzt werden sollte:

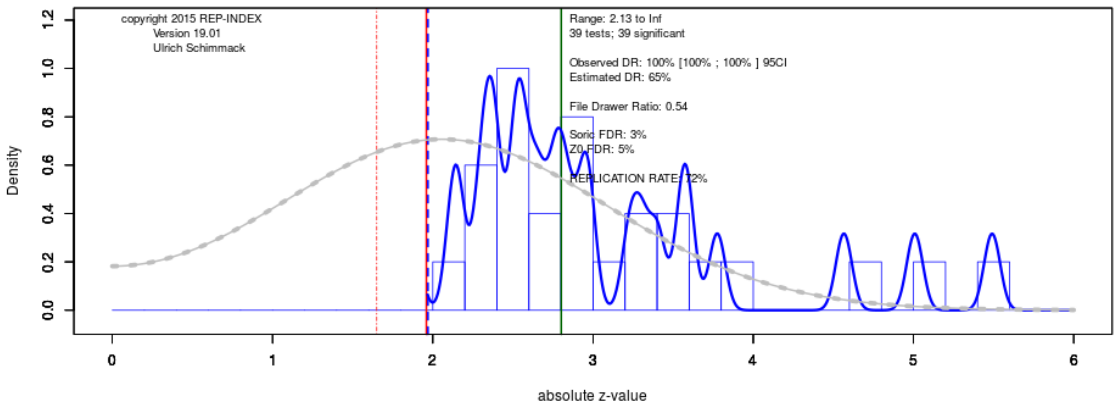

Abbildung 7. *Z*-Kurve der Allgemeinen Psychologie. Die solide vertikale rote Linie trennt nichtsignifikante Ergebnisse auf der linken Seite von signifikanten Ergebnissen auf der rechten Seite (α = .05). Die gepunktete rote Linie markiert die Schwelle für marginal signifikante Ergebnisse (α = .10). Hätten die Studien eine Teststärke von 80 %, dann wäre der häufigste Wert (Modus) bei der grünen Linie. Das blaue Histogramm gibt die Häufigkeiten der zugrundeliegenden Teststatistiken an. Die blaue Linie zeigt die Dichteverteilung der signifikanten *Z*-Werte. Die graue Linie geht in den nichtsignifikanten Bereich und gibt somit eine Schätzung darüber ab, wie viele nichtsignifikante Ergebnisse nicht veröffentlicht wurden (*file drawer*). Die *Z*-Kurve schätzt die Replizierbarkeit in der Allgemeinen Psychologie auf 72 %.

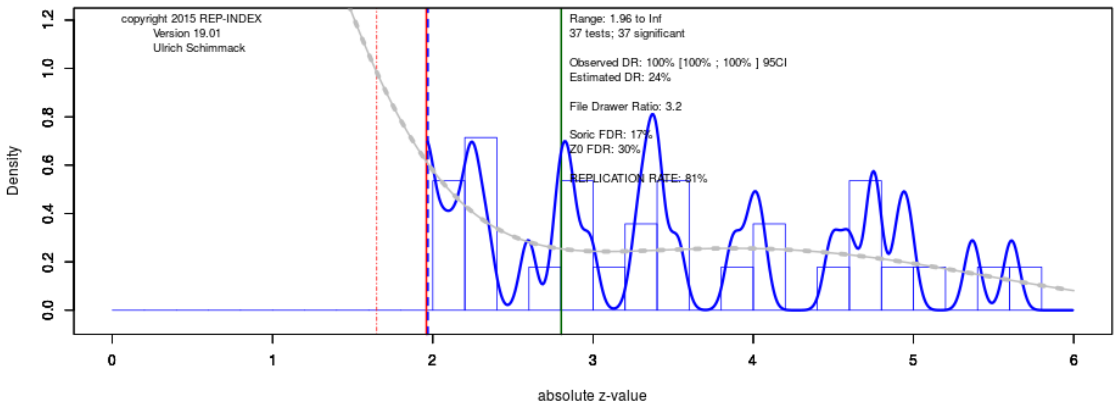

Abbildung 8. *Z*-Kurve der Sozialpsychologie. Die solide vertikale rote Linie trennt nichtsignifikante Ergebnisse auf der linken Seite von signifikanten Ergebnissen auf der rechten Seite (α = .05). Die gepunktete rote Linie markiert die Schwelle für marginal signifikante Ergebnisse (α = .10). Hätten die Studien eine Teststärke von 80 %, dann wäre der häufigste Wert (Modus) bei der grünen Linie. Das blaue Histogramm gibt die Häufigkeiten der zugrundeliegenden Teststatistiken an. Die blaue Linie zeigt die Dichteverteilung der signifikanten *Z*-Werte. Die graue Linie geht in den nichtsignifikanten Bereich und gibt somit eine Schätzung darüber ab, wie viele nichtsignifikante Ergebnisse nicht veröffentlicht wurden (*file drawer*). Die *Z*-Kurve schätzt die Replizierbarkeit in der Sozialpsychologie auf 81 %.

Diskussion

In der vorliegenden Studie ging es darum, zu analysieren, ob sich die Fachgebiete Allgemeine Psychologie und Sozialpsychologie in der Verteilung ihrer *p*-Werte unterscheiden. Daraus sollte abgeleitet werden, ob sich an der Forschungs- und Publikationspraxis in der Psychologie – insbesondere der Sozialpsychologie – seit Beginn der Replikationskrise etwas geändert hat. Anhand der erhobenen Daten konnte nicht nachgewiesen werden, dass es signifikante Unterschiede zwischen den Fachgebieten gibt. Hypothese C konnte somit nicht bestätigt werden. Allerdings ist anzunehmen, dass der Stichprobenumfang zu gering ausfiel, um valide Aussagen darüber zu treffen, wie die *p*-Werte zwischen den Fachgebieten tatsächlich verteilt sind. Der Umfang der Stichprobe war zwar ausreichend, um statistische Analysen durchführen zu können; angesichts der Masse potentiell auswertbarer Literatur ist die Stichprobe aber keinesfalls repräsentativ.

Diskussion der Allgemeinen Psychologie

Hypothese A, wonach die Verteilung der *p*-Werte in der Allgemeinen Psychologie linkssteil ausgeprägt sei, konnte bestätigt werden. Dieses Ergebnis ist vermutlich mit der Art des Forschungsgegenstandes begründbar, da die Allgemeine Psychologie häufig Untersuchungen auf der Mikroebene vornimmt, von denen angenommen werden kann, dass diese entweder relativ eindeutige Effekte erzielen oder keine, da der Untersuchungsgegenstand klar definiert bzw. eng umrissen ist. Man denke hier an das Konzept der Reliabilität aus der Messtheorie. Die Messverfahren in der Allgemeinen Psychologie gelten als eher reliabel, insbesondere verglichen mit dem Erschließen von Prozessen durch Fragebögen oder gar qualitativer Auswertungen in den Sozialwissenschaften. Eine reliable Messung kann zwar etwas anderes messen als gemessen werden sollte; dann wäre die aufgestellte Hypothese aber nicht signifikant. Wird hingegen ein signifikanter Unterschied bezüglich der Hypothese festgestellt, dürfte dieser aufgrund der reliableren Messverfahren deutlich signifikant ausfallen. Natürlich werden in der Allgemeinen Psychologie nicht ausschließlich biologisch-physiologische Messverfahren eingesetzt, sondern auch erschließende Methoden. Allerdings dürfte es einfacher sein, beispielsweise nach mehreren Monaten den Abruf bestimmter Wörter aus dem Gedächtnis zu testen, als den Einfluss einer deutlich globaleren sozialen Situation, die sowohl während der Manipulation als auch danach mit mehr Störvariablen konfundiert wird.

Es sind jedoch nicht nur die vermeintlich „schwammigen“ Untersuchungsgegenstände der Sozialpsychologie von geringer Reliabilität betroffen, sondern auch die kognitiven Neurowissenschaften (Bell, 2010; Bennett & Miller, 2010; Carp, 2012a, 2012b; Chambers, 2017, S. 31-33; Neuroskeptic, 2012; Poldrack et al., 2017), die mit der Allgemeinen Psychologie eng verwandt sind und von denen man, wie bereits ausgeführt wurde, eher Gegenteiliges annehmen würde. Dies würde einen Erklärungsbaustein liefern, warum sich die *p*-Kurven der beiden Fachgebiete nicht nennenswert voneinander unterscheiden. So glaubt beispielsweise Dirnagl (zitiert in Hubert, 2015), „dass 80 bis 90 % der neurowissenschaftlichen Studienergebnisse nicht reproduzierbar (…) sind“.

In Bezug auf die *p*-Kurve der Allgemeinen Psychologie fällt auf, dass deutlich mehr *p*-Werte im Bereich von $.01 \leq p < .02$ liegen als dies in der Sozialpsychologie der Fall ist (*Abbildungen 3* und *5*: 16 % vs. 3 %). Aufgrund des mutmaßlich weniger reliabel messbaren Forschungsgegenstandes der Sozialpsychologie ist dieser Unterschied unerwartet. Ebenfalls unerwartet ist die im Rahmen der *Z*-Kurven-Analyse geschätzte Replizierbarkeit in der Allgemeinen Psychologie, die mit 72 % geringer ausfiel als in der Sozialpsychologie (81 %). Allerdings werden mit einer *file drawer ratio* von 0.54 weniger nichtsignifikante unveröffentlichte Ergebnisse vermutet (*Abbildungen 7* und *8*).

Diskussion der Sozialpsychologie

Hypothese B, wonach die Verteilung der *p*-Werte in der Sozialpsychologie rechtssteil ausgeprägt sei, da Studien aus diesem Fachgebiet seltener repliziert werden können als jene aus der Allgemeinen Psychologie, konnte nicht bestätigt werden. Die Ergebnisse lassen nicht den Schluss zu, dass das Fachgebiet signifikant von fragwürdigen Forschungspraktiken betroffen wäre. Wie bereits ausgeführt wurde ist die vorliegende Stichprobe zu gering, um eine repräsentative Aussage über ein gesamtes Fachgebiet treffen zu können. Nichtsdestotrotz wäre es verwunderlich, wenn nach den spektakulären, öffentlichkeitswirksamen Betrugsfällen sowie der generellen Replikationskrise weiterhin massenhaft geschönte Ergebnisse veröffentlicht würden. Vielmehr ist anzunehmen, dass die Sozialpsychologie je nach Untersuchungsgegenstand naturgemäß größeren Aufwand betreiben muss, um weniger konfundierte und damit statistisch belastbarere Ergebnisse zu erzielen, als dies zum Beispiel bei der Allgemeinen Psychologie der Fall ist (Brookshire, 2018). Fälle von gezielter Manipulation dürften jedoch die Ausnahme sein. Es könnte sich letztlich auch um ein Phänomen der Wahrnehmungsverzerrung

handeln, wonach einige wenige spektakuläre Fälle das Ansehen des Fachgebiets beschädigten und damit zur Übergeneralisierung einluden und es ermöglichten, alle Forscher des Fachgebiets unter Generalverdacht zu stellen (*pars pro toto*-Effekt). Wie die Ergebnisse zeigen, liegt keine systematische Verzerrung vor, wonach die Sozialpsychologie eindeutig oder gar überwiegend Ergebnisse knapp unter der Signifikanzschwelle veröffentlichen würde. Die beiden Fachgebiete unterscheiden sich nicht signifikant. Dennoch fällt auf, dass in der Sozialpsychologie mit 5 % mehr knapp-signifikante Ergebnisse berichtet wurden als in der Allgemeinen Psychologie (0 %) (vgl. *Abbildungen 3* und *5*). Allerdings wurden auch marginal mehr hochsignifikante Ergebnisse berichtet als in der Allgemeinen Psychologie (*Abbildungen 3* und *5*: 76 % vs. 81 %).

Entgegen der aufgestellten Hypothesenbegründung fiel die im Rahmen der *Z*-Kurven-Analyse geschätzte Replizierbarkeit von Studien der Sozialpsychologie mit 81 % höher aus als in der Allgemeinen Psychologie (72 %). Allerdings befinden sich mit einer *file drawer ratio* von 3.2 schätzungsweise über drei Mal mehr nichtsignifikante Ergebnisse im *file drawer* als es signifikante Publikationen gibt (*Abbildungen 7* und *8*).

Schwierigkeiten bei der Datenerhebung

Bei der Auffindung und Auswertung der Studienergebnisse bereitete Schwierigkeiten, dass ein Zugriff auf Publikationen häufig selbst aus dem Netzwerk der Universität nicht möglich war. Zudem mangelte es – an manchen Lehrstühlen mehr als an anderen – an empirischen Arbeiten (z. B. randomisiert-kontrollierte Studien), was vor allem in der Sozialpsychologie der Fall war. Häufig hatten Studien außerdem keine klare Haupthypothese, sondern waren eher explorativer Natur. Dadurch blieb teilweise unklar, welcher *p*-Wert die Arbeit am besten repräsentiert. Teils gab es auch unterschiedlich signifikante Haupthypothesen innerhalb der gleichen Arbeit, wodurch die Entscheidung für einen bestimmten *p*-Wert einem Versuchsleitereffekt oder *Bias* gleichgekommen wäre, und die deshalb in der Auswertung unberücksichtigt blieben. Zudem konnten unveröffentlichte Studienergebnisse nicht in die *p*-Kurven-Analyse einbezogen werden. In der Regel werden „kleinere“ Erhebungen, die keine neuen Erkenntnisse versprechen, oder studentische Replikationsstudien nicht veröffentlicht, obwohl auch diese empirische Arbeiten darstellen.

Praktikabilität der Methode

Im Rahmen dieser Studie sollte auch die Praktikabilität der *p*-Kurven-Analyse eruiert werden. Hierbei ist insbesondere festzuhalten, dass das Herauslesen der Art des statistischen Ergebnisses, des Forschungsdesigns, der Haupthypothese und des dazugehörigen *p*-Wertes, der die Haupthypothese am besten repräsentiert, äußerst zeitaufwendig ist und fundierte statistisch-methodische Kenntnisse erfordert. Die Umsetzung dieses Vorhabens wurde auch dadurch erschwert, dass einige Autoren bei der Formulierung und Aufbereitung ihrer Ergebnisse offenbar nicht berücksichtigten, dass ihre Studien für den alleinigen Zweck einer Aufnahme in Meta-Analysen nicht vollständig rezipiert werden können. Unabhängig von der Aufbereitung kam es aber auch vor, dass zur effektiven Auswertung eine tiefere Fachkenntnis erforderlich gewesen wäre. Dies war insbesondere bei neurowissenschaftlichen Studien der Fall.

Für eine gewissenhafte Auswertung der *p*-Kurven-Analyse genügt es zudem nicht, möglichst passende *p*-Werte hinzuzufügen, die mit der Haupthypothese verbunden sind. Um die Ergebnisse der *p*-Kurven-Analyse nicht zu verzerren (Simonsohn et al., 2014a, S. 543) und um eine effiziente Begutachtung durch Kollegen zu gewährleisten, muss in der *p*-Wert-Tabelle zusätzlich angegeben werden, welches Forschungsdesign der ausgewerteten Haupthypothese zugrunde lag. Diese Anforderungen machen eine Aggregation von *p*-Werten mittels *Text-Mining-Software* faktisch unmöglich, was den Zeitaufwand nochmals deutlich erhöht (Bishop & Thompson, 2016).

Kritik an der Methode

Die *p*-Kurven-Analyse kann in ihrer Aussagekraft beeinträchtigt sein, wenn sich unter den zugrundeliegenden Studien einige befinden, deren Ergebnisse gefälscht oder fehlerhaft berichtet wurden. Problematisch ist es auch, wenn falsche *p*-Werte in die Analyse einbezogen wurden (die z. B. nicht zur Haupthypothese passen), oder wenn nach Kenntnis der Ergebnisse andere Studien als zuvor geplant analysiert wurden (Nelson et al., 2018). Baguley (2012) merkt zudem an, dass Betrüger eher an beeindruckenden *p*-Werten, wie z. B. $p < .0001$, interessiert seien, während sich die *p*-Kurven-Analyse in erster Linie auf knapp signifikante Ergebnisse fokussiere. Bishop und Thompson (2016) betonen ferner, dass eine flachere *p*-Kurve nicht bedeuten müsse, dass kein *p-Hacking* stattgefunden habe. Eine rechtssteile *p*-Kurve könne auch nicht *per se* als Indikator für

das Ausmaß von *p-Hacking* betrachtet werden, sofern kein spezifisches Modell vorliege, das zu den zugrundeliegenden *p*-Werten passe. Insbesondere bleibe *p-Hacking*, welches unter Ausnutzung so genannter *ghost variables* zustande kam, in der *p*-Kurven-Analyse häufig unentdeckt. Bei *ghost variables* handelt es sich um Variablen, die zwar erhoben wurden, jedoch nicht signifikant waren. Indem diese Variablen im Forschungsbericht verschwiegen werden, wird suggeriert, dass die Wahrscheinlichkeit, ein falsch-positives Ergebnis erhalten zu haben, deutlich geringer sei als sie es wäre, hätte man den tatsächlich beobachteten Effekt berichtet. Zudem ist in solchen Fällen zu vermuten, dass auch keine Korrektur für multiples Testen angewendet wurde (Bishop & Thompson, 2016).

Insgesamt betrachtet ist die *p*-Kurven-Analyse ein konservatives Verfahren (Simmons & Simonsohn, 2017, S. 2; Simonsohn et al., 2014a, S. 546; Ulrich & Miller, 2015): Studien, denen es an empirischer Evidenz mangelt, bleiben demnach häufiger unentdeckt als umgekehrt. *p*-Kurven sind linkssteil, wenn ein Effekt existiert, aber nur geringfügig rechtssteil, wenn die Studien manipuliert wurden (*p-Hacking*), oder wenn die Datenbasis aus wenigen echten und zahlreichen falschen Effekten kombiniert wurde (Simonsohn et al., 2014a). Dies trifft ebenfalls auf konfundierte Studienlagen zu (z. B. der positive Effekt von Eikonsum auf die Gesundheit, wenn in Wahrheit der Eikonsum mit mehr Sport einherging) (Simonsohn et al., 2014a).

Schimmack (2018a) gibt zu bedenken, dass die *p*-Kurven-Analyse noch keinem *Peer-Review*-Verfahren unterzogen wurde. Zudem kritisieren Brunner und Schimmack (2018), dass die *p*-Kurven-Analyse die durchschnittliche Teststärke der ihr zugrundeliegenden Studien – und damit deren Replizierbarkeit – überschätze, wenn die Studien heterogene Effekt- und Teststärken aufweisen. Simmons, Nelson und Simonsohn (2018) entgegnen, dass dies nur bei Einbeziehung von Studien mit extrem hohen Teststärken der Fall sei. Die *p*-Kurven-Analyse sei nicht geeignet, die Effektgröße in der Population zu schätzen, wenn auch nichtsignifikante Ergebnisse einbezogen werden (Schimmack, 2017). Dies sei allerdings irrelevant, wenn (z. B. im Kontext der Replikationskrise) die Replizierbarkeit bereits veröffentlichter Studien eingeschätzt werden soll; bei Heranziehung signifikanter Ergebnisse schätze die *p*-Kurven-Analyse die durchschnittliche Effektstärke der in die Analyse einbezogenen Studien überraschend gut ein (Schimmack, 2017). Es gehe Schimmack allerdings darum, die Replizierbarkeit von Studien z. B. in der Sozialpsychologie insgesamt und damit auf Grundlage heterogener Effekt- und Stich-

probengrößen (und dementsprechend heterogener Teststärken) beurteilen zu können, weshalb er die *Z*-Kurven-Methode entwickelte (Brunner & Schimmack, 2018; Schimmack, 2017). Diese nutzt die Verteilung beobachteter *Z*-Werte, um daraus einen Populationsparameter zu errechnen, der den Publikationsbias korrigiert (Schimmack, 2017).

Grenzen dieser Studie

Verglichen mit der Menge potentiell auswertbarer *p*-Werte fiel die Stichprobe in dieser Studie eher gering aus. Dementsprechend sollte von den Ergebnissen dieser Studie nicht ohne Weiteres auf die tatsächliche Verteilung der *p*-Werte in ihrer Gesamtheit geschlossen werden. Darüber hinaus wurden die Studien nicht auf Basis ihrer zugrundeliegenden Teststärke ausgewählt, was Simonsohn et al. (2014a) empfohlen hätten. Teststärken hätten somit auch unter 80 % gelegen haben können, was die Aussagekraft der errechneten *p*-Kurven hätte beeinträchtigen können. Ferner musste die Analyse von Kovariaten aus Zeitgründen unterbleiben und es wurden keine Tests auf Robustheit durchgeführt, wozu aufgrund der vorliegenden *p*-Werte allerdings keine Veranlassung bestand.

Ausblick

Die Methode der *p*-Kurven-Analyse nach Simonsohn et al. (2014a) ist ein vielversprechendes und relativ niedrigschwelliges Verfahren, um sich einen Eindruck davon zu verschaffen, ob die Forschungsergebnisse zu einem bestimmten Thema oder von einer bestimmten Personengruppe empirische Evidenz aufweisen oder nicht, also ob sie wahr sind oder ob sie auf selektives Berichten zurückzuführen sind (Simonsohn et al., 2014a). Somit lässt die *p*-Kurven-Analyse in gewisser Hinsicht auch erahnen, was sich im so genannten *file drawer* – der sprichwörtlichen Schublade mit den negativen Forschungsergebnissen – zu einem bestimmten Forschungsgegenstand verbirgt und wie prall diese Schublade wahrscheinlich gefüllt ist. Als eine für diesen Zweck geeignetere Methode hat sich die im Rahmen dieser Studie explorativ durchgeführte *Z*-Kurven-Methode (Brunner & Schimmack, 2018) erwiesen, die konkrete Schätzungen zum Ausmaß des *file drawers* und zur Replizierbarkeit abgibt. Allerdings kann die Anwendung beider Methoden mit zunehmendem Stichprobenumfang sehr schnell äußerst zeitaufwendig werden. Die vorliegende Studie konnte somit lediglich das Fundament für zukünftige umfangreichere Erhebungen legen, die mit größerer Personalstärke durchgeführt werden sollten, um mögliche Unterschiede zwischen den Fachgebieten zu entdecken.

Es mag ironisch anmuten, unter Heranziehung von *p*-Werten eine Reihe von *p*-Werten auf ihre Aussagekraft hin zu überprüfen. In einer zukünftigen Auswertung wäre es darum sinnvoll, eine alternative Auswertung mittels Bayes'schem Testen vorzunehmen. Dies würde eine Aussage darüber zulassen, wie wahrscheinlich die berichteten Ergebnisse von Publikationsbias und/oder *p-Hacking* betroffen sind oder nicht.

Weitere Anwendungsgebiete

Die *p*-Kurven-Analyse ließe sich auch einsetzen, um im Rahmen von größer angelegten Erhebungen weitere Hypothesen zu generieren und zu überprüfen; zum Beispiel inwiefern sich die *p*-Werte zwischen Ländern, Universitäten oder Lehrstühlen unterscheiden. Zudem kann die *p*-Kurven-Analyse eingesetzt werden, um die Verteilung von *p*-Werten in den Arbeiten einzelner Wissenschaftler zu analysieren, sofern entsprechend zahlreiche Publikationen vorliegen. Dies sollte jedoch mit Bedacht geschehen (Baguley, 2012; Srivastava, 2012a). Das vermutlich sinnvollste Einsatzgebiet für die Durchführung einer *p*-Kurven-Analyse ist es, den Forschungsstand zu psychologischen Konstrukten bzw. Theorien auf empirische Evidenz hin zu analysieren, um eine Aussage darüber treffen zu können, wie wahrscheinlich die berichteten Effekte wahr sind oder nicht. Sollte die *p*-Kurven-Analyse zum Standard werden (auch und insbesondere für die Analyse einzelner Forscher oder Forschergruppen) wird *p-Hacking* erschwert. Diejenigen aber, die unbedingt signifikante Ergebnisse erzielen wollen, weil sich beispielsweise an ihren beruflichen Rahmenbedingungen nichts Wesentliches ändert, werden künftig womöglich mehr Energie darauf verwenden, „noch signifikantere" Ergebnisse zu erzielen. Um das zu erreichen, könnte es sein, dass erst recht kriminelle Energie aufgewendet wird. An dieser Stelle ist jedoch anzumerken, dass es in dieser Studie nicht darum ging, *p-Hacking* (im Sinne eines explorierenden Umgangs mit Daten) zu kriminalisieren, da es oftmals sinnvoll ist, um aus Daten zu lernen (Simmons, Nelson & Simonsohn, 2013). In den meisten Fällen dürfte es zudem nicht bewusst-vorsätzlich, sondern eher unbewusst-fahrlässig geschehen (vgl. Kunda, 1990; Nelson et al., 2018).

Maßnahmen gegen die Replikationskrise

Der Kritik an der bisherigen Forschungs- und Veröffentlichungspraxis sollen nun konkrete Vorschläge zu ihrer Verbesserung folgen (siehe auch Munafò et al., 2017).

Präregistrierung von Studien. Das wohl wirksamste Mittel gegen fragwürdige Forschungspraktiken und *Publikationsbias* ist die konsequente Präregistrierung geplanter Studien (z. B. unter *aspredicted.org*) (Chambers, 2017, S. 177 ff., vgl. Kupfersmid, 1988; Newcombe, 1987; Walster & Cleary, 1970). Noch besser wäre es, so Chambers, wenn Fachzeitschriften Artikel zur Veröffentlichung annehmen würden, ohne bereits die Ergebnisse zu kennen, indem sie lediglich die Wichtigkeit der Forschungsfrage, die Sorgfalt der Methodik, die Plausibilität der Hypothesen und der Analyseverfahren (einschließlich der angestrebten Teststärke/Stichprobengröße) sowie die Replizierbarkeit beurteilten. Die Begutachtung durch Kollegen, das so genannte *Peer-Review*-Verfahren oder Kreuzgutachten, müsse außerdem zweigeteilt werden: Im ersten Schritt würde die Theorie, die daraus abgeleiteten Hypothesen, die Methodik und die darin festgelegten Analyseverfahren ausführlich beschrieben. Dann würden diese vorab z. B. im *Open Science Framework* (unter *osf.io/prereg*) registriert und könnten somit im Nachhinein nicht mehr verändert werden. Erst nach der Präregistrierung würden die Daten erhoben und wie zuvor festgelegt ausgewertet. Im zweiten Schritt würde durch kollegiale Begutachtung geprüft, inwieweit sich an das vorab registrierte Vorgehen gehalten wurde. Zudem müssten die Daten öffentlich archiviert werden, um die korrekte Durchführung der vorab festgelegten Analysemethode überprüfen zu können (Chambers, 2017).

Indem die zu überprüfenden Hypothesen vorab ausformuliert und das geplante Analyseverfahren festgelegt werden würde, könnte eindeutig zwischen konfirmatorischen und explorativen Analysen unterschieden werden. Explorative Analysen könnten selbstverständlich weiterhin genutzt werden, um Hypothesen für eine weitergehende Forschung zu generieren. Ein nachträgliches Verändern der Hypothesen (*HARKing*) wäre jedoch effektiv unterbunden. Durch die Festlegung auf bestimmte Analyseverfahren würde zudem *p-Hacking* effektiv begrenzt, indem die Möglichkeiten zum mehrfachen Testen reduziert würden. Um das Prinzip der Präregistrierung zu verbreiten, wäre es Chambers (2017) zufolge hilfreich, es im Rahmen der Vergabe von Forschungsgeldern vorauszusetzen.

Bayes'sches Testen. Eine weitere Möglichkeit, fragwürdige Forschungspraktiken einzudämmen, besteht im Umstieg vom klassischen *null hypothesis significance testing (NHST)* zum Bayes'schen Testen (vgl. Chambers, 2017, S. 68 ff.). Entgegen dem, was man eigentlich wissen will, sagt das klassische *NHST* nichts über die Wahrscheinlichkeit darüber aus, mit der eine zu überprüfende Hypothese zutreffend ist. *NHST* sagt lediglich etwas darüber aus, mit welcher Wahrscheinlichkeit ein „Ergebnis" (die Alternativhypothese) im Test als „wahr" bzw. „vorhandener Effekt" angegeben wird, *obwohl* in Wahrheit die Nullhypothese gilt, also kein Effekt vorliegt. Deswegen gibt der *p*-Wert nicht die Wahrscheinlichkeit an, mit der eine Hypothese zutrifft. Stattdessen möchte man möglichst kleine *p*-Werte erzielen, weil damit auch die Wahrscheinlichkeit geringer ausfällt, das Ergebnis irrtümlich erhalten zu haben. Zudem irrt man sich beim *NHST* mit jeder Testung – auf einem Signifikanzniveau von $\alpha = .05$ mit einer Wahrscheinlichkeit von 5 %. Wird korrekt vorgegangen, wird die für ein Experiment benötigte Stichprobengröße zuvor (*a priori*) festgelegt, und die Daten erst im Anschluss – also nach Erreichen der benötigten Stichprobengröße – analysiert. Bayes hingegen gibt die Wahrscheinlichkeit an, mit der eine Hypothese wahr ist, und erlaubt kontinuierliches Testen auch bei erweitertem Stichprobenumfang ohne Alphafehlerkumulierung. Dafür erfordert es die Angabe diverser *prior odds*, die theoretisch fundiert und schlüssig kommuniziert werden müssen. Dies bedeutet einen Mehraufwand, der jedoch das Potential hat, die bisherige Konkurrenzsituation in der Forschergemeinschaft zu einer Kollaborationssituation hin zu verschieben, bei der es weniger um die Pflege der eigenen Reputation als vielmehr um das gemeinschaftliche Erarbeiten eines möglichst akkuraten und vollständigen Abbildes der Wahrheit geht (Chambers, 2017, S. 73). Statt also seine Reputation auf ein einzelnes statistisches Ergebnis zu stützen, das möglicherweise durch Zufall entstanden ist, und dieses um jeden Preis zu verteidigen, sollte die Energie besser auf gemeinsames Erarbeiten immer wahrheitsgetreuerer, möglichst akkurater *prior odds* bzw. *probabilities* verwendet werden – gewonnen durch zahlreiche Replikationsstudien und gemeinschaftlich zusammengetragenes Wissen, um sich immer mehr der Wahrheit anzunähern. Besonders wirkungsvoll gegen *Bias* und gezielte Manipulationsversuche wäre Bayes'sches Testen in Kombination mit der Präregistrierung von Studien (Chambers, 2017, S. 73).

Open Access. Wissenschaft basiert auf der freien Weitergabe von Wissen. Obwohl es von der Öffentlichkeit finanziert wurde, wird dieses Wissen der Öffentlichkeit nicht wieder zur Verfügung gestellt: Daten werden nicht geteilt, nichtsignifikante Ergebnisse verschwinden in der Schublade, signifikante Ergebnisse werden in kommerziellen Zeitschriften veröffentlicht und verschwinden hinter einer *Paywall*, sodass der Abruf einzelner Artikel teils über 30 Euro kostet. Universitäten können sich die Abonnements wesentlicher Publikationen immer häufiger nicht mehr leisten, wodurch selbst Wissenschaftler des jeweiligen Fachgebiets keinen Zugriff mehr auf die Forschungsergebnisse ihrer Kollegen haben und bei den jeweiligen Autoren eine Kopie erfragen müssen. Auch hier ist der Antreiber das eigene Prestige: Es geht nicht um die Veröffentlichung an sich, sondern darum, in der angesehensten Zeitschrift zu veröffentlichen. „Veröffentlichen“ bedeutet jedoch, „der Öffentlichkeit zugänglich machen“ (Duden, 2019). Genau das geschieht aber nicht. Stattdessen sehen sich Wissenschaftler sogar dazu gezwungen, privates Geld zu investieren (Taylor, 2013). Insbesondere jungen, aufstrebenden Wissenschaftlern bleibt möglicherweise keine andere Wahl (Colquhoun, 2012; Chambers, 2017, S. 131f.), wenn sie ihre Forschung einem größeren Fachpublikum zugänglich machen möchten, da die wenigsten Psychologie-Journale ein volles *Open Access*-Modell anbieten – von den renommiertesten Zeitschriften keine einzige (Chambers, 2017, S. 129). Dabei korreliert das Renommee einer Zeitschrift (oftmals ausgedrückt durch den *impact factor*) nicht unbedingt mit der Qualität der darin (nicht) veröffentlichen Forschungsergebnisse (Brembs, Button & Munafò, 2013).

Das *Peer-Review*-Verfahren hat den Zeitschriften zu ihrem Ansehen verholfen; gleichwohl werden die Gutachter kaum bis gar nicht für ihre Arbeit entlohnt, während sich die Herausgeber mehrfach für ein Produkt bezahlen lassen, das fast vollständig von anderen produziert wurde (Buranyi, 2017; Chambers, 2017, S. 135). Chambers (2017, S. 130) beschreibt, dass die Vorabveröffentlichung der eigenen Manuskripte zwar meistens erlaubt bleibe, die Herausgeber der Zeitschriften aber ein „sauberes Narrativ“ bevorzugten, was dazu führe, dass die Hypothesen den gefundenen Ergebnissen angepasst werden sollten. Würden die Manuskripte vor dieser Veränderung veröffentlicht, könnte *HARKing* schnell entlarvt werden – mit den entsprechenden Konsequenzen für das Ansehen der betroffenen Wissenschaftler.

Chambers (2017, S. 208 ff.) empfiehlt Wissenschaftlern, nur noch in Zeitschriften zu veröffentlichen, die vom *RoMEO*-Projekt (siehe *sherpa.ac.uk/romeo*) als grün oder blau gekennzeichnet wurden. Das bedeutet, dass die Artikel frei und ohne Zeitverzögerung („Embargo“) zugänglich sein müssen, und zum Beispiel auf der eigenen Webseite veröffentlicht werden dürfen. Geldgeber, Institutionen und Regierungen sollten es Chambers zufolge ebenfalls zur Bedingung machen, dass ihre Forschung in *Open Access*-Zeitschriften veröffentlicht wird. Langfristig plädiert Chambers (2017, S. 208 ff.) dafür, dass Universitätsbibliotheken die Veröffentlichung und Archivierung der Forschungsergebnisse übernehmen. Zudem schlägt Chambers (2017, S. 210) vor, die Nennung der Autoren alphabetisch vorzunehmen und die konkreten Beiträge zu einer Publikation detailliert (prozentual, S. 211) und nach Sektionen unterteilt zu benennen. Dies solle die Nutzung „oberflächlicher Zitationsmetriken“ (Chambers, 2017, S. 210) bei der Beurteilung von Wissenschaftlern und ihrer Forschung eindämmen.

Open Data. Das Veröffentlichen der eigenen Rohdaten habe Chambers (2017, S. 77) zufolge zahlreiche Vorteile. So könnten unabhängige Forscher die Analysen wiederholen, was dazu beitrage, ehrliche Fehler zu finden und zu korrigieren. Zudem ermögliche die Veröffentlichung der Materialien die Durchführung direkter Replikationen. Eine Offenlegung von Rohdaten ermögliche es auch, fragwürdige Forschungspraktiken und Betrug aufzudecken, was deren Nutzung in Zukunft eindämmen könnte. Simonsohn (2013) schlägt deshalb vor, dass Verlage die Veröffentlichung der dazugehörigen Rohdaten voraussetzen sollten. Ein weiterer Nebeneffekt wäre, dass durch Vervielfältigung von Daten versehentlichen Datenverlusten aufgrund von Hardwaredefekten oder veralteten Datenträgern (siehe Diskette, VHS und CD, oder Daten als schwer auffindbarer Teil alter Sicherungskopien) vorgebeugt werden würde, und ein Zugriff auf die Daten selbst dann noch möglich wäre, wenn die Autoren aufgrund geänderter Kontaktdaten oder im Abwesenheits-, Krankheits- oder Todesfall nicht (mehr) erreichbar wären (vgl. Wicherts & Bakker, 2012).

Möglicherweise werden verbesserte Analysemethoden dazu führen, dass künftig weniger die Ergebnisse von Interesse sind, als vielmehr die zugrundeliegenden Daten. Auch für Meta-Analysen und *„Big Data“*-Anwendungen ist ein Zugriff auf Rohdaten wichtig (Chambers, 2017, S. 78). Forschung wird in der Regel mit öffentlichen Geldern finanziert und basiert insbesondere in der Psychologie häufig auf der Teilnahmebereit-

schaft unbezahlter Freiwilliger. Somit sollten die erhobenen Daten auch der Allgemeinheit zugute kommen. Wer für moralische Argumente weniger empfänglich ist, den weist Chambers (2017) darauf hin, dass *Open Data* die Anzahl der Zitierungen signifikant erhöht und somit substantielle Karrierevorteile verspricht (vgl. Piwowar, Day & Fridsma, 2007; Piwowar & Vision, 2013).

Auch die *APA* erfordert in ihren Richtlinien die Veröffentlichung von Daten oder zumindest deren Herausgabe an Interessierte, da dies den wissenschaftlichen Fortschritt sowie eine Kultur der Offenheit und Verantwortlichkeit in der Forschung fördere (APA, 2015). Zudem ermögliche es Menschen, die keine eigenen Studien durchführen können (aufgrund geographischer Abgeschiedenheit oder mangelnder finanzieller Ressourcen), die Erforschung von für sie relevanten Fragestellungen sowie die Nutzung neuer, besserer Analysemethoden, die zum Zeitpunkt der Erhebung noch nicht zur Verfügung standen. Durch Aggregation von Daten könne zudem die Generierung neuer Hypothesen unterstützt, wissenschaftliche Aussagen besser generalisiert und Entscheidungen zuverlässiger getroffen werden (APA, 2015).

Höhere Teststärken. Wenn es eine Krise der Replikation gibt, dann ist der erste Schritt zur Lösung die Durchführung von Studien mit größerer Teststärke. Chambers (2017, S. 67) zufolge bedarf es dazu einer Reform der gegenwärtigen wissenschaftlichen Kultur, die mehr Wert auf Quantität als auf Qualität lege, und die die Anzahl der Publikationen höher bewerte als die Zuverlässigkeit der zugrundeliegenden Forschungsergebnisse. Um größere Teststärken zu erreichen sollten Studien „belohnt“ werden, die replizierbar sind. Das wiederum erfordere die Anerkennung der Durchführung direkter Replikationen als wesentlichen Bestandteil wissenschaftlicher Arbeit. Wenn als Voraussetzung für eine Publikation in – insbesondere einflussreichen – Zeitschriften eingeführt werden würde, dass eine Studie mindestens einmal repliziert werden konnte, könnte dem effektiv entgegen gewirkt werden (vgl. Gervais, 2014). So hat Srivastava (2012b) vorgeschlagen, dass Zeitschriften – das könnten insbesondere jene sein, die an Renommee gewinnen möchten – verpflichtet werden sollten (oder sich zumindest selbst dazu verpflichten sollten), zu jeder publizierten Studie auch eine dazugehörige Replikationsstudie zu veröffentlichen. Dies hätte nicht nur ein Korrektiv zu den Publikationen zur Folge, sondern schaffe auch einen Anreiz für andere Wissenschaftler, Replikationsstudien durchzuführen. In einer Studie von Martin und Clarke (2017) stellte sich heraus, dass nur 3 %

der psychologischen Fachzeitschriften explizit angaben, auch Replikationsstudien aufzunehmen. Ein Beispiel findet sich bei *BMC Psychology:* Dort wolle man auch negative Ergebnisse und Replikationen explizit willkommen heißen, sofern diese einen „nützlichen Beitrag zur Forschung“ leisteten (Chambers, 2017, S. 68; Laws, 2013).

Feindliche Zusammenarbeit. In einer Welt, in der die Sorge um die eigene Reputation schwerer wiegt als das Interesse an der Wahrheit, schlägt Chambers (2017, S. 73) das Prinzip der *Adversarial Collaborations* vor. Hierbei würden sich zwei Forschergruppen, die geteilter Ansicht über die Existenz eines Effekts sind, auf ein Forschungsdesign einigen. Dieses Forschungsdesign würde dann umgesetzt, die Daten zusammengetragen und nach einem zuvor definierten Vorgehen gemeinsam ausgewertet. Wer das Ergebnis auf seiner Seite hat, „gewinnt“, und beide Parteien wären aufgrund des zuvor gemeinsam festgelegten Vorgehens gezwungen, das Ergebnis zu akzeptieren, anstatt mit einer vermeintlich mangelhaften Methodik des Gegenübers zu argumentieren. Ein Beispiel findet sich bei Matzke et al. (2015).

Auszeichnungen. Eine weitere Initiative geht auf das von Brian Nosek geführte *Center for Open Science* zurück, bei der Zeitschriften Auszeichnungen, *Badges* oder digitale Medaillen für transparente Forschungspraktiken vergeben – zum Beispiel für die Offenlegung der Rohdaten, der Materialien und für die Präregistrierung der Studie (Chambers, 2017, S. 73 f.; Center for Open Science, 2019).

Replikationszähler. Da Metriken sich großer, wenn auch zweifelhafter, Beliebtheit erfreuen, wird beispielsweise mit dem *journal impact factor (JIF)* versucht, den Einfluss und damit das Renommee von Fachzeitschriften zu messen, indem man jedes Jahr die Anzahl von Zitaten auf Artikel einer Zeitschrift mit der Anzahl der darin veröffentlichten Artikel ins Verhältnis setzt – bezogen auf die vorangegangenen zwei Jahre. Dies geschehe allerdings nach „unwissenschaftlichen“ und „willkürlichen“ Maßstäben (PLoS Medicine Editors, 2006, S. 707; Rossner, van Epps & Hill, 2017). Curry (2012), Marcus und Oransky (2013) sowie Chambers (2017, S. 74, S. 151 ff., S. 199) kritisieren diese Metrik unter anderem auch deshalb, weil sie *Goodhart's law* zufolge falsche Anreize setze: *„When a measure becomes a target, it ceases to be a good measure.“* (Strathern, 1997, S. 308).

Bereits vor über 25 Jahren zeigte Seglen (1992), dass der *impact factor* verzerrte Ergebnisse liefert, da die Hälfte aller Zitationen auf lediglich etwa 15 % der Publi-

kationen entfällt. Das bedeutet, dass wenige einschlägige Veröffentlichungen den Einfluss einer Zeitschrift (verzerrt) abbilden: 85 % der veröffentlichten Artikel werden seltener zitiert als der Durchschnitt. Zudem stellte sich heraus, dass ein hoher *impact factor* eher mit einer hohen Zurücknahmerate korreliert (Fang & Casadevall, 2011) und weniger mit dem, was er eigentlich abbilden soll, nämlich Zitationen (Brembs, 2011). Darüber hinaus konnte keine Korrelation zwischen dem *impact factor* einer Zeitschrift und der Höhe der Teststärken der darin publizierten Studien gefunden werden (Brembs et al., 2013). Marcus und Oransky (2013) schlagen als sinnvolleres Maß deshalb die Anzahl erfolgreicher Replikationen vor – den so genannten *Replication Index*. Dieser könne zudem zu Indizes zusammengefasst werden, die die Replikationsleistung eines bestimmten Fachgebiets, einer Universität oder einer Zeitschrift widerspiegelten (Chambers, 2017, S. 199). Eine weitere Maßnahme zur Stärkung der Replikationskultur bestehe in der freiwilligen Selbstverpflichtung von Zeitschriften, zu bereits veröffentlichten Studien immer auch deren Replikationen zu veröffentlichen, sobald diese zur Verfügung stehen. Und auch Geldgeber könnten einen Teil ihrer Fördermittel für Replikationsvorhaben reservieren (Chambers, 2017, S. 200).

Weitere Vorschläge sind so genannte *Altmetriken* und der *h*-Index, wobei an letzterem deutliche Kritik aufgrund einfacher Manipulierbarkeit geübt wurde (Labbé, 2010). *Altmetriken* messen beispielsweise, wie oft Artikel heruntergeladen (und z. B. in Literaturverwaltungsprogrammen wie *Zotero* integriert) wurden, wie häufig diese in der *Wikipedia*, in Blogs oder in den so genannten „sozialen Medien“ erwähnt wurden (Adie & Roe, 2013; Priem, Groth & Taraborelli, 2012; Priem, Taraborelli, Groth & Neylon, 2010; Thelwall, Haustein, Larivière & Sugimoto, 2013).

Offenlegungspflichten. Ein weiterer Baustein für mehr Transparenz in der Wissenschaft sind Chambers (2017, S. 73 f.) zufolge Checklisten, auf denen die Autoren angeben müssen, ob (und an welcher Stelle) in ihrer Veröffentlichung nachzulesen ist, welche Beobachtungen ausgeschlossen wurden und warum, ob sämtliche Ergebnisse (positive wie negative) berichtet wurden und ob dargelegt wurde, wie die Größe der Stichprobe bestimmt wurde, etc. – was letztlich auch den Umfang der Methodensektion reduziere. Beispiele finden sich seit 2013 bei *Nature Neuroscience* (Nature Neuroscience Editors, 2013) und seit 2014 bei *Psychological Science* (Chambers, 2017, S. 74).

In ihrem Artikel über *„False-Positive Psychology“* schlagen Simmons et al. (2011) sechs Offenlegungspflichten für Autoren vor. Demnach sollen Autoren (1.) bereits vor dem Beginn der Datenerhebung entscheiden, wann die Erhebung beendet wird, und diese Entscheidungsregel in ihrem Artikel angeben; (2.) zumindest 20 Beobachtungen pro Zelle erheben oder einen triftigen Grund dafür nennen, warum dies z. B. aus Kostengründen nicht möglich war; (3.) alle Variablen, die in einer Studie erhoben wurden, benennen; (4.) alle Bedingungen des Experiments berichten, einschließlich fehlgeschlagener Manipulationen; (5.) offenlegen, ob Beobachtungen aus der Analyse ausgeschlossen wurden und wie die statistischen Ergebnisse aussehen würden, wenn sie einbezogen werden; (6.) angeben, ob eine Kovariate in die Analyse einbezogen wurde, und die Ergebnisse auch ohne diese Kovariate berichten.

Zudem plädieren Simmons, Nelson und Simonsohn (2012, S. 4) für die Nutzung des folgenden, 21 Worte umfassenden Transparenzhinweises: „Wir berichten, wie wir die Größe unserer Stichprobe bestimmt haben, alle ausgeschlossenen Daten (sofern zutreffend), alle Manipulationen und alle durchgeführten Analysen.“

Sci-Hub. Im Jahr 2011 startete Alexandra Elbakyan *Sci-Hub,* eine Internetseite, die wissenschaftliche Publikationen ohne Beachtung des Urheberrechts zugänglich macht, um damit gegen die hohen Kosten zu protestieren, die kommerzielle Verlage verursachen. So erzielte der Wissenschaftsverlag *Elsevier* im Jahr 2018 eine Profitmarge von 37 % und damit einen Reingewinn von einer Milliarde Euro (Page, 2019). In Zeiten, in denen selbst so genannte Elite-Universitäten wie *Harvard* von Abonnements absehen müssen, würden insbesondere Entwicklungsländer benachteiligt, so Elbakyan (van der Sar, 2015). Aber auch an großen Universitäten in Russland und in westlichen Ländern ist ein Zugriff auf Forschungsergebnisse keineswegs selbstverständlich, so Chambers (2017, S. 139). Sowohl Elbakyan als auch Chambers berichten, dass Forscher (also die Urheber) teilweise sogar gezwungen seien, für ihre eigenen Zeitschriftenartikel Geld zu bezahlen (Chambers, 2017, S. 137; van der Sar, 2015)

Elbakyan sieht in *Sci-Hub* deshalb eine konsequente Manifestation von Artikel 27 der Allgemeinen Erklärung der Menschenrechte, wonach jeder das Recht habe, „am kulturellen Leben der Gemeinschaft frei teilzunehmen, sich an den Künsten zu erfreuen und am wissenschaftlichen Fortschritt und dessen Errungenschaften teilzuhaben“. Zudem habe jeder „das Recht auf Schutz der geistigen und materiellen Interessen, die ihm als Ur-

heber von Werken der Wissenschaft, Literatur oder Kunst erwachsen“ (United Nations, 1948). Ob ein Geschäftsmodell mit Artikel 27 vereinbar ist, bei dem mit der wissenschaftlichen Leistung anderer Geld verdient wird, ohne die Urheber zu entlohnen, sondern sich sogar von ihnen bezahlen zu lassen, sei demnach fraglich (van der Sar, 2015).

Ein *Wiki* für Forschungsergebnisse. Dank des Internets wäre es möglich, Forschungsergebnisse in einem *Wiki*-System, angelehnt an *Wikipedia* und *Wikidata* (siehe *wikidata.org*), zu veröffentlichen. Dies würde keine umfangreichen Publikationen erfordern. Forscher, die eine Meta-Analyse durchführen möchten, würden z. B. Übersichtstabellen vorfinden mit Ergebnissen, die sich auf bestimmte Forschungsgegenstände beziehen. Beispielsweise könnte man dem Themenbaum „Psychologie > Allgemeine Psychologie > Generierungseffekt“ folgen, unter dem sämtliche Ergebnisse mit Bezug zum Generierungseffekt eingestellt werden könnten. Dazu könnten sie sich Details anzeigen lassen und daraufhin entscheiden, ob sie die Ergebnisse in ihre Analysen einbeziehen möchten oder nicht. Ein an diese Gedanken angelehntes Projekt auf *Wiki*-Basis findet sich unter *replication.uni-goettingen.de*.

Auf diese Weise könnten auch Studierende z. B. im Rahmen ihres empirischen Praktikums erhobene Daten und Ergebnisse zugänglich zu machen. Die Option, bestimmte Daten nicht zu verwenden, bestünde natürlich immer. Daten zu nutzen, die nie veröffentlicht wurden, ist dagegen unmöglich. Insbesondere da die Psychologie trotz ihres Status als empirische Naturwissenschaft häufig als *soft science* bezeichnet wird, ist es wichtig, dass die Forscher kollaborieren statt konkurrieren, veröffentlichen statt geheim halten, teilen statt zurückhalten, „denn geschönte Ergebnisse verlangsamen oder verhindern den wissenschaftlichen Fortschritt“ (Schimmack & Heene, 2016, o. S.).

Initiativen wie das *Journal of Articles in Support of the Null Hypothesis (JASNH)* (siehe *jasnh.com*) und *PsychFileDrawer* (siehe *psychfiledrawer.org*) versuchen, das *file-drawer problem* zu lösen, indem sie eine Plattform zur Verfügung stellen, auf der man auch nichtsignifikante Ergebnisse veröffentlichen kann. Ohne eine solche Plattform würden Forscher ihre Zeit damit vergeuden, empirische Fragestellungen zu untersuchen, die längst untersucht worden seien, so der Herausgeber des *JASNH* (Reysen, 2002). Zudem hilft die Veröffentlichung von Nulleffekten dabei, das Risiko falsch-positiver Ergebnisse einzuschätzen (Schimmack, 2019; Sterling, 1959).

Positive Entwicklungen

Seit dem Beginn der Replikationskrise in den Jahren 2010-2012 habe sich viel getan. Darum sprechen Nelson et al. (2018) inzwischen von einer Renaissance der Psychologie. Die Replikationskrise scheine abzuklingen: Autoren würden mittlerweile bereitwillig Daten teilen, renommierte Zeitschriften würden sowohl erfolgreiche als auch erfolglose Replikationsversuche veröffentlichen und zahlreiche Forscher würden ihre Studien sowohl präregistrieren als auch mit deutlich größerem Stichprobenumfang durchführen. Insgesamt werde vermehrt Wert auf Replizierbarkeit gelegt – eine Einschätzung, die von den Ergebnissen dieser Studie gestützt wird.

Zudem bildeten sich zwischenzeitlich so genannte *Open Science*-Initiativen heraus, in denen sich Forscher zusammenschließen, um gemeinsam für eine transparente Wissenschaft einzustehen (vgl. Glöckner, Fiedler & Renkewitz, 2018, S. 25). Beispielsweise bekennen sich die Unterzeichner der Initiative *Commitment to Research Transparency and Open Science* (Schönbrodt, Maier, Heene & Zehetleitner, 2015) zu den Werten Offenheit und Transparenz in der Wissenschaft, da sie davon überzeugt seien, dass entsprechende Forschungspraktiken die Glaubwürdigkeit, den Wert und die Wirkung ihrer Forschung erhöhen. Dies schlägt sich auch in der Personalauswahl nieder. So wurde zum Beispiel folgender Passus in Ausschreibungen für Professuren im Fach Psychologie an der Ludwig-Maximilians-Universität München eingefügt: „Das Department Psychologie legt Wert auf transparente und replizierbare Forschung und unterstützt diese Ziele durch Open Data, Open Material und Präregistrierungen. Bewerber werden daher gebeten, in ihrem Anschreiben darzulegen, auf welche Art und Weise sie diese Ziele bereits verfolgt haben und in Zukunft verfolgen möchten.“ (Schönbrodt & Scheel, 2017, S. 138 f.).

Insgesamt scheint die Replikationskrise dazu geführt zu haben, dass die Psychologie begonnen hat, ihre bisherige Forschungs- und Veröffentlichungspraxis umfassend zu reflektieren. Es bleibt abzuwarten, welche Maßnahmen gegen die Replikationskrise letztlich umgesetzt werden und wie nachhaltig diese Änderungen sein werden.

Literaturverzeichnis

Adie, E., & Roe, W. (2013). Altmetric: Enriching scholarly content with article-level discussion and metrics. *Learned Publishing, 26*(1), 11-17.

American Psychological Association. (2002). Ethical principles of psychologists and code of conduct. *American Psychologist, 57*(12), 1060-1073.

American Psychological Association. (2015). *Data sharing: Principles and considerations for policy development* [PDF]. Abgerufen unter https://www.apa.org/science/leadership/bsa/data-sharing-report.pdf

Baguley, T. (2012, 16. Februar). Simulating p curves and detecting dodgy stats [Blogartikel]. *Psychological Statistics.* Abgerufen unter http://psychologicalstatistics.blogspot.com/2012/02/ simulating-p-curves-and-detecting-dodgy.html

Bakker, M., & Wicherts, J. M. (2011). The (mis) reporting of statistical results in psychology journals. *Behavior research methods, 43*(3), 666-678.

Begley, C. G., & Ioannidis, J. P. A. (2015). Reproducibility in science: Improving the standard for basic and preclinical research. *Circulation Research, 116*(1), 116-126.

Bell, V. (2010, 4. März). How reliable are fMRI results? [Blogartikel]. *Mind Hacks.* Abgerufen unter https://mindhacks.com/2010/03/04/how-reliable-are-fmri-results

Bennett, C. M., Baird, A. A., Miller, M. B., & Wolford, G. L. (2012). Neural correlates of interspecies perspective taking in the post-mortem Atlantic Salmon: An argument for multiple comparisons correction. *Journal of Serendipitous and Unexpected Results, 1*(1), 1-5.

Bennett, C. M., & Miller, M. B. (2010). How reliable are the results from functional magnetic resonance imaging?. *Annals of the New York Academy of Sciences, 1191*(1), 133-155.

Bestätigungstendenz. (2019). In *Dorsch: Lexikon der Psychologie.* Abgerufen unter https://portal.hogrefe.com/dorsch/bestaetigungstendenz

Bezeau, S., & Graves, R. (2001). Statistical power and effect sizes of clinical neuropsychology research. *Journal of clinical and experimental neuropsychology, 23*(3), 399-406.

Bishop, D. (2013, 7. Juni). Interpreting unexpected significant results [Blogartikel]. *BishopBlog.* Abgerufen unter https://deevybee.blogspot.com/2013/06/interpreting-unexpected-significant.html

Bishop, D. V., & Thompson, P. A. (2016). Problems in using p-curve analysis and text-mining to detect rate of p-hacking and evidential value. *PeerJ, 4.*

Brembs, B. (2011, 18. August). Retractions correlate better with 'impact factor' than citations [Blogartikel]. *Björn Brembs' Blog.* Abgerufen unter http://blogarchive.brembs.net/news.php?item.766.11

Brembs, B., Button, K., & Munafò, M. (2013). Deep impact: Unintended consequences of journal rank. *Frontiers in Human Neuroscience, 7*(291), 1-12.

Brookshire, B. (2018, 16. Dezember). Sometimes a failure to replicate a study isn't a failure at all [Blogartikel]. *Science News.* Abgerufen unter https://www.sciencenews.org/blog/scicurious/willpower-ego-depletion-social-psychology-replication

Brunner, J., & Schimmack, U. (2018). Estimating population mean power under conditions of heterogeneity and selection for significance. *Meta-Psychology,* 1-24.

Buranyi, S. (2017, 27. Juni). Is the staggeringly profitable business of scientific publishing bad for science?. *The Guardian.* Abgerufen unter https://www.theguardian.com/science/2017/jun/27/profitable-business-scientific-publishing-bad-for-science

Carp, J. (2012a). On the plurality of (methodological) worlds: Estimating the analytic flexibility of fMRI experiments. *Frontiers in Neuroscience, 6*(149), 1-13.

Carp, J. (2012b). The secret lives of experiments: Methods reporting in the fMRI literature. *NeuroImage, 63*(1), 289-300.

Center for Open Science. (2019). Openness is a core value of scientific practice [Webseite]. Abgerufen unter https://osf.io/tvyxz/wiki/home

Cesario, J. (2014). Priming, replication, and the hardest science. *Perspectives on Psychological Science, 9*(1), 40-48.

Chambers, C. (2017). *The seven deadly sins of psychology: A manifesto for reforming the culture of scientific practice.* New Jersey, USA: Princeton University Press.

Chase, L. J., & Chase, R. B. (1976). A statistical power analysis of applied psychological research. *Journal of Applied Psychology, 61*(2), 234-237.

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *The Journal of Abnormal and Social Psychology, 6*5(3), 145-153.

Colquhoun, D. (2012, 29. Juni). Queen Mary University of London trying to commit scientific suicide? [Blogartikel]. *DC's Improbable Science.* Abgerufen unter http://www.dcscience.net/2012/06/29/is-queen-mary-university-of-london-trying-to-commit-scientific-suicide

Comte, A. (1830). *Cours de philosophie positive.* Paris: Bachelier.

Cuddy, A. J., Schultz, S. J., & Fosse, N. E. (2018). P-curving a more comprehensive body of research on postural feedback reveals clear evidential value for power-posing effects: Reply to Simmons and Simonsohn (2017). *Psychological Science, 29*(4), 656-666.

Curry, S. (2012, 13. August). Sick of impact factors [Blogartikel]. *Reciprocal Space.* Abgerufen unter http://occamstypewriter.org/scurry/2012/08/13/ sick-of-impact-factors

Dang, J., Liu, Y., Liu, X., & Mao, L. (2017). The ego could be depleted, providing initial exertion is depleting. *Social Psychology, 48*(4), 242-245.

Deutsche Gesellschaft für Psychologie. (2019). Übersicht über Studienorte Psychologie [Webseite]. Abgerufen unter https://studium.dgps.de/studienorte/uebersicht-ueber-studienorte-psychologie

Doyen, S., Klein, O., Pichon, C. L., & Cleeremans, A. (2012). Behavioral priming: It's all in the mind, but whose mind?. *PLoS ONE, 7*(1), 1-7.

Earp, B. D., & Trafimow, D. (2015). Replication, falsification, and the crisis of confidence in social psychology. *Frontiers in Psychology, 6*(621), 1-11.

Everett, J. A. C., & Earp, B. D. (2015). A tragedy of the (academic) commons: Interpreting the replication crisis in psychology as a social dilemma for early-career researchers. *Frontiers in Psychology, 6*(1152), 1-4.

Fanelli, D. (2010). “Positive” results increase down the hierarchy of the sciences. *PLoS ONE, 5*(4), 1-10.

Fanelli, D. (2012). Negative results are disappearing from most disciplines and countries. *Scientometrics, 90*(3), 891-904.

Fanelli, D. (2013). Positive results receive more citations, but only in some disciplines. *Scientometrics, 94*(2), 701-709.

Fang, F. C., & Casadevall, A. (2011). Retracted science and the Retraction Index. *Infection and Immunity, 79*(10), 3855–3859.

Fang, F. C., Steen, R. G., & Casadevall, A. (2012). Misconduct accounts for the majority of retracted scientific publications. *Proceedings of the National Academy of Sciences, 109*(42), 17028-17033.

Ferguson, C. J., & Heene, M. (2012). A vast graveyard of undead theories: Publication bias and psychological science's aversion to the null. *Perspectives on Psychological Science, 7*(6), 555-561.

Fiedler, K., & Schwarz, N. (2015). Questionable research practices revisited. *Social Psychological and Personality Science, 7*(1), 45–52.

Fisher, R. A. (1926). The arrangement of field experiments. *Journal of the Ministry of Agriculture of Great Britain, 33,* 503-513.

Franz, C. (2019). *Package 'cramer'* [PDF]. Abgerufen unter https://cran.r-project.org/web/packages/cramer/cramer.pdf

Friese, M., Loschelder, D. D., Gieseler, K., Frankenbach, J., & Inzlicht, M. (2019). Is ego depletion real? An analysis of arguments. *Personality and Social Psychology Review, 23*(2), 107-131.

Gervais, W. (2014, 5. März). More Power [Blogartikel]. *Will Gervais.* Abgerufen unter https://willgervais.com/blog/2014/3/5/more-power

Glöckner, A., Fiedler, S., & Renkewitz, F. (2018). Belastbare und effiziente Wissenschaft. *Psychologische Rundschau, 69*(1), 22-36.

GNU Project (2019). PSPP [Software]. Abgerufen unter https://gnu.org/software/pspp

Greenwald, A. G. (1975). Consequences of prejudice against the null hypothesis. *Psychological Bulletin, 82*(1), 1-20.

Grieneisen, M. L., & Zhang, M. (2012). A comprehensive survey of retracted articles from the scholarly literature. *PLoS ONE, 7*(10), 1-15.

Hagger, M. S., Chatzisarantis, N. L., Alberts, H., Anggono, C. O., Batailler, C., Birt, A. R., ... & Calvillo, D. P. (2016). A multilab preregistered replication of the ego-depletion effect. *Perspectives on Psychological Science, 11*(4), 546-573.

Hemmelder, V. (2016, 21. Januar). The lessons we learned from a dead fish [Blogartikel]. *Harvard Neuro Blog.* Abgerufen unter https://harvardneuro.wordpress.com/2016/01/21/tbt-the-lessons-we-learned-from-a-dead-fish

Hoekstra, R., Finch, S., Kiers, H. A., & Johnson, A. (2006). Probability as certainty: Dichotomous thinking and the misuse of p values. *Psychonomic Bulletin & Review, 13*(6), 1033-1037.

Hubert, M. (2015, 28. Januar). Neurowissenschaften in der Glaubwürdigkeitskrise. *Deutschlandfunk.* Abgerufen unter https://www.deutschlandfunk.de/qualitaetssicherung-neurowissenschaften-in-der.676.de.html?dram:article_id=310005

IBM. (2017). SPSS Statistics [Software]. Abgerufen unter https://ibm.com/products/spss-statistics

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS Medicine, 2*(8), 696-701.

Ioannidis, J. P. (2008). Why most discovered true associations are inflated. *Epidemiology, 19*(5), 640-648.

Jarrett, C. (2016, 16. September). Ten famous psychology findings that it's been difficult to replicate. *Research Digest.* Abgerufen unter https://digest.bps.org.uk/2016/09/16/ten-famous-psychology-findings-that-its-been-difficult-to-replicate

JASP Team (2019). JASP [Software]. Abgerufen unter https://jasp-stats.org

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science, 23*(5), 524-532.

Johnson, J. A. (2013, 22. Juli). Are research psychologists more like detectives or lawyers? [Blogartikel]. *Psychology Today.* Abgerufen unter https://psychologytoday.com/us/blog/cui-bono/201307/are-research-psychologists-more-detectives-or-lawyers-0

Jussim, L. (2016, 26. Februar). Are most published social psychology findings false? [Blogartikel]. *Psychology Today.* Abgerufen unter https://psychologytoday.com/intl/blog/rabble-rouser/201602/are-most-published-social-psychology-findings-false

Kara, S. (2015, 23. April). Zu schön, um wahr zu sein. *DIE ZEIT.* Abgerufen unter https://web.archive.org/web/20151028234746/https://www.zeit.de/2015/17/sozialpsychologie-professor-daten-manipulation

Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review, 2*(3), 196-217.

Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams Jr, R. B., Alper, S., … & Batra, R. (2018). Many Labs 2: Investigating variation in replicability across samples and settings. *Advances in Methods and Practices in Psychological Science, 1*(4), 443-490.

Kuhn, D. (1989). Children and adults as intuitive scientists. *Psychological Review, 96*(4), 674-689.

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin, 108*(3), 480-498.

Kupfersmid, J. (1988). Improving what is published: A model in search of an editor. *American Psychologist, 43*(8), 635-642.

Labbé, C. (2010). *Ike Antkare one of the great stars in the scientific firmament* [PDF]. Abgerufen unter https://hal.archives-ouvertes.fr/hal-01354123/document

Laws, K. R. (2013). Negativland - a home for all findings in psychology. *BMC Psychology, 1*(2), 1-8.

Leggett, N. C., Thomas, N. A., Loetscher, T., & Nicholls, M. E. (2013). The life of p: "Just significant" results are on the rise. *The Quarterly Journal of Experimental Psychology, 66*(12), 2303-2309.

Levelt, W. J. M., Noort, E., & Drenth, P. J. D. (2012). *Falende wetenschap: De frauduleuze onderzoekspraktijken van sociaal-psycholoog Diederik Stapel* [PDF]. Tilburg: Tilburg University. Abgerufen unter https://pure.mpg.de/rest/items/item_1569964_8/component/file_1569967/content

Mahoney, M. J. (1977). Publication prejudices: An experimental study of confirmatory bias in the peer review system. *Cognitive Therapy and Research, 1*(2), 161-175.

Makel, M. C., Plucker, J. A., & Hegarty, B. (2012). Replications in psychology research: How often do they really occur?. *Perspectives on Psychological Science, 7*(6), 537-542.

Marcus, A., & Oransky, I. (2013, 7. April). Time for a Reproducibility Index: What's behind paper retractions?. *Lab Times online.* Abgerufen unter https://web.archive.org/web/20171117160551/http://www.labtimes.org/labtimes/ranking/dont/2013_04.lasso

Martin, G. N., & Clarke, R. M. (2017). Are psychology journals anti-replication? A snapshot of editorial practices. *Frontiers in Psychology, 8*(523), 1-6.

Masicampo, E. J., & Lalande, D. R. (2012). A peculiar prevalence of p values just below .05. *The Quarterly Journal of Experimental Psychology, 65*(11), 2271-2279.

Maslany, A. J., & Campbell, J. I. D. (2013). Failures to replicate hyper-retrieval-induced forgetting in arithmetic memory. *Canadian Journal of Experimental Psychology, 67*(1), 72-77.

Matzke, D., Nieuwenhuis, S., van Rijn, H., Slagter, H. A., van der Molen, M. W., & Wagenmakers, E. J. (2015). The effect of horizontal eye movements on free recall: A preregistered adversarial collaboration. *Journal of Experimental Psychology: General, 144*(1), 1–15.

Munafò, M. R., Nosek, B. A., Bishop, D. V., Button, K. S., Chambers, C. D., du Sert, N. P., … & Ioannidis, J. P. (2017). A manifesto for reproducible science. *Nature Human Behaviour, 1*(21), 1-9.

Nature Neuroscience Editors. (2013). Raising standards. *Nature Neuroscience, 16*(5), 517.

Nelson, L. D., Simmons, J., & Simonsohn, U. (2018). Psychology's renaissance. *Annual Review of Psychology, 69*(1), 511-534.

Neuroskeptic. (2012, 30. Juni). False Positive Neuroscience? [Blogartikel]. *Neuroskeptic.* Abgerufen unter http://blogs.discovermagazine.com/neuroskeptic/2012/06/30/false-positive-neuroscience

Newcombe, R. G. (1987). Towards a reduction in publication bias. *British Medical Journal, 295*(6599), 656-659.

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology, 2*(2), 175-220.

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., … & Yarkoni, T. (2015). Promoting an open research culture. *Science, 348*(6242), 1422-1425.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251), 1-8.

Page, B. (2019, 21. Februar). Elsevier records 2% lifts in revenue and profits. *The Bookseller.* Abgerufen unter https://www.thebookseller.com/news/elsevier-records-2-lifts-revenue-and-profits-960016

Palus, S. (2015, 8. Dezember). Diederik Stapel now has 58 retractions [Blogartikel]. *Retraction Watch.* Abgerufen unter http://retractionwatch.com/2015/12/08/diederik-stapel-now-has-58-retractions

Pashler, H., & Harris, C. R. (2012). Is the replicability crisis overblown? Three arguments examined. *Perspectives on Psychological Science, 7*(6), 531-536.

Pashler, H., & Wagenmakers, E. J. (2012). Editors' introduction to the special section on replicability in psychological science: A crisis of confidence?. *Perspectives on Psychological Science, 7*(6), 528-530.

Piwowar, H. A., Day, R. S., & Fridsma, D. B. (2007). Sharing detailed research data is associated with increased citation rate. *PLoS ONE, 2*(3), 1-5.

Piwowar, H. A., & Vision, T. J. (2013). Data reuse and the open data citation advantage. *PeerJ, 1*(175), 1-25.

PLoS Medicine Editors. (2006). The impact factor game. *PLoS Medicine, 3*(6), 707-708.

Priem, J., Groth, P., & Taraborelli, D. (2012). The altmetrics collection. *PLoS ONE, 7*(11), 1-2.

Priem, J., Taraborelli, D., Groth, P., & Neylon, C. (2010). Altmetrics: A manifesto [Webseite]. Abgerufen unter http://altmetrics.org/manifesto

Poldrack, R. A., Baker, C. I., Durnez, J., Gorgolewski, K. J., Matthews, P. M., Munafò, M. R., … & Yarkoni, T. (2017). Scanning the horizon: Towards transparent and reproducible neuroimaging research. *Nature Reviews Neuroscience, 18*(2), 115-143.

R Core Team (2013). R: A language and environment for statistical computing. *R Foundation for Statistical Computing.* Abgerufen unter https://r-project.org

RStudio Team (2018). RStudio: Integrated development for R. *RStudio, Inc.* Abgerufen unter https://rstudio.com

Reuning, A. & Meyer, A. (2019, 20. Januar). Wenn Studien einem zweiten Blick nicht standhalten. *Deutschlandfunk.* Abgerufen unter https://deutschlandfunk.de/signifikant-oder-nicht-wenn-studien-einem-zweiten-blick.740.de.html?dram:article_id=438216

Reysen, S. (2002). *Journal of Articles in Support of the Null Hypothesis: Home.* Abgerufen unter https://web.archive.org/web/20020328103810/http://www.jasnh.com/

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin, 86*(3), 638-641.

Rossner, M., van Epps, H., & Hill, E. (2007). Show me the data. *The Journal of Cell Biology, 179*(6), 1091-1092.

Russell, B. (1929). *Our knowledge of the external world.* London: George Allen & Unwin.

Schimmack, U. (2016, 18. April). Replicability report no. 1: Is ego-depletion a replicable effect? [Blogartikel]. *Replicability-Index.* Abgerufen unter https://replicationindex.com/2016/04/18/is-replicability-report-ego-depletionreplicability-report-of-165-ego-depletion-articles

Schimmack, U. (2017, 16. November). Z-curve: A method for estimating replicability based on test statistics in original studies [Blogartikel]. *Replicability-Index.* Abgerufen unter https://replicationindex.com/2017/11/16/preprint-z-curve-a-method-for-the-estimating-replicability-based-on-test-statistics-in-original-studies-schimmack-brunner-2017

Schimmack, U. (2018a, 25. März). P-curve handles heterogeneity just fine [Blogartikel]. *Replicability-Index.* Abgerufen unter https://replicationindex.com/2018/03/25/open-discussion-forum-67-p-curve-handles-heterogeneity-just-fine

Schimmack, U. (2018b, 28. Dezember). An introduction to anti-social psychology [Blogartikel]. *Replicability-Index.* Abgerufen unter https://replicationindex.com/2018/12/28/an-introduction-to-anti-social-psychology

Schimmack, U. (2018c, 28. Dezember). Social psychology textbook audit: Ego depletion [Blogartikel]. *Replicability-Index.* Abgerufen unter https://replicationindex.com/2018/12/28/socpsytb-ego-depletion

Schimmack, U. (2019a, 11. Januar). Replicability audit of social psychology [Blogartikel]. *Replicability-Index.* Abgerufen unter https://replicationindex.com/2019/01/11/replicability-audit

Schimmack, U. (2019b, 12. Januar). How many false discoveries are published in psychology? [Blogartikel]. *Replicability-Index.* Abgerufen unter https://replicationindex.com/2019/01/12/false-discovery-risk

Schimmack, U. (2019c, 27. August). No crisis in social psychology [Blogartikel]. *Replicability-Index.* Abgerufen unter https://replicationindex.com/2019/08/27/no-crisis-in-social-psychology

Schimmack, U., & Brunner, J. (2019). Estimating replicability with z-curve [Webseite]. Abgerufen unter https://zcurve.shinyapps.io/zcurve19

Schimmack, U., & Heene, M. (2016, 20. April). Die Verdrängung des selektiven Publizierens: 7 Fallstudien von prominenten Sozialpsychologen [Blogartikel]. *Replicability-Index.* Abgerufen unter https://replicationindex.com/2016/04/20/die-verdrangung-des-selektiven-publizierens-7-fallstudien-von-prominenten-sozialpsychologen

Schönbrodt, F. D. (2018). P-checker: One-for-all p-value analyzer [Webseite]. Abgerufen unter http://shinyapps.org/apps/p-checker

Schönbrodt, F. D., Maier, M., Heene, M., & Zehetleitner, M. (2015). Voluntary commitment to research transparency [Webseite]. Abgerufen unter http://researchtransparency.org

Schönbrodt, F. D., & Scheel, A. (2017). FAQ zu Open Data und Open Science in der Sportpsychologie. *Zeitschrift für Sportpsychologie, 24*(4), 134-139.

Schramm, S. (2013, 3. Juni). Das Dilemma der Sozialpsychologie. *Deutschlandfunk.* Abgerufen unter https://www.deutschlandfunk.de/das-dilemma-der-sozialpsychologie.676.de.html?dram:article_id=248720

Schuyt, C. J. M., & van Arkel, E. G. (2012). *Suspicion of scientific misconduct by Dr. Jens Förster* [PDF]. Abgerufen unter https://retractionwatch.files.wordpress.com/2014/04/report_foerster.pdf

Schuyt, C. J. M., & van Arkel, E. G. (2014). *LOWI Report* [PDF]. Abgerufen unter https://retractionwatch.files.wordpress.com/2014/05/translation_lowi.pdf

Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies?. *Psychological Bulletin, 105*(2), 309-316.

Seglen, P. O. (1992). The skewness of science. *Journal of the American Society for Information Science, 43*(9), 628-638.

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*(285), 30-34.

Sterling, T. D., Rosenbaum, W. L., & Weinkam, J. J. (1995). Publication decisions revisited: The effect of the outcome of statistical tests on the decision to publish and vice versa. *The American Statistician, 49*(1), 108-112.

Storer, N. W. (1967). The hard sciences and the soft: Some sociological observations. *Bulletin of the Medical Library Association, 55*(1), 75-84.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22*(11), 1359-1366.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2012). A 21 word solution. *Dialogue, 26*(2), 4-7. Abgerufen unter http://spsp.org/sites/default/files/ dialogue_26%282%29.pdf

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2013). Life after p-hacking. Meeting of the Society for Personality and Social Psychology (SPSP), New Orleans, LA, 17. - 19. Januar 2013. Abgerufen unter http://dx.doi.org/10.2139/ssrn.2205186

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2018, 8. Januar). P-curve handles heterogeneity just fine [Blogartikel]. *Data Colada.* Abgerufen unter http://datacolada.org/67

Simmons, J. P., & Simonsohn, U. (2017). Power posing: P-curving the evidence. *Psychological Science, 28*(5), 687-693.

Simonsohn, U. (2013). Just post it: The lesson from two cases of fabricated data detected by statistics alone. *Psychological Science, 24*(10), 1875-1888.

Simonsohn, U. (2017a). P-curve app 4.06 [Webseite]. Abgerufen unter http://p-curve.com/app4

Simonsohn, U. (2017b, 30. November). *The R Code behind the p-curve app 4.06* [Quelltext]. Abgerufen unter http://p-curve.com/app4/pcurve_app4.06.r

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014a). P-curve: A key to the file-drawer. *Journal of Experimental Psychology: General, 143*(2), 534-547.

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014b). P-curve and effect size: Correcting for publication bias using only significant results. *Perspectives on Psychological Science, 9*(6), 666-681.

Simonsohn, U., Simmons, J. P., & Nelson, L. D. (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*(6), 1146-1152.

Smith, R. A., Levine, T. R., Lachlan, K. A., & Fediuk, T. A. (2002). The high cost of complexity in experimental design and data analysis: Type I and type II error rates in multiway ANOVA. *Human Communication Research, 28*(4), 515-530.

Srivastava, S. (2012a, 10. Februar). Does your p-curve weigh as much as a duck? [Blogartikel]. *The Hardest Science.* Abgerufen unter https://thehardest science.com/2012/02/10/does-your-p-curve-weigh-as-much-as-a-duck

Srivastava, S. (2012b, 27. September). A pottery barn rule for scientific journals [Blogartikel]. *The Hardest Science.* Abgerufen unter https://thehardest science.com/2012/09/27/a-pottery-barn-rule-for-scientific-journals

Strathern, M. (1997). ‘Improving ratings’: Audit in the British university system. *European Review, 5*(3), 305-321.

Taylor, M. (2013, 17. Januar). Hiding your research behind a paywall is immoral. *The Guardian.* Abgerufen unter https://www.theguardian.com/science/blog/2013/jan/17/open-access-publishing-science-paywall-immoral

Thelwall, M., Haustein, S., Larivière, V., & Sugimoto, C. R. (2013). Do altmetrics work? Twitter and ten other social web services. *PLoS ONE, 8*(5), 1-7.

Ulrich, R., & Miller, J. (2015). p-hacking by post hoc selection with multiple opportunities: Detectability by skewness test?. Comment on Simonsohn, Nelson, and Simmons (2014). *Journal of Experimental Psychology: General, 144*(6), 1137-1145.

United Nations (1948). *Universal declaration of human rights*. Abgerufen unter https://www.un.org/en/universal-declaration-human-rights

van der Heijden, A. J., Groenen, P. J. F., Zeelenberg, R., & te Lindert, R. (2014). *Report of the Smeesters follow-up investigation committee* [PDF]. Rotterdam: Erasmus University. Abgerufen unter https://web.archive.org/web/20170311224705/http://www.eur.nl/fileadmin/ASSETS/press/2014/maart/Report_Smeesters_follow-up_investigation_committee.final.pdf

van der Sar, E. (2015, 27. Juni). Sci-Hub Tears Down Academia's "Illegal" Copyright Paywalls [Blogartikel]. *TorrentFreak.* Abgerufen unter https://torrentfreak.com/sci-hub-tears-down-academias-illegal-copyright-paywalls-150627

Veröffentlichen. (2019). In *Duden Online-Wörterbuch.* Abgerufen unter https://www.duden.de/rechtschreibung/veroeffentlichen

Wagenmakers, E. J., Beek, T., Dijkhoff, L., Gronau, Q. F., Acosta, A., Adams Jr, R. B., ... & Bulnes, L. C. (2016). Registered replication report: Strack, Martin, & Stepper (1988). *Perspectives on Psychological Science, 11*(6), 917-928.

Walster, G. W., & Cleary, T. A. (1970). A proposal for a new editorial policy in the social sciences. *The American Statistician, 24*(2), 16-19.

Wason, P. C. (1960). On the failure to eliminate hypotheses in a conceptual task. *Quarterly Journal of Experimental Psychology, 12*(3), 129-140.

Wason, P. C. (1968). Reasoning about a rule. *Quarterly Journal of Experimental Psychology, 20*(3), 273-281.

Weir, K. (2015). A reproducibility crisis? The headlines were hard to miss: Psychology, they proclaimed, is in crisis. *Monitor on Psychology, 46*(9), 39.

Whewell, W. (1840). *The philosophy of the inductive sciences: Founded upon their history* (Bd. 1). London: John W. Parker.

Wicherts, J. M., & Bakker, M. (2012). Publish (your data) or (let the data) perish! Why not publish your data too?. *Intelligence, 40*(2), 73-76.

Wicherts, J. M., Bakker, M., & Molenaar, D. (2011). Willingness to share research data is related to the strength of the evidence and the quality of reporting of statistical results. *PLoS ONE, 6*(11), 1-7.

Wicherts, J. M., Borsboom, D., Kats, J., & Molenaar, D. (2006). The poor availability of psychological research data for reanalysis. *American Psychologist, 61*(7), 726-728.

Wicherts, J. M., Veldkamp, C. L., Augusteijn, H. E., Bakker, M., van Aert, R., & van Assen, M. A. (2016). Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid p-hacking. *Frontiers in Psychology, 7*(1832), 1-12.

Zhang, M., & Grieneisen, M. L. (2013). The impact of misconduct on the published medical and non-medical literature, and the news media. *Scientometrics, 96*(2), 573-587.

Zwaan, R. A., Pecher, D., Paolacci, G., Bouwmeester, S., Verkoeijen, P., Dijkstra, K., & Zeelenberg, R. (2017). Participant nonnaiveté and the reproducibility of cognitive psychology. *Psychonomic Bulletin & Review, 25*(5), 1968-1972.